Study Notes inspired by the FDP Exam Study Guide[1] March 16 - April 4, 2020

Terence Lim[2]

Last Updated: December 2019

[1]The FDP Charter and Study Guide copyrights are owned by the FDP Institute.

[2]https://terence-lim.github.io.

Contents

Topic 1. Introduction to Data Science and Big Data

Reading 1.1 *Data Science for Business (Ch. 1, 2)*

Provost, F. and T. Fawcett. (2019). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc. Chapters 1 & 2.

Ch. 1. Introduction: Data-Analytic Thinking

Ch. 2. Business Problems and Data Science Solutions

Keywords

Data mining (p. 2)
Data mining is the extraction of knowledge from data, via technologies that incorporate the fundamental principles of data science. Data mining is used: (1) for general customer relationship management to analyze customer behavior in order to manage attrition and maximize expected customer value; (2) by the finance industry for credit scoring and trading, and in operations via fraud detection and workforce management; and (3) by major retailers from Walmart to Amazon throughout their businesses, from marketing to supply-chain management.

Data science (p. 4)
Data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data.

Churn (p. 4)
Customers switching from one company to another is called churn.

Data-driven decision making (p. 5)
Data-driven decision-making (DDD) refers to the practice of basing decisions on the analysis of data, rather than purely on intuition. For example, a marketer could select advertisements based purely on her long experience in the field and her eye for what will work. Or, she could base her selection on the analysis of data regarding how consumers react to different ads.

Data engineering (p. 5, 7)
There is a lot to data engineering and processing that are not data science, but are more general. New "big data" technologies (such as Hadoop, HBase, and MongoDB), which process datasets that are too large for traditional systems, are used for many tasks, including data engineering in support of data mining techniques and other data science activities,

Data-analytic thinking (p. 12)
Data-analytic thinking is the ability to approach problems "data-analytically." Such a perspective entails a set of fundamental concepts and principles that facilitate careful thinking, and frameworks to structure the analysis so that it can be done systematically.

Target (p. 24)
The target refers to a specific purpose thas has been defined on the population of individuals that we want to predict or group by.

Label (p.24)
The value for the target variable for an individual is often called the individual's label, emphasizing that often (not always) one must incur expense to actively label the data. Acquiring data on the target often is a key data science investment.

Unsupervised data mining (p. 24)
When there is no such target, the data mining problem is referred to as unsupervised. Clustering, an unsupervised task, produces groupings based on similarities, but there is no guarantee that these similarities are meaningful or will be useful for any particular purpose.

Supervised data mining (p. 25)
When a specific target defined, the data mining task is being done for a specific reason of predicting likelihood. This is called a supervised data mining problem.

1.1.1 Data analytic thinking (Ch. 1)

A. Discuss the ubiquity of data opportunities.

With vast amounts of data now available, companies in almost every industry are focused on exploiting data for competitive advantage. In the past, firms could employ teams of statisticians, modelers, and analysts to explore datasets manually, but the volume and variety of data have far outstripped the capacity of manual analysis. At the same time, computers have become far more powerful, networking has become ubiquitous, and algorithms have been developed that can connect datasets to enable broader and deeper analyses than previously possible.

B. Define data science, engineering, and data-driven decision making.

Data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data. In this book, we will view the ultimate goal of data science as improving decision making, as this generally is of direct interest to business. Data-driven decision-making (DDD) refers to the practice of basing decisions on the analysis of data, rather than purely on intuition. Data science needs access to data and it often benefits from sophisticated data engineering that data processing technologies may facilitate, but these technologies are not data science technologies per se. They support data science, but they are useful for much more, such as efficient transaction processing, modern web system processing, and online advertising campaign management. Big data – datasets that are too large for traditional data processing systems – require new "big data" technologies (such as Hadoop, HBase, and MongoDB).

C. Explain data and data science capability as a strategic asset.

Firms in many traditional industries are exploiting new and existing data resources for competitive advantage. They employ data science teams to bring advanced technologies to bear to increase revenue and to decrease costs. In addition, many digital companies are being developed with data mining as a key strategic component (e.g. Amazon, Google, Facebook and Twitter get tremendous value from their data assets). Increasingly, managers need to oversee analytics teams and analysis projects, marketers have to organize and understand data-driven campaigns, venture capitalists must be able to invest wisely in businesses with substantial data assets, and business strategists must be able to devise plans that exploit data.

D. Describe data-analytic thinking.

Data-analytic thinking is the ability to approach problems "data-analytically." When faced with a business problem, you should be able to assess whether and how data can improve performance. Understanding the fundamental concepts, and having frameworks for organizing data-analytic thinking not only will allow one to interact competently, but will help to envision opportunities for improving data-driven decision-making, or to see data-oriented competitive threats.

E. Compare data science and the work of the data scientist.

Data science, like computer science, is a young field, whose particular concerns are fairly new and general principles are just beginning to emerge. A working data scientist is also proficient with certain sorts of specific programming languages and tools, such data mining techniques (e.g., random forests, support vector machines), specific application areas (recommendation systems, ad placement optimization), alongside popular software tools for processing big data. In 10 years' time the predominant technologies will likely have changed, while the general principles of data science likely will change little over the coming decades.

1.1.2 Business problems and data science solutions (Ch. 2)

A. Describe how one transitions from business problems to data mining tasks.

In collaboration with business stakeholders, data scientists decompose a business problem into subtasks. The solutions to the subtasks can then be composed to solve the overall problem. Some of these subtasks are unique to the particular business problem, but others are common data mining tasks. A critical skill in data science is the ability to decompose a data-analytics problem into pieces such that each piece matches a known task for which tools are available.

B. Compare supervised methods to unsupervised methods.

If a specific target can be provided, the problem can be phrased as a supervised one. Supervised tasks require different techniques than unsupervised tasks do, and the results often are much more useful. A supervised technique is given a specific purpose for the grouping – predicting the target. An unsupervised task produces groupings based on similarities, but there is no guarantee that these similarities are meaningful or will be useful for any particular purpose. Classification, regression, and causal modeling generally are solved with supervised methods. Similarity matching, link prediction, and data reduction could be either. Clustering, co-occurrence grouping, and profiling generally are unsupervised.

C. Describe the difference between data mining and using the results of data mining.

Data mining find patterns in historical data to produce a model. Importantly, the historical data have the target value specified. The results of the data mining are used when the model is applied to new data for which we do not know the class value. The model predicts both the class value and the probability that the class variable will take on that value.

D. Describe key aspects of the data mining process, including business understanding, data understanding, data preparation, modeling, and evaluation.

The Business Understanding stage casts the business problem as one or more data science problems. In Data Understanding we need to dig beneath the surface to uncover the structure of the business problem and the data that are available, and then match them to one or more data mining tasks for which we may have substantial science and technology to apply. In the Data Preparation phase, the data are manipulated and converted into forms that yield better results. One very general and important concern during data preparation is to beware of "leaks", where a variable collected in historical data gives information on the target variable but is not actually available when the decision has to be made. The output of Modeling, where data mining techniques are applied to the data, is some sort of model or pattern capturing regularities in the data. The purpose of the Evaluation stage is to assess the data mining results rigorously and to gain confidence that they are valid and reliable, and equally important, to help ensure that the model satisfies the original business goals.

Reading 1.2 *Big Data is a Big Deal*

Getman, G. and Q. Hasan. (2019). Big Data is a Big Deal: An investor's guide to the applications and challenges of alternative data, New York, NY: Lombard Odier.

Retrieved from https://www.lombardodier.com/contents/news/world-in-transition/2019/april/big-data-is-a-big-deal.html

Keywords

Alternative data (p. 5)
Any dataset that does not meet the criteria for traditional financial data (e.g. income statements, balance sheets, press releases) or market data (e.g. pricing, volumes, factors). It is highly diverse and predominantly noise, but does offer a window into real-world commercial activity

Big Data (p. 5)
Complex data sources, where the massive scale, frequency and abstract nature of the data is beyond the ability of traditional relational databases to capture, manage and analyze.

Data Science (p. 5)
Advanced analytical techniques that require knowledge of data sourcing, data mining, data warehousing, programming algorithms, statistical modelling, machine learning, natural language processing and IT visualization tools.

Data Analytics (p. 5)
Synonymous with the term "Data Science."

Moore's Law (p. 5)
Observation made by Intel co-founder Gordon Moore that the complexity and computing power of hardware doubles every 18-24 months.

Bezos Law (p. 5)
"Over the history of the cloud, a unit of computing power price is reduced by 50% approximately every three years."

Unstructured (p. 7)
Information that either does not have a pre-defined data model or is not organized in a pre-defined manner.

Discretionary (p. 9)
Investment approach based on idea that stocks follow earnings and earnings follow revenues: companies that continue to beat on revenues tend to relate to companies that see share price appreciation over time. Discretionary managers develop predictive models using Financial Data, as well as adjustments based on intuition from reading sell-side research, attending conferences and conducting management team interviews.

Quantitative (p. 9)
An investing approach: Predicts security prices, using mathematical and statistical modelling to identify links between pricing, valuation, and other Market Data over time

Financial Data (p. 9)
Income statement, Balance sheet, Cash flows, Ratio analysis, Corporate filings, Competitive analysis, Market sizing.

Market Data (p. 9)
Security pricing, Derivative pricing, Short interest, Volumes, Put/Call spreads, Implied vol, Factor analysis.

Tagging (p. 11)
Converts unstructured datasets into structured indices that are ready for analysis. Mostly systematic, using machine learning and AI tools.

Incrementalism (p.12)
Piecing together rolling cycles recent data can be a strong indicator of the future inflections.

1.2.1 Defining big data

A. Define alternative data, big data, data science and data analytics.

Alternative Data can track real-time operational and commercial activity, to make predictive links to company-specific business trends. The data tracks earnings and inherently speaks to fundamental trends, but the skills necessary to process it is a subset of Data Science and Data Analytics (data sourcing, data mining, data warehousing, programming algorithms, statistical modelling, machine learning, natural language processing and IT visualization tools) and therefore inherently quantitative.

While not all Alternative Data necessarily comes from Big Data, nearly all of the growth in alternative data has come from Big Data sources, such as consumer spending, online behavior, geolocation, government statistics and contributory databases. It is based on the analysis of new, messy Alternative Data.

Data Science, which is synonymous with Data Analytics, refers to advanced analytical techniques that require knowledge of data sourcing, data mining, data warehousing, programming algorithms, statistical

modelling, machine learning, natural language processing and IT visualization tools.

B. Discuss the recent phenomenon of data proliferation in terms of:

- **Moore's Law and Bezos law.** Over the last 50 years, Moore's Law has solidified itself as the golden rule for the technology industry, with exponential growth of technological innovations necessary to perform advanced data science. Bezos law is its anecdotal cousin for the revolutionary ease of access to affordable cloud computing and storage.
- **Digitization of consumer lifestyles and business processes.** In the last five years, Internet access has gone up by 1.3 billion people, with usage now reaching 47% of the world's population. Global internet traffic will continue to grow 30% annually, increasing threefold in just five years, with nearly all this growth from smartphones, which are projected to account for 50% of all internet traffic by 2022. Smartphones are filled with a myriad of sensors that collect data on a very granular level: our locations, spending habits and online behaviors have been datafied, anonymized and indexed for analysis.
- **Business intelligence tools.** To derive actionable investment insights while the data is still relevant, users need tools that can help them to digest and present information in a way that is seamless and natural to our decision-making processes. Advancements in data analysis and affordable access to external vendors (such as QlikView, Tableau, and Splunk) have recently paved the way for the business intelligence and visualization software necessary to sort, summarize and present terabytes of data in a concise format, and make Big Data more practical.
- **Big data intermediaries.** Vendors began by selling analytics directly to businesses as ways to enhance organizational return (e.g. hotel statistics to lodging and leisure companies, ad spend data for the marketing industry, image recognition and diagnostic databases within the healthcare industry). These vendors eventually realized that this data might offer commercial insights for investing and began offering datasets to buy-side firms. The USD 32 billion global Big Data market and ecosystem, is expected to continue to grow at an annual rate of nearly 20%, reaching USD 156.7 billion by 2026. But the total buy-side firms spend on alternative data only accounts for an estimated 1.3% of the current market.

C. Compare types and sources of alternative data.

We tend to focus on five broad categories of alternative data:

1. Consumer spending: Credit cards, Emails, Point of sale terminals.
2. Government statistics: Trade data, DMV registrations, Building permits, Personal bankruptcy filings, Local number portability, Bill of lading.
3. Online behavior: Web traffic, Web searches, Web scraping, App downloads, Clickstream, Cookies.
4. Geolocation: Satellites, Drones, Smartphones.
5. Contributory databases: Point of sale, Hardware distribution, Hotel stays, Media buys, Flight bookings, Drug prescriptions.

This list is not all encompassing. While there are a number of other alternative data sources such as social media, we generally find them to be less reliable or relevant.

D. List five characteristics used to assess the quality of a big data source.

1. Do we want to do more work with the data?
2. Does this data make us smarter?
3. Does this data match up against problems we are trying to solve for?
4. What could this data mean for company revenues and business models?
5. Do the insights exceed the cost of acquiring and managing the data?

1.2.2 The benefits and limitations of big data for investment decisions

A. Explain how big data provides a new source of fundamental insights, relative to discretionary and quantitative strategies, by using the four data quadrants of any investment framework (financial data, alternative data, market data and internal data).

Instead of a common misconception of Big Data as a tool turns discretionary managers (who traditionally develop predictive models using Financial Data, adjusted based on intuition) into "quantamental" or adds to the vast toolkit of quantitative managers (who traditionally use mathematical and statistical modelling to identify links between pricing, valuation and other Market Data), we prefer to think of Big Data as a new wave of thought that is worthy of a new investment framework, based on the analysis of new, messy Alternative Data. Building an objective, systematic and scientific process underpinned by evidence – comprising data from app usage, web traffic, credit card transactions, IP information and other alternate date – can provide very precise, real-time insights into near-term earnings inflections while keeping personal biases at bay.

B. Describe a process of harnessing data-driven insights including

- **Sourcing:** Sourcing is finding vendors that have something interesting; there are now thousands of datasets aimed at the buy-side, but access to more data for analysis does not mean you gain more insight. Sourcing data successfully requires a fundamental knowledge of what moves stocks as well as thinking creatively about what specific types of data could tell us about different types of companies. Properly evaluating, tagging and mapping a single dataset can take from weeks to months without the proper systems in place, therefore crucial to develop filters to maximize time value from a commercial point of view and short list data sets based on fundamental need.
- **Backtesting:** Once the more promising datasets are narrowed down, the data needs to be run through systems to see where it applies and if there is value in implementing it. The reality is most data is garbage, apt to giving users false or inconsistent signals. The backtesting process is both fundamental and quantitative, is not perfectly linear, and requires robust evaluation criteria and understanding of data biases and limitations to be successful
- **Tagging:** Converts unstructured datasets into structured series that are ready for analysis, mostly using systematic machine learning and AI tools. It is an extremely time consuming exercise to develop the proper infrastructure and managers have to make the decision as to whether they will focus the resources on tagging the data in house, or leveraging the vendor ecosystem.
- **Mapping:** Once the data is structured into a clean index it can be mapped to specific companies, and categorized by relevance to a particular segment or financial metric of a company.

- **Visualization:** Front-end visualization tools, which systematize and contextualize the data, are essential for a fundamental analyst to extract and understand the data quickly enough to take action while signals are still topical.

C. Discuss the scope and limitations of big data including the backward nature of data and the concept of incrementalism.

An obvious drawback of data is that it is inherently backward looking. If modeled well, recent data trends can be extrapolated into the near future, but one needs to embrace the limited scope of data relevancy over time. This is where the concept of incrementalism comes in – data can be a strong indicator of the future inflections.

1.2.3 Challenges and unique skill sets needed to translate big data into actionable insights

A. Describe barriers to entry and the learning curve.

Building out a robust data science platform and proper visualization tools requires considerable capital investments and roughly 12-18 months of development before investors can reap any benefits. Costs can run into the multiple millions of dollars just to maintain the talent and datasets necessary to continue to run the operation. Even when managers commit to the buildout and costs, they often find that they cannot extract enough consistent alpha to justify the upkeep due to a mismatch to their investment culture/framework.

B. Explain the need to critically evaluate data sources.

The most critical distinction among investment managers using Big Data is their ability to gauge the accuracy and value of a broad range of datasets. Not all types of alternative data can be trusted. Even within more trusted types of data, investment managers must be able to appraise the incremental value between various vendors. Furthermore, it is essential to appreciate the costs associated with time-spend. Structuring datasets in house incurs meaningful human capital to clean. The real competitive edge is tied to the ability to discern the quality and potential return of a new dataset at an individual stock and KPI level. Data is not a cure-all, and 95% of it is often incorrect or misleading when taken at face value. The key is developing the processes and experience necessary to filter out the noise and extract the 2-3% of data that actually provides meaningful insights.

C. Discuss three major facets of refining data.

1. Normalizing data: address issues in the data such as compatibility, changes, and bias.
2. Treatment of errors and outliers: how to validate data and treat outlier transactions on a company-by-company basis.
3. Reflecting change: Underlying constituents go in and out of data indices periodically. Companies themselves can make several changes ranging from accounting policies, transformative acquisitions, to moving into new product segments or geographies.

D. Compare discretionary and quantitative managers in the context of aligning investment culture with big data requirements.

Discretionary managers invest in a certain way based on bottom-up fundamental analysis and gut instinct, and often resist the push to drastically change the way in which they research ideas. They often hire a group of token data scientists in siloed roles. Whereas alternative data speaks to real-time trends lasting only a few weeks, discretionary manager form long-term conviction and hypothesis.

Quantitative strategies are used to working with very long historical datasets over multiple market cycles. Alternative data on the other hand will often have only 4-5 years of history at most. Quantitative strategies generally spot patterns across thousands of stocks, diversifying smaller bets across a huge number of positions. Alternative datasets on the other hand tend to be very limited in scope. Fully systematic processes find alternative data to be too unruly due to its unstructured nature that requires a very labor intensive process.

1.2.4 Implementation and costs involved in utilizing alternative data in the investment process

A. Describe considerations in developing realistic expectations of the benefits of blending artificial intelligence with discretionary decision making.

1. Quantifying mental models and the logic from fundamental research (i.e. develop a mathematical model for the "gut feeling") helps reference and iterate on successful work, keep personal biases at bay and use technology to provide scalability
2. Data, which speaks only to to one aspect of the top-line growth drivers, tells humans where to focus, what questions to ask, and of whom, but personal experience is key to optimize the models for current market biases driving stock prices and unquantifiable risks.
3. There is only so much a model can infer about the future from looking back at price or other data. It helps us understand the fundamental topline trends of a company systematically, but the rest requires discretionary judgement
4. Find the balance in constantly battling human biases based on past experiences while avoiding being led astray by blind trust of machines.

B. Discuss resource requirements and methods of reducing their costs.

One of the largest misconceptions when it comes to alternative data is the belief that an investment manager's edge is related to their ability to access the most expensive, robust, or niche datasets and process them in-house. We can leverage the deep level of specialization within the external vendor ecosystem, who often have dozens of data scientists dedicated to a single dataset and can deliver better quality of tagging. Shifting to cloud computing and open source software allows us to be more nimble and faster to market, while incurring lower up-front costs. Price of data can vary greatly based on willingness to accept reduced frequency, slight lag, or a smaller subset of data that may be sufficient. With the proper evaluation tools and some creativity, managers can blend multiple smaller lower-cost datasets across different types of data.

Reading 1.3 *Big Data and Investment Management*

Kaul, S. (2016). Chapter 8. Big Data and Investment Management. In Citi GPS: Global Perspectives & Solutions, (pages 58-65)

Keywords

Quantitative fundamental analysis (p. 60, 62)
The expansion of new data sets used as a stand-in for the opinions, behaviors and physical responses that drive much of the fundamental 'feel' that helps inform investment decisions from discretionary portfolio managers. This 'datafication' of previously interpretive and predictive data will allow quantitative modeling to delve much more deeply into a broader set of hypothesis that were previously only used by discretionary fundamental managers and achieve the same outcome.

1.3.1 Facets of the big data phenomenon

A. Describe three facets of the big data phenomenon

1. The volume of data that can be incorporated into investment models
2. The velocity at which that data can be processed
3. The variety of data that can now be accessed and the fact that many of these data sets did not exist a few years ago.

1.3.2 Investment managers' use big of data

A. Describe the spectrum of big data adoption.

Currently the spectrum of big data adoption is broad. Early adopters are investing heavily in developing a whole technology stack and hiring data scientists to drive investment research and make it an integral part of the front offices. Another segment of funds is experimenting with big data by either enlisting big data techniques that extend their existing research capabilities through proofs of concept or by piloting content from third-party providers utilizing outsourced access to big data technology and new data sets. Most investment firms are not yet focused on big data because they lack the institutional momentum, the skill set and the business case to build out these capabilities in the short-term.

B. Identify the players driving the adoption of big data in investment firms.

Experimentation and usage of big data technology is being driven by the front office and not IT. A specific research analyst may put in a request to analyze a new data set to understand its relationship to time-series data leading IT to accommodate the request tactically.

C. Explain why there will be pressure to incorporate big data principles in investment research.

Because early adopters are already locking up access to semi-private data sets to provide their models with an information edge. This is allowing these firms to create a real-time view of supply and demand fundamentals compared to increasingly 'stale' fundamental data used in traditional quantitative models. Firms that employ big data techniques could thus gain an advantage over late adopters for some time until these techniques are utilized by more organizations.

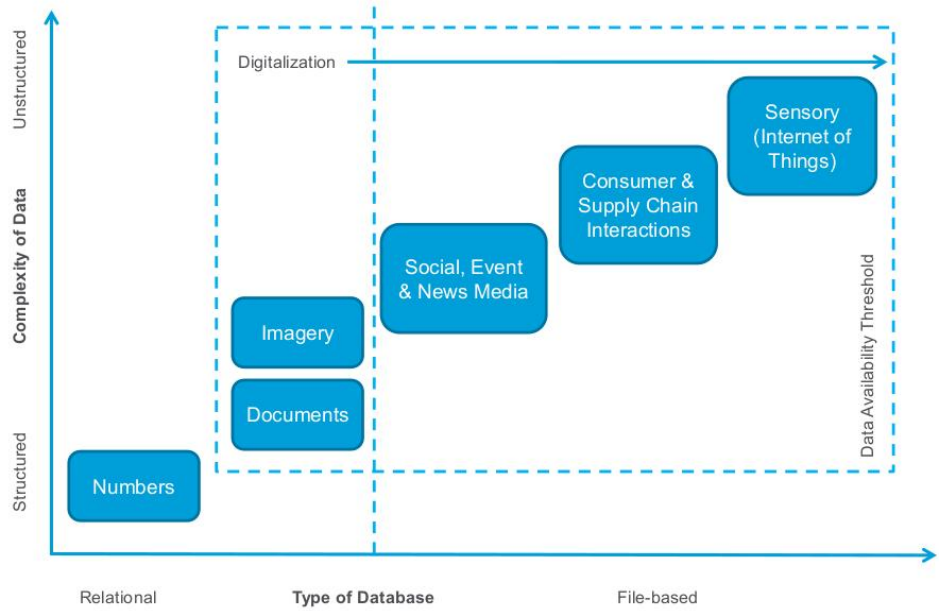
D. Describe third party providers that facilitate efforts to accelerate adoption of big data principles.

A marketplace of third-party providers and data vendors make accessing and investigating new data sets much easier. This allows a broader swath of investment managers to acquire some basic big data capabilities without full-scale infrastructure and staff investments.

1.3.3 The expansion and transformation of quantitative fundamental analysis

A. Describe the evolution of systematic investment management in terms of the complexity of data used and the type of databases used.

Figure 61. Evolution of Systematic Investment Management



Source: Citi Business Advisory Services

B. Compare the traditional decision-making process of a fundamental analyst with the potential decision-making processes that incorporate big data fundamentals.

Traditionally, a discretionary fundamental portfolio manager might be able to talk to corporate executives and pick up from their body language, read more about recent activities of the firm, visit key offices or facilities, and call contacts within the industry. Through these efforts, the discretionary fundamental

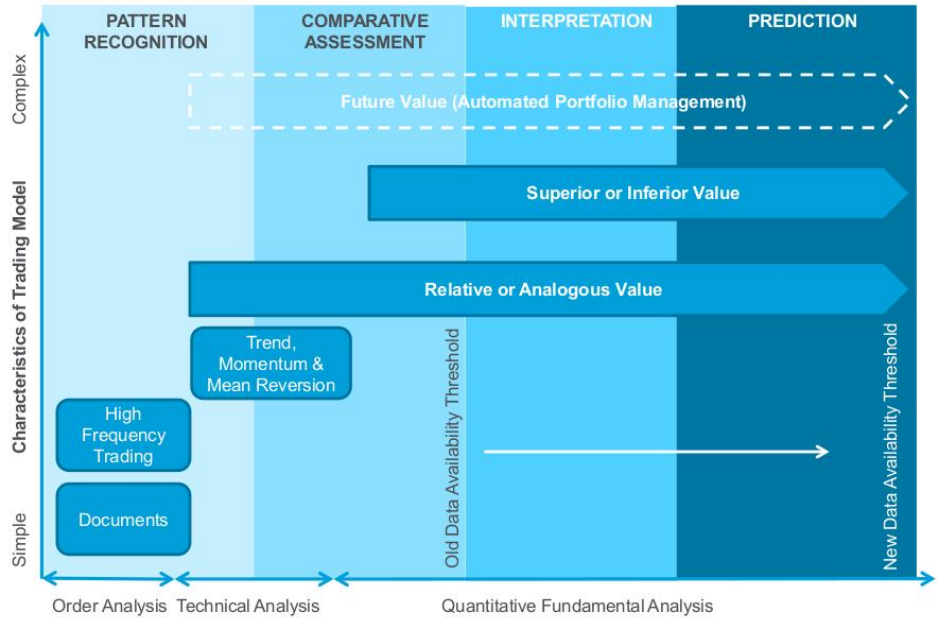
portfolio to predict that the ccompany may soon be announcing a future event, such as opening a new production facility.

'Datafication' of previously interpretive and predictive data can be used to achieve the same outcome: A linguistic program utilizing sentiment analysis flags that there has been a noticeable shift in the ratio of positive and negative words being used in regard to a company. Parsing programs identify that company executives increase use of certain terms (e.g. "research & development", "exploration" and "expansion". A quantitative model may use the correlation of these words to query public records to determine if any new filings have been made, call up satellite imagery and run a time series image analysis to determine where there were noticeable changes, identify individuals that work in the company's new site development team based on an organization chart published in an investor presentation on Google and reference their Facebook and Instagram accounts of to look at the geo-tagging of all photo postings. This combination of factors would be fed into predictive models that indicate a high likelihood of opening a new production facility, and even when it is likely to open.

1.3.4 Emerging portfolio management models that utilize big data principles

A. Describe the evolution of systematic investment management in terms of characteristics of trading models as it progresses from order analysis, then technical analysis to quantitative fundamental analysis.

Figure 62. Evolution of Systematic Investment Management



Source: Citi Business Advisory Services

B. Describe a new type of systematic portfolio management model that bases trade selection on the likely 'future value' of a company.

Based on predicting a likely future event, the model would look for equivalent announcements from the historic record, and quantify the average size and duration of a price responses. It could then put on a position in anticipation of the future event most in line with that model, reassess the price pattern each

day, and determine if the size of the position needs to be increased or decreased based on predictive analysis changes.

Other types of 'future value' systematic portfolio management programs could be given signals about two companies within the same sector that are both likely to have future events, and institute a pairs trade based on historic patterns in anticipation of a change in the relative value between the two companies – in essence, a forward-looking mean reversion trade.

C. Describe intermediate approaches to big data adoption that use a subset of alternative data to improve portfolio management.

Firms might use only one aspect of this new big data landscape to assist in their investment process as they begin to understand the applicability of new data sets. For example, companies may have an advanced capability in analyzing social media, satellite images, or consumer transaction data from a credit card aggregation service.

D. Describe how fundamental discretionary managers can benefit from big data models.

Discretionary fundamental managers may prove even more adept at using the big data tool set as their ability to create hypothesis in a fluid manner is already a core skill. Big data models built to mimic the hypothesis building that discretionary fundamental traders explore could begin to provide discretionary fundamental traders an expanded opportunity pool to assess and choose from – equaling the work of a team of analysts at a faster speed and across a broader universe.

Reading 1.4 *Big Data and Machine Learning in Quantitative Investments (Ch. 2, 4 & 5)*

Guida, T. (2019). Big Data and Machine Learning in Quantitative Investments. West Sussex, UK: John Wiley & Sons Ltd. Chapters 2, 4 & 5.

2. "Taming Big Data", Rado Lipus and Daryl Smith
4. "Implementing Alternative Data in an Investment Process", Vinesh Jha
5. "Using Alternative and Big Data to Trade Macro Assets", Saeed Amen and Iain Clark

Keywords

Quant quake (p.110)
In August 2007 when many (systematic investing) quants suffered their worst losses – before or since – over a three-day period.

Fundamental prediction (p.124)
Prediction of fundamental values, which tend to be more stable than price predictions.

Fundamental law of active management (p.127)
A manager's information ratio (IR), a measure of his or her risk-adjusted active return, is shown to be a function of two things: information coefficient (IC), which is the correlation between the manager's predictions and subsequent realized returns – a measure of skill; and the number of independent bets (N) – a measure of breadth. Captures distinction of depth versus breadth in investment philosophies.

Quantamental investing (p. 127)
Quantamental investing takes many forms, including embracing alternative data, and the usage of traditional quantitative techniques such as backtesting, risk management and portfolio attribution in the context of a fundamental analysis-driven portfolio.

Alternative data life cycle (p.145)
Similar to life cycle of a product from the perspective of innovative technology vendors – noting that alternative data hasn't 'crossed the chasm' (at the time of this writing) and that the toughest part of the adoption cycle is moving from visionary 'early adopters'.

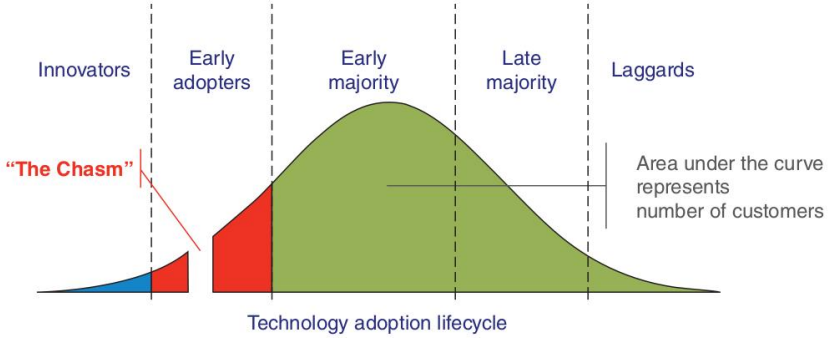


FIGURE 4.1 Technology Adoption Lifecycle
Source: Wikimedia Commons. Author: Craig Chelius https://commons.wikimedia.org/wiki/File:Technology-Adoption-Lifecycle.png

Exhaust data (p.151)
Large amounts of (secondary) data that are generated and collected by companies as part of their everyday business.

Nowcasts (p.151)
The basic principle is that signals about the direction of longer-run economic data releases (e.g. quarterly change in GDP) can be extracted from a large and heterogeneous set of information sources (for example, industrial orders and energy consumption, or other alternative data) that can be generated throughout the month before the release itself is published.

1.4.1 Taming big data (Ch. 2)

A. Discriminate between alternative data and big data.

The term 'alternative data' refers to novel data sources which can be used for investment management analysis and decision-making purposes Alternative data refers to data which was, in the main, created in the past several years and which until very recently has not been available to the investment world. Alternative is not always big and big is not always alternative

B. Contrast drivers of adoption of alternative data with its challenges in the investment community.

The majority of innovators and early adopters are based in the US, with a small percentage of European and an even lower number of Asian funds. Most of the innovators and early adopters have systematic and quantitative investment strategies, and, to a significant degree, consumer-focused discretionary funds. The adoption of alternative data is at the cusp of transitioning into an early majority phase.

The adoption of alternative data within the investment community has been driven by the advancements of financial technology and has improved technological capabilities for analyzing different datasets. Many investors, hedge funds, and asset managers alike view these developments as a complementary tool alongside conventional investment methodologies, offering an advantage over investment managers that have not deployed such capabilities.

It is more challenging for firms driven by fundamental strategies to integrate and research alternative datasets given that the required technical and data infrastructure needed is often not adequate, and that research teams frequently have significant skill set gaps. For large, established traditional asset managers, one significant obstacle is the slow internal process of providing the research team with test data. This procedure often requires (i) due diligence on the new data provider, (ii) signing legal agreements for (in most cases free) test data, and (iii) approval by compliance teams

C. Identify the largest categories of alternative data types in use today.

1. Crowd sourced: Data has been gathered from a large group of contributors, typically using social media or smartphone apps
2. Economic: Data gathered is relevant to the economy of a particular region. Examples include trade flow, inflation, employment, or consumer spending data
3. ESG: Data is collected to help investors identify environmental, social, and governance risks across different companies
4. Event: Any dataset that can inform users of a price-sensitive event for equities. Examples include takeover notification, catalyst calendar or trading alert offerings
5. Financial products: Any dataset related to financial products. Examples include options pricing, implied volatility, ETF, or structured products data
6. Fund flows: Any datasets related to institutional or retail investment activity
7. Fundamental: Data is derived from proprietary analysis techniques and relates to company fundamentals
8. Internet of things: Data is derived from interconnected physical devices, such as Wi-Fi infrastructures and devices with embedded internet connectivity
9. Location: Dataset is typically derived from mobile phone location data
10. News: Data is derived from news sources including publicly available news websites, news video channels or company-specific announcement vendors
11. Price: Pricing data sourced either on or off exchange
12. Surveys and Polls: Underlying data has been gathered using surveys, questionnaires or focus groups

13. Satellite and aerial: Underlying data has been gathered using satellites, drones or other aerial devices
14. Search: Dataset contains, or is derived from, internet search data
15. Sentiment: Output data is derived from methods such as natural language processing (NLP), text analysis, audio analysis, or video analysis
16. Social media: Underlying data has been gathered using social media sources
17. Transactional: Dataset is derived from sources such as receipts, bank statements, credit card, or other data transactions
18. Weather: Data is derived from sources that collect weather-related data, such as ground stations and satellites
19. Web scraping: Data is derived from an automated process that collects specific data from websites on a regular basis
20. Web and app tracking: Data is derived from either (i) an automated process that archives existing websites and apps and tracks specific changes to each website over time or (ii) monitoring website visitor behaviour

D. Evaluate the usefulness of an alternative data set.

Key criteria for assessing alternative data usefulness:

1. Data history length: The earliest point from which historical point in time data is available
2. Data frequency: The frequency with which data can be delivered
3. Universe coverage: How many investable companies the dataset relates to
4. Market obscurity: How well-known this dataset is to institutional investors
5. Crowding factor: How many hedge funds and asset management clients are using this dataset
6. Uniqueness: How unique this specific dataset is
7. Data quality: Completeness, structure, accuracy and timeliness of data
8. Annual price: Subscription price charged by the data provider

E. Describe the likely attributes that differentiate alternative data sets in terms of cost.

It can be very difficult to work out a price, for two reasons. The first is that in many cases new providers' understanding and knowledge of peer or comparable data subscription pricings is non-existent or very limited. Second, data providers do not know how their data will be used by the buy side and how much value or alpha a dataset provides to an asset manager. For some free data sources there might be the indirect cost of data retrieval, cleaning, normalizing, mapping to identifiers, and other preparations to make these data sources useful for research and production at a fund manager

F. Discuss some of the biggest alternative data trends.

1. Is alternative data for equities only? It is applicable to all asset classes (including non-listed equities or privately held firms) and not just to listed equities.

2. Supply-side? In 2017 we saw a large increase in dataset launches of location, web, and app tracking sources.
3. Most common queries? With regard to demand, the top categories enquired about were ESG, Transactional, Sentiment, and Economic data in the majority of months in 2017.

1.4.2 Implementing alternative data in an investment process (Ch. 4)

A. Describe the "quant quake" and how it motivated the search for alternative data.

After poor but not hugely unusual performance in July 2007, many quantitative strategies experienced dramatic losses – 12 standard deviation events or more by some accounts – over the three consecutive days of 7, 8 and 9 August. It turned out that they were all trading very similar strategies. Most equity market-neutral quants traded within a similar universe, controlling risk with similar risk models and for the most part betting on the same alphas built on the same data sources. The search for alternative data sources started during those days.

B. Discuss reasons for "the chasm" in the alternative data adoption life cycle and reasons that the chasm has been difficult to cross for many fund managers.

Buy-side respondents would like to adopt alternative data as part of their process. Relatively few have made significant progress, though the ranks continue to grow. The adoption has fallen a bit short of the hype, even though data and quantitative techniques are far more prevalent currently than at the time of the Quant Quake. The early adopters tend to be those quantitative fund management companies that are already especially data-savvy and that command the resources to experiment with new datasets.

- Some fund managers have expressed concern about crowdedness in alternative datasets
- figuring out which datasets are useful is difficult and turning them into alphas is difficult.
- Perhaps the holdouts who have not embraced alternative data are hoping that value, momentum and mean reversion aren't very crowded, or that their take on these factors is sufficiently differentiated
- It is possible that a behavioural explanation is at work: herding. Some managers to be a better outcome than adopting an alternative data strategy that is innovative but has a short track record and is potentially harder to explain to an allocator or to internal bureaucracy, particularly if it doesn't go well.

C. Discuss methods for improving the efficiency of evaluating datasets for the purpose of finding alpha.

1. Allocating increased research resources specifically to new datasets, setting a clear time horizon for evaluating each, and then making a definitive decision about the presence or absence of added value from a dataset.
2. Building a turnkey backtesting environment which can efficiently evaluate new alphas and determine their potential added value to the existing process. The more mundane data processing, evaluation and reporting aspects can be automated to expedite the process

3. Assigning an experienced quantitative analyst to be responsible for evaluating new datasets – someone who has seen a lot of alpha factors before and can think about how the current one might be similar or different.
4. Increasing outreach to innovative data suppliers rather than what's available from the big data providers, whose products are harder to consider truly alternative.
5. Giving priority to datasets which are relatively easy to test, in order to expedite one's exposure to alternative alpha.
6. Gaining more comfort with the limited history length that we often see with alternative datasets. One can't necessarily judge these datasets by the same criteria of 20-year backtests as we can with more traditional factors

D. Describe issues involved with selecting a data source for evaluation within the context of a quant equity process.

1. First, one must dedicate resources to collect the data, or source it from data providers.
2. It's unclear to most managers which providers' datasets have investment value at first glance
3. At a minimum, a dataset should have sufficient history and breadth; it should be possible to transform the data into something approximating point in time; and it should be tagged, or taggable, to securities.
4. Once a vendor is chosen for evaluation, one needs to carefully examine their datasets. These datasets have not been as thoroughly combed over.
5. One must develop one's own hypotheses on why these datasets may be predictive or useful rather than leveraging published or working papers

E. Explain why and under what circumstances a fundamental prediction may be more appropriate than an asset price prediction when working with alternative data.

Fundamental predictions tend to be more stable than asset price predictions. Therefore, one can potentially build a robust prediction of fundamental values with a relatively short history.

F. Apply the fundamental law of active management and describe how it applies differently to discretionary managers than quant managers.

$IR = IC \times \sqrt{N}$. A manager's information ratio (IR), a measure of his or her risk-adjusted active return, is shown to be a function of two things: The information coefficient (IC), which is the correlation between the manager's predictions and subsequent realized returns – a measure of skill; and the number of independent bets (N) – a measure of breadth. Simplistically, discretionary managers focus on IC and quant managers focus on breadth; a quant strategy is replicable across many assets, but rarely provides high conviction on any particular trade, whereas a fundamental analyst theoretically can provide a high, but unscalable, IC through in-depth research.

G. Describe the transition from fundamental analysis to quantamental analysis.

One use for alternative data among fundamental managers is to gather even deeper insights about a company, without necessarily increasing the number of total bets. A little further up the data-adoption curve, some fundamental teams are ingesting data through user interfaces (UIs) designed to provide visualization, screening and alerts regarding alternative datasets. Finally, some fundamental teams have recently brought in teams to both manage vendor relationships and provide data science tools in-house. One challenge is getting the portfolio managers and analysts to pay attention to the in-house products generated by the data science team. Quantamental teams will need to shift some of their attention away from the typical tools in order to best leverage new datasets.

H. Describe how the alternative data can be used to generate a trading signal using examples including blogger sentiment, online consumer demand, transactional data, and environmental, social and governance (ESG) data.

- TipRanks collects online advice from a variety of sources, including news articles and several financial blog sites. Its proprietary Natural Language Processing algorithm is employed to generate sentiment for each article. From event studies to understand stock price behaviour before and after the publication of a blog article, we can then wrap these signals into a simple stock-scoring algorithm
- Demand for a company's products can be proxied by the amount of attention which is paid, such as by bloggers, to the company's web presence. Alpha-DNA has developed a proprietary scoring system to rank approximately 2000 companies on their overall performance strength across digital platforms (site, search, social) and consumer effectiveness (penetration, engagement, popularity). This digital demand predicts revenue surprises with regularity, hence also leads to profitable portfolios built with the same scores.
- Sandalwood Advisors has collected several unique high-value datasets which capture both online and offline mainland Chinese retail transactions. Aggregate up to the company level, monthly changes in sales or market share are representative of of actual reported quarterly growth rates of revenue growth. Top-ranked stocks also outperformed the universe.
- The CFPB maintains a consumer complaint database, updated daily, which logs complaints by consumers relating to retail financial services. Our hypothesis is that companies with relatively more complaints face greater business risk which should impact stock volatility as well. An ESG-enhanced risk model could be used in several ways. The new factor could be used as a constraint in an optimization process in order to mitigate ESG risk at the portfolio level; portfolio and stock-level ESG risks can be monitored; and one can measure returns residualized to ESG factors for use in, for example, mean reversion stock selection models.

1.4.3 Using alternative and big data to trade macro assets (Ch. 5)

A. Define general concepts within big data and alternative data including "exhaust data".

There are certain characteristics of what constitutes big data, which are collectively known as the 4Vs: volume (big data can range from many gigabytes to terabytes or even petabytes), variety (big data, such as web content, consists of text and other media, not purely numerical data), velocity (it can be generated at high frequencies and at irregular time intervals, such as tick data for traded assets) and veracity (data,

which can emanate from unverified sources or try to actively spread disinformation, can often still require cleaning to remove invalid observations).

Alternative data need not always consist of big data. They are datasets which are not commonly used in finance and not in the mainstream. 'Exhaust data', generated by companies as part of their everyday business, can be the source of alternative datasets. For example, a media streaming company is also likely to collect secondary data, which is a by-product of users listening to music or watching TV.

B. Compare traditional model building approaches and machine learning.

Traditionally, when developing a trading strategy or indeed any sort of forecast, we try to find a hypothesis first. We can then validate (or indeed invalidate) our hypothesis using statistical analysis. The idea of machine learning techniques is that we don't need to know the form of the relationship between variables beforehand; instead machine learning can help us model the function, even if it is highly nonlinear. The difficulty, however, is that we might end up finding patterns in what is essentially noise. Furthermore, the nature of financial problems is not stable: financial time series are non-stationary and markets experience changing regimes. We can instead apply machine learning to pre-processing and cleaning the dataset (such as classifying parts of the dataset, applying techniques such as sentiment analysis or topic identification of text), not purely focusing on forecasting the time series of the asset itself.

C. Discuss how big data and alternative data can be used to improve economic forecasts and "nowcasts."

We might seek to use alternative datasets to improve longer-run economic forecasts or there can be alternative datasets which directly give us a forecast, which may be useful for broader-based investing. We can also trade on a short-term basis, around economic data releases, if we can generate reasonable estimates in real-time, to improve our forecasts for monthly economic data releases (such as monthly change in nonfarm payrolls) that can be generated throughout the month to aid our trading strategy.

D. Describe how case studies show that alternative data is related to the following types of macro data: US Treasury yields, Implied volatility in the foreign exchange market, and investor anxiety

- Cuemacro's Federal Reserve sentiment index attempts to quantify the communications in a systematic manner. The raw input data consists of text extracted from Federal Reserve communications (speeches, statements and minutes), which is of a relatively small size. Sentiment scores. which are generated using natural language processing, are aggregated into a time series to represent an index which tracks the overall sentiment of the Federal Reserve over time to give a representative view of Fed sentiment over recent weeks. There is a strong relationship between this series and moves in FX or bond markets 10Y yields
- Articles from the BN newswire, whilst typically consumed by users of the Bloomberg Terminal, are also available in a machine-readable form. After filtering the dataset for the assets we are trading, the volume of news can also be useful in itself. There is significant positive correlation between implied volatility in various crosses and news volume related to those currencies. This suggests that we can use news volume as an input to model implied volatility.
- Investopedia is a financial education website. The principle behind its Investor Anxiety Index (IAI) is to track search terms, focussing on those terms related to investor anxiety such as 'short selling',

made by users which results in Investopedia page views. The index shows an intuitive relationship against rises and falls of VIX option prices, which is often referred to as the 'Wall Street Fear Gauge'. Returns for an active trading rule for S&P 500 futures based upon IAI are higher than one using VIX.

Topic 2. Data Mining & Machine Learning: Introduction

Reading 2.1 *An Introduction to Statistical Learning (Ch. 1 & 2)*

James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). An Introduction to Statistical Learning: with applications in R. New York, NY: Springer. Chapters 1, 2.1 & 2.2.

Ch. 1. Introduction

Ch. 2. Statistical Learning

Keywords

Statistical learning (p. 1)
Statistical learning refers to a vast set of tools for understanding data. Supervised statistical learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs. With unsupervised statistical learning, there are inputs but no supervising output; nevertheless we can learn relationships and structure from such data.

Quantitative variables (p. 28)
Quantitative variables take on numerical values, such as a person's age, height, or income, the value of a house, and the price of a stock.

Qualitative response (p. 28)
Qualitative variables take on values in one of K different classes, or categories.

Binary response (p.28)
A binary response involves a two-class qualitative, response.

Regression (p. 28)
Regression problems involve a quantitative response.

Classification problems (p.28)
Classification problems involve qualitative response.

Semi-supervised learning (p. 28)
Semi-supervised learning uses a statistical learning method that can incorporate the observations for which response measurements are available as well as the the other observations for which they are not.

Predictors (p. 29)
Predictors are variables, other than the response variable, that are measured of each observation.

Mean squared error (MSE) (p. 29)
MSE measures the average squared difference of the predicted and true responses. It quantifies the extent to which the predicted response value for a given observation is close to the true response value for that observation.

Training MSE (p. 30)
The training data are that used to fit the model.

Test data (p. 30)
Test data are previously unseen data, not used to train the model.

Test MSE (p. 30)
Test MSE is the accuracy of the predictions that we obtain when we apply our method to previously unseen test data.

Degrees of freedom (p. 32)
The degrees of freedom is a quantity that summarizes the flexibility of a function, or curve.

Cross validation (p. 33)
Cross-validation is a method for estimating test MSE using the training data.

Expected test MSE (p. 34)
The expected test MSE refers to the average test MSE that we would obtain if we repeatedly estimated f using a large number of training sets, and tested each over all possible values in the test set.

Bias (p. 35)
Bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.

Bias-variance trade-off (p. 36)
Good test set performance of a statistical learning method requires low variance (the amount by which $\hat{f}$ would change if we estimated it using a different training data set) as well as low squared bias.

Indicator variable (p. 37)
An indicator variable takes on a value of 1 or 0, and measures, for example, if an observation was classified correctly by our classification method; otherwise it was misclassified.

Error rate (p.37)
The proportion of mistakes that are made if we apply our estimate $\hat{f}$ to the observations.

Training error (p. 37)
The error computed based on the data that was used to train our classifier.

Test error (p.37)
The error computed based on a set of test observations that were not used in training.

Bayes classifier (p. 37)
A very simple classifier that assigns each observation to the most likely class, given its predictor values.

Conditional probability (p. 37)
The probability that Y belongs to class j, given the observed predictor vector values x: $\Pr(Y = j|X = x)$.

Bayes decision boundary (p. 38)
The line in the multidimensional spare consisting of the predictor variables which represents the points where the conditional probability is exactly 50% of predicting either class one or two in a two-class Bayes classier.

Bayes error rate (p. 38)
The Bayes classifier produces the lowest possible test error rate, called the Bayes error rate.

K-nearest neighbors (p. 39)
Given a positive integer K and a test observation x, the KNN classifier first identifies the K points in the training data that are closest to x, then estimates the conditional probability for class j as the fraction of those points whose response values equal j. Finally, KNN applies Bayes rule and classifies the test observation x to the class with the largest probability.

2.1.1 Organization and resources of the book (Ch. 1)

Statistical learning should not be viewed as a series of black boxes. No single approach will perform well in all possible applications. Without understanding all of the cogs inside the box, or the interaction between those cogs, it is impossible to select the best box. While it is important to know what job is performed by each cog – the model, intuition, assumptions, and trade-offs behind each of the methods – it is not necessary to have the skills to construct the machine inside the box.

2.1.2 Statistical learning (Ch. 2.1)

A. Explain why we estimate a function with data, including the role of input and output variables and their synonyms, as well as error terms (reducible and irreducible), expected values and variance.

There are two main reasons:

1. Prediction. When inputs X are readily available and the error term averages to zero, but the output Y cannot be easily obtained, we can predict Y using our estimate of the function $\hat{Y} = \hat{f}(X)$. The accuracy of $\hat{Y}$ as a prediction for Y depends on the reducible error and the irreducible error. We can potentially improve the accuracy of $\hat{f}$ by using the most appropriate statistical learning technique to reduce the error. estimate f. However, if Y is also a function of an error term ϵ (which contains unmeasured variables useful in predicting Y), then this is known as the irreducible error, because no matter how well we estimate f, we cannot reduce the error. We can show that the expected (or average) squared different between the predicted and actual value is:

$$\begin{aligned} E[(Y - \hat{Y})^2] &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= [f(X) - \hat{f}(X)]^2 + Var(\epsilon) \\ &= \text{Reducible Error} + \text{Irreducible Error} \end{aligned}$$

   where $Var(\epsilon)$ represents the variance associated with the error term.

2. Inference. We instead want to understand the relationship between X and Y:

- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

B. Compare and contrast parametric and non-parametric learning methods.

Parametric methods involve a two-step model-based approach:

1. First, we make an assumption about the functional form, or shape, of f (e.g. that f is linear in X)
2. After a model has been selected, we need a procedure that uses the training data to fit or train the model (e.g. finding the linear coefficients by the ordinary least squares approach).

Non-parametric methods do not make explicit assumptions about the functional form of f. Instead they seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly. By avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a wider range of possible shapes for f.

C. Describe the trade-offs between prediction accuracy, flexibility and model interpretability, including the role of overfitting.

In general, as the flexibility of a method increases, its interpretability decreases. Surprisingly, even if we are only interested in prediction, more accurate predictions are often obtained by a less flexible method, due to the potential for overfitting in highly flexible methods.

D. Determine when a supervised learning model is preferable to unsupervised or semi-supervised learning models.

When this is an associated response measurement for each observation of the predictor measurement(s), supervised learning can fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference). In the somewhat more challenging situation in which for every observation, a vector of measurements but no associated response are observed, then unsupervised learning is used to understand the relationships between the variables or observations. When response measurements are only available for some of the observations, semi-supervised learnings methods are used which incorporate observations for which response measurements are available as well observations for which they are not.

E. Explain how the appropriateness of regression problems relative to classification problems may be related to whether responses are quantitative or qualitative.

Variables can be characterized as either quantitative take on numerical values, or qualitative (also known as categorical) which take on values in one of K different classes, or categories. Problems with a quantitative response tend to be referred to as regression problems, while those involving a qualitative response as classification problems.

2.1.3 Assessing Model Accuracy (Ch. 2.2)

A. Recognize and explain the equation for mean squared error.

In the regression setting, the most commonly-used measure of model accuracy is the mean squared error: $MSE = \frac{1}{n}\sum_i^n (y_i - \hat{f}(x_i)^2$, where $\hat{f}(x_i)$ is the prediction for the ith observation. The MSE will be small if the predicted responses are very close to the true responses, and will be large if for some of the observations, the predicted and true responses differ substantially.

B. Explain the goal of measuring the quality of fit by minimizing training and test mean square errors (MSEs) and the implications of different levels of flexibility (degrees of freedom) for both training and test MSEs.

We are really interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data. The problem is that many statistical methods specifically estimate coefficients so as to minimize the training set MSE, but the test MSE is often much larger. The degrees of freedom is a quantity that summarizes the flexibility of a function. As model flexibility increases, training MSE will decrease, but the test MSE may not.

C. Explain the purpose of cross validation.

In practice, one can usually compute the training MSE with relative ease, but estimating test MSE is considerably more difficult because usually no test data are available. Cross-validation, is a method for estimating test MSE using the training data.

D. Explain the bias-variance trade-off with an MSE decomposition into three fundamental quantities.

Expected test MSE can always be decomposed into the sum of three fundamental quantities: variance of $\hat{f}(x)$, the squared bias of $\hat{f}(x)$ and the variance of the error term ϵ.

$$E[y - \hat{f}(x)]^2 = \text{Var}[\hat{f}(x)] + \text{Bias}[\hat{f}(x)]^2 + \text{Var}[\epsilon]$$

Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set. Bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model. The challenge, or trade-off, lies in finding a method for which both the variance and the squared bias are low.

E. Explain the salient features of a simple Bayes classifier (for two classes) including the Bayes decision boundary and Bayes error rate.

The simple Bayes classifier assigns each observation to the most likely class, given its predictor values. In a two-class problem, this corresponds to predicting class one if $Pr(Y = 1|X = x)) > 0.5$, and class two otherwise. The Bayes decision boundary, which is plotted in the multi-dimensional space consisting of the predictors X, represents the set of points for which the probabilities are 50%. The Bayes classifier produces the lowest possible test error rate, called the Bayes error rate. The Bayes error rate is analogous to the irreducible error,

F. Explain how the K-nearest neighbors classifier is related to the Bayes classifier and how the choice of K impacts results.

The KNN classifier first identifies the K points in the training data that are closest to x. It then estimates the conditional probability for class j as the fraction of points whose response values equal j. Finally, KNN applies Bayes rule and classifies the test observation to the class with the largest probability. As K grows, the method becomes less flexible and produces a decision boundary that is close to linear. This corresponds to a low-variance but high-bias classifier.

Reading 2.2 *Using Neural Networks to Recognize Handwritten Digits*

Nielsen, M. A. (2015). Using Neural Networks to Recognize Handwritten Digits. In Neural Networks and Deep Learning, Determination Press.

Retrieved from http://neuralnetworksanddeeplearning.com/chap1.html

Keywords

Perceptron neurons (p. 3)
A perceptron takes several binary inputs and produces a single binary output.

Weights (p. 4)
Weights are real numbers expressing the importance of the respective inputs to the output.

Threshold value (p. 4)
The threshold is a real number which is a parameter of the neuron, that determines how much evidence, the weighted sum of the input factors, is needed to make a decision.

Bias (p. 6)
A measure of how easy it is to get the perceptron to fire.

Layer (p. 6)
A column of perceptrons. A many-layer network of perceptrons can engage in sophisticated decision making.

NAND gate (p. 7)
An elementary logical function. A NAND gate is universal for computation, that is, we can build any computation up out of NAND gates.

Input layer (p. 9)
The input layer encodes the inputs, and are special units which are simply defined to output the desired values.

Learning algorithms (p. 10)
Learning algorithms automatically tune the weights and biases of a network of artificial neurons.

Sigmoid neuron (p. 11)
Sigmoid neurons are similar to perceptrons, but modified so that small changes in their weights and bias cause only a small change in their output.

Sigmoid function (p. 12)
The mathemetical function defined by $\sigma(z) \equiv \frac{1}{1+e^{-z}}$

Activation function (p. 14)
Neurons can output other activation functions, with different shapes than the sigmoid function.

Input neurons (p. 16)
The leftmost layer in a neural network is called the input layer, and the neurons within the layer are called input neurons.

Output neurons (p. 16)
The rightmost or output layer in a neural network contains the output neurons.

Hidden layer (p. 16)
The middle layer is called a hidden layer, since the neurons in this layer are neither inputs nor outputs. Some networks have multiple hidden layers.

Multilayer perceptrons (p. 16)
Multiple layer networks are sometimes called multilayer perceptrons (despite being made up of sigmoid neurons, not perceptrons).

Feedforward neural networks (p. 17)
Neural networks where the output from one layer is used as input to the next layer. This means there are no loops in the network – information is always fed forward, never fed back.

Recurrent networks (p. 17)
Mmodels of artificial neural networks in which feedback loops are possible. The idea in these models is to have neurons which fire for some limited duration of time, before becoming quiescent. That firing can stimulate other neurons, which may fire a little while later, also for a limited duration. That causes still more neurons to fire, and so over time we get a cascade of neurons firing.

Cost function (p. 24)
To quantify how well we are achieving the goal of finding weights and biases so that the output from the network approximates the output for all training inputs.

Loss function (p. 24)
Cost function is sometimes referred to as a loss function.

Objective function (p. 24)
Cost function is sometimes referred to as a objective function

Quadratic cost function (p. 25)
$C(w,b) \equiv \frac{1}{2n} \sum_x ||y(a) - a||^2$, where w denotes the collection of all weights in the network, b all the biases, n is the total number of training inputs, a is the vector of outputs from the network when x is input, and the sum is over all training inputs, x.

Mean Squared error (MSE) (p. 25)
The quadratic cost function is also sometimes known as the mean squared error or just MSE.

Gradient descent algorithm (p. 25)
Given a function of many variables, gradient descent is a technique which can be used to find values of the variables that minimize tha function.

Learning rate (p. 30)
A small, positive parameter chosen for the gradient descent algorithm so that its approximation is good.

Stochastic gradient descent (p. 34)
Stochastic gradient descent is a commonly used and powerful technique for learning in neural networks. It samples small mini-batches rather than apply gradient descent to the full batch. It speeds up estimates compared to an exact computation of the gradient.

Mini-batch (p. 34)
A small random sample from the full batch.

Epoch (p. 35)
After randomly chosen mini-batches to train with have exhausted the training inputs, that is said to complete an epoch of training.

Validation set (p. 37)
A subset of training data used in figuring out how to set certain hyper-parameters of the neural network.

Hyper-parameters (p. 37)
Things like the learning rate, and so on, which aren't directly selected by our learning algorithm.

Deep neural networks (p.55)
Deep nets build up a complex hierarchy of concepts, by breaking down a very complicated question into very simple questions answerable at the level of single pixels. It does this through a series of many layers, with early layers answering very simple and specific questions about the input image, and later layers building up a hierarchy of ever more complex and abstract concepts.

2.2.1 Motivation for using neural nets to recognize handwritten digits

A. Explain the use of a training set as an alternative to a rules-based program to recognize digits.

Simple intuitions about how we recognize shapes turn out to be not so simple to express algorithmically. When you try to make such rules precise, you quickly get lost in a morass of exceptions and caveats and special cases. The neural network uses a large number of handwritten digits, known as training examples, automatically infer rules for recognizing handwritten digits. Furthermore, by increasing the number of training examples, the network can learn more about handwriting, and so improve its accuracy.

2.2.2 Perceptron neurons

A. Calculate the output of a perceptron neuron.

$$\text{output} = 0, \text{ if } \sum_j w_j x_j \leq \text{threshold}$$
$$1, \text{ if } \sum_j w_j > \text{threshold}$$

B. Describe the intuition of a perceptron as a decision-making device.

A device that makes decisions by weighing up evidence.

C. Describe a perceptron as a NAND gate and what it implies for perceptron networks with respect to computing logical functions.

We can use perceptrons to compute simple logical functions like a NAND gate. In fact, we can use networks of perceptrons to compute any logical function at all.

D. Explain how perceptron neurons are more than new types of NAND gates.

We can devise learning algorithms which can automatically tune the weights and biases of a network of artificial neurons. Instead of explicitly laying out a circuit of NAND and other gates, our neural networks can simply learn to solve problems, sometimes problems where it would be extremely difficult to directly design a conventional circuit.

2.2.3 Sigmoid neurons

A. Recognize a limitation of perceptron neurons that can be overcome by sigmoid neurons.

A small change in the weights or bias of any single perceptron in the network can sometimes cause the output of that perceptron to completely flip, say from 0 to 1. That flip may then cause the behaviour of the rest of the network to completely change in some very complicated way. That makes it difficult to see how to gradually modify the weights and biases so that the network gets closer to the desired behaviour. We can overcome this problem by introducing new type of artificial neuron called a sigmoid neuron

B. Recognize and differentiate perceptron neurons from sigmoid neurons.

Sigmoid neurons are similar to perceptrons, but modified so that small changes in their weights and bias cause only a small change in their output.

C. Calculate the output of a sigmoid function which is also referred to as a logistic function.

The output is not 0 or 1, but a smooth function $\sigma(w \cdot x + b) = \frac{1}{1+\exp(-\sum_j w_j x_j - b)}$

D. Explain why the smoothness of the sigmoid function is important.

The smoothness of the sigmoid function means that small changes in the weights and bias will produce a small change in the output from the neuron. This allows a network of sigmoid neurons to learn.

2.2.4 The architecture of neural networks

A. Identify components of a simple network with appropriate terminology.

The leftmost layer in a simple network is called the input layer, and the neurons within the layer are called input neurons. The rightmost or output layer contains the output neurons, or, as in this case, a single output neuron. The middle layer is called a hidden layer, since the neurons in this layer are neither inputs nor outputs. Some networks have multiple hidden layers.

B. Describe the central feature of a feed forward network.

Neural networks where the output from one layer is used as input to the next layer. There are no loops in the network – information is always fed forward, never fed back.

C. Compare and contrast feedforward networks with recurrent networks.

The idea in these models is to have neurons which fire for some limited duration of time, before becoming quiescent. That firing can stimulate other neurons, which may fire a little while later, also for a limited duration. That causes still more neurons to fire, and so over time we get a cascade of neurons firing. Loops dontt cause problems in such a model, since a neuronts output only affects its input at some later time, not instantaneously.

2.2.5 A simple network to classify handwritten digits

A. Argue for a natural order for solving the two problems of segmenting digits and classifying digits.

First we break an image containing many digits into a sequence of separate images, each containing a single digit. Once the image has been segmented, the program then needs to classify each individual digit

B. Calculate the required input neurons for classifying an individual digit in an image of a certain size in pixels.

If training data for the network will consist of many 28 by 28 pixel images of scanned handwritten digits, the input layer contains $784 = 28 \times 28$ neurons.

C. Explain the choice to use ten output neurons instead of four for classifying an individual digit.

If we had 4 outputs, then the first output neuron would be trying to decide what the most significant bit of the digit was. And there's no easy way to relate that most significant bit to simple shapes in the image.

2.2.6 Learning with gradient descent

A. Recognize a quadratic cost function of weights and biases and alternative terminology for the cost function.

The quadratic cost function is also sometimes known as the mean squared error or just MSE: $C(w,b) \equiv \frac{1}{2n}\sum_x ||y(a) - a(b,w,x)||^2$, where w denotes the collection of all weights in the network, b all the biases, n is the total number of training inputs, a is the vector of outputs from the network when x is input, and the sum is over all training inputs, x. The output a depends on x, w and b.

B. Explain why minimizing a quadratic cost function is preferable to different types of cost functions.

It is easy to use a smooth cost function like the quadratic cost to figure out how to make small changes in the weights and biases so as to get an improvement in the cost.

C. Recognize an equation for an update rule that defines the gradient descent algorithm and explain the purposed of each component in the equation.

The update rule for the variables (weights) v is: $v \to v' = v - \eta\nabla C$ where ∇C is the gradient of C and η is a small, positive parameter (known as the learning rate).

D. Explain how quickly stochastic gradient descent can speed up learning given a training set size n and a mini-batch size, m.

For example, if we have a training set of size $n = 60{,}000$, and choose a mini-batch size of (say) $m = 10$, this means we'll get a factor of 6,000 speedup in estimating the gradient.

2.2.7 Implementing a network to classify digits

A. Understand the role of hyper-parameters and their impact output for each epoch.

Hyper-parameters of the neural network are things like the learning rate, and so on, which aren't directly selected by our learning algorithm. If we choose our hyper-parameters poorly, we can get bad results. Even though we initially made a poor choice of hyper-parameters, we at least got enough information from the performance of the network over each epoch to help us improve our choice of hyper-parameters.

2.2.8 Why deep learning matters

A. Describe deep learning in terms of neural networks and their performance relative to networks that are not based on deep learning methods

Compared to shallow neural networks, i.e. networks with just a single hidden layer, deep nets build up a complex hierarchy of concepts, by breaking down a very complicated question into very simple questions answerable at the level of single pixels. It does this through a series of many layers, with early layers answering very simple and specific questions about the input image, and later layers building up a hierarchy of ever more complex and abstract concepts. Networks with this kind of many-layer structure – two or more hidden layers – are called deep neural networks.

Topic 3. Data Mining & Machine Learning: Regression, LASSO, Predictive Models, Times Series & Tree Models

Reading 3.1 *Data Science for Business (Ch. 3, 4, 5, & 9)*

Provost, F. and T. Fawcett. (2019). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapters 3, 4, 5, & 9.

Ch. 3. Introduction to Predictive Modeling: From Correlation to Supervised Segmentation.

Ch. 4. Fitting a Model to Data.

Ch. 5. Overfitting and Its Avoidance.

Ch. 9. Evidence and Probabilities.

Keywords

Information (p. 43)
Information is a quantity that reduces uncertainty.

Tree induction (p. 44)
Tree induction incorporates the idea of supervised segmentation in repeatedly selecting informative attributes.

Predictive model (p. 45)
A formula for estimating the unknown value of interest: the targe. It could be mathematical, a logical statement such as a rule, or a hybrid of the two.

Prediction (p. 45)
To estimate an unknown value.

Descriptive modeling (p. 46)
The primary purpose of descriptive modeling is not to estimate a value but instead to gain insight into the underlying phenomenon or process.

Target variable (p. 46)
The target variable, whose values are to be predicted, is commonly called the dependent variable in statistics

Attributes or features (p. 46)
The features (table columns) have many different names as well. Statisticians speak of independent variables or predictors as the attributes supplied as input. In operations research you may also hear explanatory variable.

Model induction (p. 47)
The creation of models from data is known as model induction. Induction is a term from philosophy that refers to generalizing from specific cases to general rules (or laws, or truths). the procedure that creates the model from the data is called the induction algorithm or learner.

Deduction (p. 47)
Deduction starts with general rules and specific facts, and creates other specific facts from them.

Training data (p. 47)
The input data for the induction algorithm, used for inducing the model, are called the training data.

Labeled data (p. 47)
Data are called labeled because the value for the target variable (the label) is known.

Supervised segmentation (p. 48)
An intuitive way of thinking about extracting patterns from data in a supervised manner is to try to segment the population into subgroups that have different values for the target variable (and within the subgroup the instances have similar values for the target variable). If the segmentation is done using values of variables that will be known when the target is not, then these segments can be used to predict the value of the target variable.

Information gain (p. 51)
Information gain, based on a purity measure called entropy, is the most common splitting criterion for evaluating how well each attribute splits a set of examples into segments, with respect to a chosen target variable.

Entropy (p. 51)
Entropy is a measure of disorder that can be applied to a set. Disorder corresponds to how mixed (impure) the segment is with respect to these properties of interest such as values of the target variable.

Parent set (p. 52)
The original set of examples.

Child set (p. 52)
Suppose the attribute we split on has k different values. The result of splitting the original examples (the parent set) on the attribute values into the k children sets.

Variance (p. 56)
A natural measure of impurity for numeric (not class) values. If the set has all the same values for the numeric target variable, then the set is pure and the variance is zero. If the numeric target values in the set are very different, then the set will have high variance.

Entropy graph/chart (p. 58)
A two-dimensional description of the entire dataset's entropy as it is divided in various ways by different attributes.

Classification tree (p. 63)
A segmentation of the data to take the form of a "tree," upside down with the root at the top. Each interior node in the tree contains a test of an attribute, with each branch from the node representing a distinct value of the attribute. Following the branches from the root node down (in the direction of the arrows), each path eventually terminates at a terminal node, or leaf. The tree creates a segmentation of the data.

Decision nodes (p. 63)
The classification tree is made up of nodes, interior nodes and terminal nodes, and branches emanating from the interior nodes. Each interior node in the tree contains a test of an attribute, each branch from the node representing a distinct value of the attribute, and each path eventually terminates at a terminal node, or leaf.

Probability estimation tree (p. 64)
In some applications, we want predict the probability of membership in the class, rather than the class itself. In this case, the leaves of the probability estimation tree would contain these probabilities.

Tree induction (p. 64)
A popular method to induce a supervised segmentation from a dataset. Tree induction takes a divide-and-conquer approach, starting with the whole dataset and applying variable selection to try to create the "purest" subgroups possible using the attributes. We thenn simply take each data subset and recursively apply attribute selection to find the best attribute to partition it.

Decision surface or boundary (p. 69)
Each internal (decision) node corresponds to a split of the instance space, while each leaf node corresponds to an unsplit region of the space (a segment of the population). The lines separating the regions are known as decision lines (in two dimensions) or more generally decision surfaces (in higher dimensions), or decision boundaries.

Frequency-based estimation of class membership probability (p. 72)
We can use instance counts at each leaf to compute a class probability estimate. This is called a frequency-based estimate.

Laplace correction (p. 73)
A "smoothed" version of the frequency-based estimate, the purpose of which is to moderate the influence of leaves with only a few instances. The Laplace correction equation for binary class probability estimation becomes: $p(c) = \frac{n+1}{n+m+2}$, where n is the number of examples in the leaf belonging to class c, and m is the number of examples not belonging to class c.

Linear classifier (p. 85)
Separate the instances by class with a straight line in the (two-dimensional) instance space. It is essentially a weighted sum of the values for the various attributes.

Linear discriminant (p. 86)
Discriminates between classes, and the function of the decision boundary is a linear combination – a weighted sum – of the attributes.

Margin (p. 92)
The distance between the parallel lines of a bar between the classes.

Support vector machine (p. 92)
SVM's are linear discriminants that separate the classes, using an objection function to fit the data of choosing the fattest bar between the classes. The SVM's objective function incorporates the idea that

a wider bar is better. Then once the widest bar is found, the linear discriminant will be the center line through the bar. The distance between the dashed parallel lines is called the margin around the linear discriminant, and thus the objective is to maximize the margin.

Hinge-loss (p. 94)
Error function where the penalty for a misclassified point is proportional to the distance from the decision boundary.

Zero-one loss (p. 95)
Loss function assigns a loss of zero for a correct decision and one for an incorrect decision.

Squared error (p. 95)
Loss is proportional to the square of the distance from the boundary. Squared error loss usually is used for numeric value prediction (regression), rather than classification. The squaring of the error has the effect of greatly penalizing predictions that are grossly wrong.

Odds (p. 98)
The odds of an event is the ratio of the probability of the event occurring to the probability of the event not occurring.

Log-odds (p. 99)
Logarithm of the odds. For any number in the range 0 to ∞, its log will be between $-\infty$ to ∞.

Logistic function (p. 101)
Logistic regression's estimate of class probability as a function of $f(x)$ – the distance from the separating boundary: $p(x) = \frac{1}{1+e^{-f(x)}}$

Generalization (p. 112)
Generalization is the property of a model or modeling process, whereby the model applies to data that were not used to build the model.

Overfitting (p. 113)
When the model does not generalize at all beyond the data that were used to build it. It is tailored, or "fit,"perfectly to the training data. In fact, it is "overfit." Overfitting is the tendency of data mining procedures to tailor models to the training data, at the expense of generalization to previously unseen data points.

Fitting graph (p. 113)
A fitting graph shows the accuracy of a model as a function of complexity. Each point on a curve represents an accuracy estimation of a model with a specified complexity (as indicated on the horizontal axis).

Holdout data (p. 113)
Some data for which we know the value of the target variable, but which will not be used to build the model. Creating holdout data is like creating a "lab test" of generalization performance: we estimate the generalization performance by comparing the predicted values with the hidden true values.

Base rate (p. 115)
The performance of a classifier that always selects the majority class. A corresponding baseline for a regression model is a simple model that always predicts the mean or median value of the target variable.

Sweet spot (p. 117)
The complexity in a fitting graph at which holdout accuracy declines as the tree grows past its "sweet spot".

Cross-validation (p. 126)
Cross-validation is a more sophisticated holdout training and testing procedure, that generates statistics on the estimated performance, such as the mean and variance. The estimates are computed over all the data by performing multiple splits and systematically swapping out samples for testing.

Folds (p. 127)
Cross-validation begins by splitting a labeled dataset into k partitions called folds. Cross-validation then iterates training and testing k times. In each iteration of the cross-validation, a different fold is chosen as the test data. In this iteration, the other folds are combined to form the training data.

Learning curve (p. 131)
A plot of the generalization performance against the amount of training data is called a learning curve.

Sub-training set (p. 134)
We can take the training set and split it again into a training subset and a testing subset. Then we can build models on this training subset and pick the best model based on this testing subset. Let's call the former the sub-training set for clarity.

Validation set (p. 134)
The training set is split again into a training subset and a testing subset, then we can build models on this training subset and pick the best model based on this testing subset. The latter is called the validation set for clarity.

Nested holdout testing (p. 134)
We can take the training set and split it again into a training subset and a testing subset (called the validation set for clarity). Then we can build models on this training subset and pick the best model based on the validation set. The validation set is separate from the final test set, on which we are never going to make any modeling decisions. This procedure is often called nested holdout testing.

Nested cross validation (p. 135)
Say we would like to do cross-validation to assess the generalization accuracy of a new modeling technique, which has an adjustable complexity parameter C, but we do not know how to set it. Before building the model for each fold, we first run another entire cross-validation experiment on just that fold's training set to find the value of C estimated to give the best accuracy. The result of that experiment is used only to set the value of C to build the actual model for that fold of the cross-validation. Then we build another model using the entire training fold, using that value for C, and test on the corresponding test fold.

Sequential forward selection (p. 135)
Sequential forward selection (SFS) of features uses a nested holdout procedure to first pick the best individual feature, by looking at all models built using just one feature. After choosing a first feature, SFS tests all models that add a second feature to this first chosen feature. The best pair is then selected. Next the same procedure is done for three, then four, and so on. When adding a feature does not improve classification accuracy on the validation data, the SFS process stops.

Sequential backward elimination (p. 135)
Sequential backward elimination is similar to sequential forward selection of features, but works in reverse. It starts with all features and discards features one at a time. It continues to discard features as long as there is no performance loss.

Independent events (p. 236)
If two events are independent, then knowing about one of them tells you nothing about the likelihood of the other.

Joint probability using conditional probability (p. 237)
The (joint) probability of A and B is the probability of A times the (conditional) probability of B given A: $p(AB) = p(A) \cdot p(B|A)$

Bayes' Rule (p. 237)
$p(B|A) = \frac{p(A|B) \cdot p(B)}{p(A)}$

Posterior probability (p. 238)
The probability that the target variable C takes on the class of interest c after taking the evidence E (the vector of feature values) into account

Prior (p. 238)
The probability we would assign to the class before seeing any evidence. In Bayesian reasoning generally, this could come from several places. It could be (i) a "subjective" prior, meaning that it is the belief of a particular decision maker based on all her knowledge, experience, and opinions; (ii) a "prior" belief based on some previous application(s) of Bayes' Rule with other evidence, or (iii) an unconditional probability inferred from data.

Likelihood (p. 240)
The joint probability of the evidence, viewed as a function of the parameters with the data held fix.

Conditional independence (p. 241)
If we A and B are conditionally independent given C, we can now combine the probabilities much more easily: $p(AB|C) = p(A|C) \cdot p(B|C)$

Naive Bayes Classifier (p. 242)
Classifies a new example by estimating the probability that the example belongs to each class and reports the class with highest probability. It assumes the attributes are conditionally independent, given the class: $p(c|E) = \frac{p(e_1|c) \cdot p(e_2|c) ... p(ek|c) \cdot p(c)}{p(E)}$

Generative Model (p. 244)
Generative models, unlike discriminative methods which try directly to discriminate different targets by minimizing loss or entropy, turns the question around and asks: "How do different targets generate feature values?" They attempt to model how the data were generated, by apply Bayes' Rule to answer the question: "Which class most likely generated this example?"

Lift (p. 244)
Lift is the amount by which a classifier concentrates the positive examples above the negative examples. Lift measures how much more prevalent the positive class is in the selected subpopulation over the prevalence in the population as a whole.

Naive-Naive Bayes (p. 245)
With a slight modification, we can adapt our Naive Bayes equation to model the different lifts attributable to the different pieces of evidence, along with a very straightforward way of combining them. The slight modification is to assume full feature independence, rather than the weaker assumption of conditional independence used for Naive Bayes. Starting at the prior probability, each piece of evidence – each feature e_i – raises or lowers the probability of the class by a factor equal to that piece of evidence's lift (which may be less than one).

3.1.1 Models, Induction and Prediction

A. Define information and tree induction.

Information is a quantity that reduces uncertainty. Tree induction is a widely used predictive modeling technique for finding informative attributes. It incorporates the idea of supervised segmentation in an elegant manner, repeatedly selecting informative attributes, and extracting tree structured models from a dataset.

B. Define prediction in the context of data science.

In data science, a predictive model is a formula for estimating the unknown value of interest: the target. The formula could be mathematical, or it could be a logical statement such as a rule. Often it is a hybrid of the two.

C. Compare and contrast predictive modeling with descriptive modeling.

Prediction more generally means to estimate an unknown value. In contrast, the primary purpose of descriptive modeling is not to estimate a value but instead to gain insight into the underlying phenomenon or process.

D. Define attributes or features.

An instance or example, representing a fact or a data point, is described by a set of attributes (fields, columns, variables, or features). An instance is also sometimes called a feature vector, because it can be represented as a fixed-length ordered collection (vector) of feature values.

E. Describe model induction.

The creation of models from data is known as model induction – generalizing general rules in a statistical sense from specific cases.

F. Compare and contrast induction with deduction.

Induction is a term from philosophy that refers to generalizing from specific cases to general rules. Our models are general rules in a statistical sense (they usually do not hold 100% of the time; often not nearly). In contrast, Deduction starts with general rules and specific facts, and creates other specific facts from them.

G. Define training data and labeled data.

The input data for the induction algorithm, used for inducing the model, are called the training data. They are called labeled data because the value for the target variable (the label) is known.

3.1.2 Supervised Segmentation

A. Describe supervised segmentation.

A supervised learning method, to select one more attributes/features/variables, to try to segment the population into subgroups that have different values for the target variable (and within the subgroup the instances have similar values for the target variable).

B. List the complications arising from selecting informative attributes.

1. Attributes rarely split a group perfectly.
2. Is it better to split off one single data into a pure subset, than another split that does not produce any pure subset, but reduces the impurity more broadly?
3. Not all attributes are binary; many attributes have three or more distinct values. How do we compare one attribute can split into two groups while another might split into three groups, or more.
4. Some attributes take on numeric values (continuous or integer). How should we think about creating supervised segmentations using numeric attributes?

C. Define entropy and information gain.

The most common splitting criterion is called information gain, and it is based on a purity measure called entropy: Entropy $= -p_1 \log(p_1) - p_2 \log(p_2) - ...$, where each p_i is the probability (the relative percentage) of property i within the set, ranging from $p_i = 1$ when all members of the set have property i, and $p_i = 0$ when no members of the set have property i.

D. Recognize and apply entropy with a set containing two distinct groups.

Since there are only two classes, we can substitute out one of the two probabilities $p_1 = 1 - p_2$

E. Recognize and apply entropy with maximum and minimum disorder.

Entropy measures the general disorder of the set, ranging from zero at minimum disorder (the set has members all with the same, single property) to one at maximal disorder (the properties are equally mixed).

F. Contrast parent set with child set.

Parent set is the original set of examples, while the result of splitting the original set on the attribute values (assuming the attribute we split on has k values) into the k children sets.

G. Calculate information gain for a child relative to a parent.

The information gain (IG) to measure how much an attribute improves (decreases) entropy is $IG(parent, children) = entropy(parent) - p(c_1) \times entropy(c_1) + p(c_2) \times entropy(c_2) + ...$. Note that the entropy for each child is weighted by the proportion of instances belonging to that child.

H. Discuss the issues with the numerical variables for supervised segmentation.

Numeric variables can be "discretized" by choosing a split point (or many split points) and then treating the result as a categorical attribute. We still are left with the question of how to choose the split point(s) for the numeric attribute. Conceptually, we can try all reasonable split points, and choose the one that gives the highest information gain.

I. Define variance and discuss its application to numeric variables for supervised segmentation.

Variance is a natural measure of impurity for numeric (target) values in the subsets.

J. Define an entropy graph/chart.

An entropy graph is a two-dimensional description of the entire dataset's entropy as it is divided in various ways by different attributes. On the x axis is the proportion of the dataset (0 to 1), and on the y axis is the entropy (also 0 to 1) of a given piece of the data. The amount of shading of an entropy corresponds to the total (weighted sum) entropy, with each bar corresponding to the entropy of one of the attribute's values, and the width of the bar corresponding to the prevalence of that value in the data.

K. Describe how an entropy chart can be used to select informative variable.

The amount of shaded area in each graph represents the amount of entropy in the dataset when it is divided by some chosen attribute. Our goal of having the lowest entropy corresponds to having as little shaded area as possible.

L. Define a classification tree, decision nodes, probability estimation tree, and tree induction.

Consider a segmentation of the data to take the form of an upside down "tree," with the root at the top. The tree is made up of nodes, interior nodes and terminal nodes, and branches emanating from the interior nodes. Each interior node in the tree contains a test of an attribute, with each branch from the node representing a distinct value of the attribute. These nonleaf nodes are often referred to as "decision nodes," because when descending through the tree, at each node one uses the values of the attribute to make a decision about which branch to follow. The tree creates a segmentation of the data: every data point will correspond to one and only one path in the tree, and thereby to one and only one leaf, which contains a classification for its segment. Such a tree is called a classification tree or more loosely a decision tree. Tree induction takes a divide-and-conquer approach, starting with the whole dataset and applying variable selection to try to create the "purest" subgroups possible using the attributes. Often we want to predict the probability of membership in the class, rather than the class itself. In this case, the leaves of the probability estimation tree would contain these probabilities rather than a simple class value.

3.1.3 Visualizing Segmentations

A. Define decision surface or decision boundaries.

Each internal (decision) node corresponds to a split of the instance space, while each leaf node corresponds to an unsplit region of the space (a segment of the population). The lines (in two dimensions) or more generally surfaces (in higher dimensions) separating the regions are known as decision boundaries.

B. Describe the relationship between the decision surface and the number of variables

For a problem of n variables, each node of a classification tree imposes an $(n-1)$- dimensional "hyper-plane"decision boundary on the instance space.

C. Define frequency-based estimation of class membership probability.

We can use instance counts at each leaf to compute a class probability estimate. This is called a frequency-based estimate of class membership probability.

D. Calculate probability at each node of a decision tree.

If a node or leaf contains n positive instances and m negative instances, the probability of any new instance being positive may be estimated as $n/(n+m)$.

E. Define frequency-based estimation of class membership probability.

see above.

F. Describe how Laplace correction is used to modify probability of a leaf node with few members.

To moderate the influence of leaves with only a few instances, the Laplace correction equation is a "smoothed" version of the frequency-based estimate, to modify class probability estimation

G. Calculate the value of the Laplace correction.

$p(c) = \frac{n+1}{n+m+2}$, where n is the number of examples in the leaf belonging to class c, and m is the number of examples not belonging to class c.

3.1.4 Classification via Mathematical Functions

A. Define a linear classifier.

A linear classifier separates the instances by class with a straight line bounday in the (two-dimensional) instance space.

B. Recognize and apply the equation of a straight line using slope and intercept.

$Y = \text{slope} \times X + \text{intercept}$

C. Define a linear discriminant.

The linear discrimant function discrimantes between the classes and is a weighted sum of the values for the various attributes.

D. Describe decision boundaries in 2-dimension, 3-dimension, and higher dimensions.

In two dimensions, the linear combination corresponds to a line. In three dimensions, the decision boundary is a plane, and in higher dimensions it is a hyperplane.

E. Interpret magnitude of a feature's weight in a general linear model.

These weights are often loosely interpreted as importance indicators of the features. Roughly, the larger the magnitude of a feature's weight, the more important that feature is for classifying the target.

F. Describe how linear discriminant functions can be used for scoring and ranking instances.

The output of the linear discriminant function itself, $f(x)$, gives an intuitively satisfying ranking of the instances by their (estimated) likelihood of belonging to the class of interest. Its value will be relatively small when x is near the boundary, and large when far in the positive direction.

G. Describe the objective function of the Support Vector Machine (SVM).

The SVM's objective function first fits the fattest bar between the classes, and incorporates the idea that a wider bar is better. Then once the widest bar is found, the linear discriminant will be the center line through the bar.

H. Describe the important ideas behind the Support Vector Machine (SVM).

1. The SVM's objective function first finds the bar between the classes. The linear discriminant will be the center line through the bar. The distance between the dashed parallel lines is called the margin around the linear discriminant, and thus the objective is to maximize the margin.
2. If the data are not perfectly linearly separable, the best fit is some balance between a fat margin and a low total error penalty. The penalty for a misclassified point is proportional to the distance from the decision boundary, so if possible the SVM will make only "small" errors.

I. Define margin for SVM

The margin is the distance between the parallel lines of the bar separating the classes. The objective of SVM is to maximize the margin.

J. Define the hinge-loss function and zero-one loss function, and squared error.

A loss function determines how much penalty should be assigned to an instance based on the error in the model's predicted value (e.g.its distance from the separation boundary).

- Hinge loss incurs no penalty for an example that is not on the wrong side of the margin. The hinge loss only becomes positive when an example is on the wrong side of the boundary and beyond the margin.
- Zero-one loss, as its name implies, assigns a loss of zero for a correct decision and one for an incorrect decision.
- Squared error specifies a loss proportional to the square of the distance from the boundary. Squared error loss usually is used for numeric value prediction (regression), rather than classification.

K. Describe the reason for not using squared loss function in classification problems.

Using squared error for classification also penalizes points far on the correct side of the decision boundary.

3.1.5 Regression via Mathematical Functions

A. Describe the major drawback of the least square regression.

Least squares regression minimizes squared error as its objective function which has a serious drawback in that it is very sensitive to the data: erroneous or otherwise outlying data points can severely skew the resultant linear function.

B. Define odds and log odds.

The odds of an event is the ratio of the probability of the event occurring to the probability of the event not occurring. Log-odds is the logarithm of the odds. For any number in the range 0 to ∞ its log will be between $-\infty$ to ∞.

C. List the important points of the logistic regression.

1. For probability estimation, logistic regression uses the same linear model as do our linear discriminants for classification and linear regression for estimating numeric target values.
2. The output of the logistic regression model is interpreted as the log-odds of class membership.
3. These log-odds can be translated directly into the probability of class membership.

D. Recognize and apply the logistic function.

Logistic regression's estimate of class probability is a function of $f(x)$ – the distance from the separating boundary: $p(x) = \frac{1}{1+e^{-f(x)}}$

E. Describe the shape of the logistic function.

The curve is called a "sigmoid" curve because of its "S" shape (as a function of the distance from the separating boundary), which squeezes the probabilities into their correct range (between zero and one).

F. Describe how objective function is formed in logistic regression.

The class probability estimates computes the "likelihood" that a particular labeled example belongs to the correct class, for a set of parameters w in the logistic regression model. The model (set of weights) that gives the highest sum is the model that gives the highest "likelihood" to the data – the "maximum likelihood" model. Maximizing this objective function "on average" gives the highest probabilities to the positive examples and the lowest probabilities to the negative examples.

G. Compare and contrast classification trees with linear classifiers.

1. A classification tree uses decision boundaries that are perpendicular to the instance-space axes, whereas the linear classifier can use decision boundaries of any direction or orientation. This is a direct consequence of the fact that classification trees select a single attribute at a time whereas linear classifiers use a weighted combination of all attributes.
2. A classification tree is a "piecewise" classifier that segments the instance space recursively when it has to, using a divide-and-conquer approach. In principle, a classification tree can cut up the instance space arbitrarily finely into very small regions. A linear classifier places a single decision surface through the entire space, hence is limited to a single division into two segments. This is a direct consequence of there being a single (linear) equation that uses all of the variables, and must fit the entire data space.

3.1.6 Overfitting and Its Avoidance

A. Define generalization, overfitting, fitting graph, holdout data, and base rate.

Generalization is the property of a model or modeling process, whereby the model applies to data that were not used to build the model. Overfitting is the tendency of data mining procedures to tailor models to the training data, at the expense of generalization to previously unseen data points. A fitting graph shows the accuracy of a model as a function of complexity. Holdout data are not be used to build the model, but for which we know the value of the target variable: we estimate the generalization performance of the model by comparing the predicted values with the hidden true values. Base rate is the performance of a classifier that always selects the majority class.

B. Apply the concept of fitting graph to find the optimal tree induction model.

A procedure that grows trees until the leaves are pure tends to overfit. The complexity of the tree lies in the number of nodes. We artificially limit the maximum size of each tree, as measured by the number of nodes it's allowed to have, indicated on the horizontal axis of the fitting graph. For each tree size we create a new tree from scratch, using the training data. We measure two values: its accuracy on the training set and its accuracy on the holdout (test) set. As the trees are allowed to get larger, the training- set accuracy continues to increase, but the holdout accuracy eventually declines.

C. Define sweet spot for a typical fitting graph.

The complexity at which holdout accuracy declines as the tree grows past its "sweet spot". Unfortunately, no one has come up with a procedure to determine this exact sweet spot theoretically, so we have to rely on empirically based techniques.

D. Apply the concept of overfitting in mathematical functions.

One way mathematical functions can become more complex is by adding more variables (more attributes). As you increase the dimensionality, you can perfectly fit larger and larger sets of arbitrary points in the training data.

E. Analyze overfitting for logistic regression and support vector machine.

As outliers are added to the training data, the fitted support vector machine line moves very little in response, but the fitted logistic regression line moves considerably. If a linear boundary exists, logistic regression will find it, even if this means moving the boundary to accommodate outliers. The SVM tends to be less sensitive to individual examples, as its training procedure incorporates complexity control.

F. Explain why overfitting is bad.

As a model gets more complex it is allowed to pick up harmful spurious correlations. These correlations are idiosyncracies of the specific training set used and do not represent characteristics of the population in general. The harm occurs when these spurious correlations produce incorrect generalizations in the model. This is what causes performance to decline when overfitting occurs.

G. Define cross-validation and folds.

Cross-validation is a more sophisticated holdout training and testing procedure, that computes statistics on the estimated performance, such as the mean and variance, over all the data by performing multiple splits and systematically swapping out samples for testing. It by splitting a labeled dataset into k partitions called folds. Cross-validation then iterates training and testing k times. In each iteration of the cross-validation, a different fold is chosen as the test data. In this iteration, the other folds are combined to form the training data.

H. Define a learning curve.

A plot of the generalization performance against the amount of training data is called a learning curve. All else being equal, the generalization performance of data-driven modeling generally improves as more training data become available, up to a point.

I. Compare and contrast a learning curve with a fitting graph.

A learning curve shows the generalization performance – the performance only on testing data, plotted against the amount of training data used. A fitting graph shows the generalization performance as well as the performance on the training data, but plotted against model complexity. Fitting graphs generally are shown for a fixed amount of training data.

J. Describe shape of learning curves for logistic regression and tree induction.

Learning curves usually are steep initially as the modeling procedure finds the most apparent regularities in the dataset. Then as the modeling procedure is allowed to train on larger and larger datasets, it finds more accurate models. However, the marginal advantage of having more data decreases, so the learning curve

becomes less steep. In some cases, the curve flattens out completely because the procedure can no longer improve accuracy even with more training data. For smaller data, tree induction will tend to overfit more; this leads logistic regression to perform better for smaller datasets. On the other hand, the flexibility of tree induction can be an advantage with larger training sets: the tree can represent substantially nonlinear relationships between the features and the target.

K. List strategies that can be used to avoid overfitting in tree induction.

1. to stop growing the tree before it gets too complex
2. to grow the tree until it is too large, then "prune" it back, reducing its size (and thereby its complexity).

L. Describe how minimum number of instances in a tree leaf can be used to limit tree size.

The idea behind this minimum-instance stopping criterion to limit tree size is that for predictive modeling, we essentially are using the data at the leaf to make a statistical estimate of the value of the target variable for future cases that would fall to that leaf. If we make predictions of the target based on a very small subset of data, we might expect them to be inaccurate. A nice propertyof controlling complexity in this way is that tree induction will automatically grow the tree branches that have a lot of data and cut short branches that have fewer data – thereby automatically adapting the model based on the data distribution.

M. Explain how hypothesis testing can be used to limit tree induction.

A hypothesis test tries to assess whether a difference in some statistic is not due simply to chance. In most cases, the hypothesis test is based on a "p-value," which gives a limit on the probability that the difference in statistic is due to chance. So, for stopping tree growth, an alternative to setting a fixed size for the leaves is to conduct a hypothesis test at every leaf to determine whether the observed difference in (say) information gain could have been due to chance. If the hypothesis test concludes that it was likely not due to chance, then the split is accepted and the tree growing continues.

N. Define sub-training set, validation set, and nested holdout testing.

Let's say we are saving the test set for a final assessment. We can take the training set and split it again into a training subset and a testing subset. Then we can build models on this training subset and pick the best model based on this testing subset. Let's call the former the sub-training set, and the latter the validation set for clarity. The validation set is separate from the final test set, on which we are never going to make any modeling decisions. This procedure is often called nested holdout testing.

O. Explain nested cross-validation.

Say we would like to do cross-validation to assess the generalization accuracy of a new modeling technique, which has an adjustable complexity parameter C, but we do not know how to set it. Before building the model for each fold, we first run another entire cross-validation experiment on just that fold's training set to find the value of C estimated to give the best accuracy. The result of that experiment is used only to set the value of C to build the actual model for that fold of the cross-validation. Then we build another model using the entire training fold, using that value for C, and test on the corresponding test fold.

P. Describe sequential forward selection and sequential backward elimination.

Sequential forward selection (SFS) of features uses a nested holdout procedure to first pick the best individual feature, by looking at all models built using just one feature. After choosing a first feature, SFS tests all models that add a second feature to this first chosen feature. The best pair is then selected. Next the same procedure is done for three, then four, and so on. When adding a feature does not improve classification accuracy on the validation data, the SFS process stops. There is a similar procedure called sequential backward elimination of features. It works by starting with all features and discarding features one at a time. It continues to discard features as long as there is no performance loss.

3.1.7 Evidence and Probabilities

A. Define independent events.

If two events are independent, then knowing about one of them tells you nothing about the likelihood of the other.

B. Calculate joint probability of two events

Let's say we have two events, A and B. If we know p(A) and p(B), can we say what is the joint probability that both A and B occur, p(AB)? There is one special case when we can: if events A and B are independent: $p(AB) = p(A) \cdot p(B)$.

C. Recognize and apply joint probability using conditional probability.

The (joint) probability of A and B is the probability of A times the (conditional) probability of B given A: $p(AB) = p(A) \cdot p(B|A)$

D. Calculate joint probability for independent and dependent events.

In the case of independent events: $p(AB) = p(A) \cdot p(B)$. The general formula for combining probabilities that takes care of dependencies between events is: $p(AB) = p(A) \cdot p(B|A)$

E. Explain the Bayes' Rule with the help of an example.

Let's consider H to be some hypothesis that we are interested in assessing the likelihood of, and E to be some evidence that we have observed. Bayes' Rule says that we can compute the probability of our hypothesis H given some evidence E by instead looking at the probability of the evidence given the hypothesis, as well as the unconditional probabilities of the hypothesis and the evidence: $p(H|E) = \frac{p(E|H) \cdot p(H)}{p(E)}$

F. Define posterior probability, prior, likelihood, and conditional independence.

With Bayes' Rule in the context of a classification problem: $p(C = c|E) = \frac{p(E|C=c) \cdot p(C=c)}{p(E)}$

- On the lefthand side is the quantity we would like to estimate. This is the posterior probability: $p(C = c|E)$

- $p(C = c|E)$ is the probability that the target variable C takes on the class of interest c after taking the evidence E (the vector of feature values) into account
- $p(C = c)$ is the "prior" probability of the class, i.e., the probability we would assign to the class before seeing any evidence.
- $p(E|C = c)$ is the likelihood of seeing the evidence E – the particular features of the example being classified – when the class $C = c$. This likelihood might be calculated from the data as the percentage of examples of class c that have feature vector E. $p(E)$ is the likelihood of the evidence: how common is the feature representation E among all examples

G. Explain the naive Bayes classifier.

It classifies a new example by estimating the probability that the example belongs to each class and reports the class with highest probability. It assumes the attributes are conditionally independent, given the class. In other words, in $p(e_1 \wedge e_2 \wedge ... \wedge e_k|c)$, each e_i is independent of every other e_j given the class c: $p(E|c) = p(e_1 \wedge e_2 \wedge ... \wedge e_k|c) = p(e_1|c) \cdot p(e_2|c)...p(e_k|c)$ Each of the $p(e_i|c)$ terms can be computed directly from the data, since now we simply need to count up the proportion of the time that we see individual feature e i in the instances of class c, rather than looking for an entire matching feature vector. $p(c|E) = \frac{p(e_1|c) \cdot p(e_2|c)...p(ek|c) \cdot p(c)}{p(E)}$

H. Explain why we do not need to calculate the denominator of the Bayes' rule for naive Bayes classifier.

Generally $p(E)$ never actually has to be calculated, for one of two reasons.

1. If we are interested in classification, what we mainly care about is: of the different possible classes c, for which one is $p(C|E)$ the greatest? In this case, E is the same for all, and we can just look to see which numerator is larger.
2. Where we would like the actual probability estimates, because the classes often are mutually exclusive and exhaustive, every instance will belong to one and only one class. With our independence assumption, we can compute $p(E)$ from the data with from terms that are each one is either the evidence "weight" of some particular individual piece of evidence, or a class prior.

I. List the advantages and disadvantages of the naive Bayes classifier.

Naive Bayes is a very simple classifier, yet it still takes all the feature evidence into account. It is very efficient in terms of storage space and computation time. Training consists only of storing counts of classes and feature occurrences as each example is seen. It performs surprisingly well for classification on many real-world tasks, because the violation of the independence assumption tends not to hurt classification performance. Another advantage is that it is naturally an "incremental learner." – an induction technique that can update its model one training example at a time.

However, the violation of the independence assumption makes probability estimates more extreme: the probability will be overestimated for the correct class and underestimated for the incorrect class(es). This does become a problem if we're going to be using the probability estimates themselves – so Naive Bayes should be used with caution for actual decision-making with costs and benefits.

J. Define generative model, lift, and Naive-Naive Bayes

Generative models, unlike discriminative methods which try directly to discriminate different targets by minimizing loss or entropy, turns the question around and asks: "How do different targets generate feature values?" They attempt to model how the data were generated by applying Bayes' Rule to answer the question: "Which class most likely generated this example?"

Lift measures how much more prevalent the positive class is in the selected subpopulation over the prevalence in the population as a whole. The Naive Bayes equation is slightly modified to assume full feature independence, rather than the weaker assumption of conditional independence. Starting at the prior probability, each piece of evidence – each feature e_i – raises or lowers the probability of the class by a factor equal to that piece of evidence's lift (which may be less than one).

Reading 3.2 *An Introduction to Statistical Learning (Ch. 3 & 6)*

James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). An Introduction to Statistical Learning: with applications in R. New York, NY: Springer. Chapters 3 & 6.

Ch. 3. Linear Regression

Ch. 6. Linear Model Selection and Regularization

Keywords

Residual (p. 62)
The difference between the observed response value and the response value predicted by our linear model

Residual sum of squares (p. 63)
Sum of squared residuals over all observations

Population regression line (p. 63)
The best linear approximation to the true relationship between X and Y

Least squares line (p. 63)
The line characterized by the least squares regression coefficient estimates

Bias (p. 65)
Do we expect the estimate to equal the true population parameter?

Unbiased (p. 65)
An unbiased estimator does not systematically over- or under-estimate the true parameter

Standard error (p. 65)
Tells us the average amount that the estimate differs from the actual value.

Residual standard error (p. 66)
The estimate of the (square root of the) variance of the residuals

Confidence interval (p. 66)
Standard errors can be used to compute confidence intervals. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.

Null hypothesis (p. 67)
The most common hypothesis test involves testing the null hypothesis of H_0: There is no relationship between X and Y.

Alternative hypothesis (p. 67)
For the most common hypothesis test, the alternative hypothesis H_a : There is some relationship between X and Y.

t-statistic (p. 67)
Measures the number of standard deviations that the estimate is away from 0.

R^2 statistic (p. 70)
Measures the proportion of variability in Y that can be explained using X. It is also identical to the squared correlation between X and Y. An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression. A number near 0 indicates that the regression did not explain much of the variability in the response.

Total sum of squares (p. 70)
TSS measures the total variance in the response Y, and can be thought of as the amount of variability inherent in the response before the regression is performed.

F-statistic (p. 75)
Tests whether all of the regression coefficients are zero.

Forward selection (p. 78)
We begin with the null model – a model that contains an intercept but no predictors. We then iteratively add one variable to the model that results in the lowest RSS for each new model (with increasing number of variables in the model), until some stopping rule is satisfied.

Backward selection (p. 79)
We start with all variables in the model, and repeatedly remove the variable with the largest p-value – that is, the variable that is the least statistically significant, until a stopping rule is reached.

Mixed selection (p. 79)
This is a combination of forward and backward selection. We start with no variables in the model, and as with forward selection, we add the variable that provides the best fit. We continue to add variables one-by-one. If at any point the p-value for one of the variables in the model rises above a certain threshold, then we remove that variable from the model. We continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

Dummy variable (p. 84)
An indicator variable that takes on two possible numerical values, 0 or 1, for incorporating a qualitative predictor that only has two levels, or possi- ble values, into a regression model.

Additive linear (p. 86)
The additive linear assumption of the standard linear regression model means that the effect of changes in a predictor on the response is independent of the values of the other predictors, and the change in the response due to a one-unit change in a predictor is constant regardless of the value of the predictor.

Hierarchical principle (p. 89)
If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

Polynomial regression (p. 90)
A very simple way to directly extend the linear model to accommodate non-linear relationships is to include a quadratic term or several polynomial functions of the predictors in the regression model.

Heteroscedasticity (p. 95)
Non-constant variances in the error terms of a regression model.

Multicollinearity (p. 101)
When collinearity exists between three or more variables even if no pair of variables has a particularly high correlation. This can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response, and reduces the accuracy of the estimates of the regression coefficients.

Power (p. 101)
The probability of a hypothesis test correctly detecting a non-zero coefficient.

Variance inflation factor (p. 101)
A better way to assess multicollinearity, than inspecting the correlation matrix of predictor variables, is to compute the variance inflation factor (VIF). The VIF is the ratio of the variance of a coefficient estimate when fitting the full model divided by the variance of of the coefficient estimate if fit on its own.

Best subset selection (p. 205)
To perform best subset selection, we fit a separate least squares regression for each possible combinations and subsets of the p predictors. We then look at all of the resulting models, with the goal of identifying the one that is best.

Deviance (p.206)
measure that plays the role of RSS for a broader class of models. The deviance is negative two times the maximized log-likelihood; the smaller the deviance, the better the fit.

Forward stepwise selection (p. 207)
begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model. At each step the variable that gives the greatest additional improvement to the fit is added to the model.

Backward stepwise selection (p. 208)
begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

C_p (p. 211)
Adds a penalty of $2d\hat{\sigma}^2$, where σ^2 is an estimate of the variance of the error associated with each response measurement, to the training RSS in order to adjust for the fact that the training error tends to underestimate the test error. The penalty increases as the number of predictors in the model increases; this is

intended to adjust for the corresponding decrease in training RSS.

Akaike information criterion (AIC) (p. 211)
Defined for a large class of models fit by maximum likelihood. In the case of the standard linear regression model with Gaussian errors, maximum likelihood and least squares are the same thing, hence C_p and AIC are proportional to each other.

Bayesian information criterion (BIC) (p. 211)
BIC is derived from a Bayesian point of view, but ends up looking similar to C_p (and AIC) as well. It generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than C_p.

Adjusted R^2 (p. 211)
The usual R^2 always increases (since residual sum of squares RSS always decreases) as more variables are added. The intuition behind the adjusted R^2 is that once all of the correct variables have been included in the model, adding noise variables will lead to a decrease in the statistic. In theory, the model with the largest adjusted R^2 will have only correct variables and no noise variables.

Ridge regression (p. 215)
Ridge regression is very similar to least squares, except that the coefficients are estimated by minimizing a slightly different quantity that has the effect of shrinking the coefficient estimates towards zero.

Tuning parameter (p. 215)
The tuning parameter λ serves to control the relative impact of these RSS and shrinkage penalty terms on the coefficient estimates in a ridge regression model

Shrinkage penalty (p. 215)
The second criteria, of shrinking coefficient estimates towards zero, that ridge regression trade-off with seeking coefficient estimates that fit the data well and makes RSS small.

l2 norm (p. 216)
A measure of the distance from coefficient values from zero, equal to the square root of the sum of squared values.

Scale equivalent (p. 217)
When multiplying a predictor variable X_j by a constant c simply leads to a scaling of the coefficient estimates $\hat{\beta}_j$ by a factor of $1/c$.

Sparse (p. 219)
Sparse models involve only a subset of the variables.

Dimension reduction (p. 229)
A class of approaches that transform the predictors and then fit a least squares model using the transformed variables, where the problem is reduced to estimating fewer coefficients.

Linear combination (p. 229)
Linear combinations Z_m of predictor variables $X_1, X_2, ..., X_p$ are represented by $Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$

Principal component analysis (p. 230)
Principal components analysis (PCA) is a popular approach for deriving a low-dimensional set of features from a large set of variables. The first principal component has been chosen so that the projected observations are as close as possible to the original observations, which is also the direction along which the data vary the most. The second principal component is a linear combination of the variables that is

uncorrelated with the first, and has largest variance subject to this constraint. One can then construct up to p distinct principal components.

Principal component regression (p. 233)
A dimension reduction method that involves constructing the first M principal components and then using these components as the predictors in a linear regression model that is fit using least squares. The key idea of PCR is that often a small number of principal components suffice to explain most of the variability in the predictor variables, and assumes that these are also the directions that are associated with Y (though this assumption is not guaranteed)

Partial least squares (p. 237)
PLS first identifies a new set of features that are linear combinations of the original features, and then fits a linear model via least squares using these new features. Unlike PCR, PLS identifies these new features in a supervised way – that is, it makes use of the response in order to identify new features that not only approximate the old features well, but also that are related to the response.

Low-dimensional (p. 238)
Setting in which the number of observations is much greater than the number of features.

High dimensional (p. 239)
The case where the number of features is larger than the number of observations.

Curse of dimensionality (p. 242)
The test error tends to increase as the dimensionality of the problem (i.e. the number of features or predictors) increases, unless the additional features are truly associated with the response.

3.2.1 Simple Linear Regression

A. Define residual and RSS.

$e_i = y_i - \hat{y}_i$ represents the ith residual – this is the difference between the ith observed response value and the ith response value that is predicted by our linear model. RSS is the estimate of the (square root of the) variance of the residuals.

B. Calculate the value of RSS.

We define the residual sum of squares as $RSS = e_1^2 + e_2^2 + ... + e_n^2$ or equivalently as $RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 ... (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$

C. Recognize and apply the least squares coefficient estimates.

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$, $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$

D. Interpret the least squares coefficients.

The least squares approach chooses the least squares coefficients β_0 andβ_1 to minimize the RSS.

E. Define population regression line and least squares line.

The model given by $Y = \beta_0 + \beta_1 X + \epsilon$ defines the population regression line, which is the best linear approximation to the true relationship between X and Y. The least squares regression coefficient estimates that minimizedd the RSS characterize the least squares line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

F. Define the concept of bias and unbiased estimators.

Do we expect the estimate to equal the true population parameter? An unbiased estimator does not systematically over- or under-estimate the true parameter. If we could average a huge number of estimates from a huge number of sets of observations, then this estimate would be spot on.

G. Define standard error and residual standard error.

Standard error tells us the average amount that the estimate differs from the actual value. The residual standard error is the estimate of the (square root of the) variance of the residuals $\hat{\sigma} \equiv RSE = \sqrt{RSS/(n-2)}$.

H. Calculate standard error of a statistic.

The standard errors associated with linear regression coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, are:
$\text{SE}(\hat{\beta}_0^2) = \sigma^2[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}]$, $\text{SE}(\hat{\beta}_1^2) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$

I. Calculate the 95% confidence interval.

For linear regression, the 95% confidence interval for β approximately takes the form $\hat{\beta} \pm 2 \cdot \text{SE}(\hat{\beta})$

J. Describe null and alternative hypothesis.

The most common hypothesis test involves testing the null hypothesis of H_0: There is no relationship between X and Y; versus the alternative hypothesis H_a : There is some relationship between X and Y.

K. Calculate the t-statistic.

$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta})}$

L. Describe how accuracy of a linear regression can be assessed.

The quality of a linear regression fit is typically assessed using two related quantities: the residual standard error (RSE) and the R^2 statistic. The RSE provides an absolute measure of lack of fit of the model The R^2 statistic provides an alternative measure of fit. It takes the form of a proportion – the proportion of variance explained – and so it always takes on a value between 0 and 1, and is independent of the scale.

M. Calculate R^2 statistic given TSS and RSS.

$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$, where TSS, the total sum of squares, is the amount of variability inherent in the response before the regression is performed; and RSS is the residual sum of square.

N. Interpret values of R^2.

An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression. A number near 0 indicates that the regression did not explain much of the variability in the response.

O. Describe the relationship between R^2 and correlation.

R^2 is identical to the squared correlation between X and Y.

P. Define total sum of squares.

TSS $= \sum(y_i - \overline{y})^2$ measures the total variance in the response Y.

3.2.2 Multiple linear regression

A. Interpret the coefficients of a multiple linear regression.

The coefficient β_j quantifies the association between the jth predictor and the response. It is the average effect on Y of a one unit increase in X_j, holding all other predictors fixed.

B. Describe how relationship between response and predictors is tested in a multiple linear regression.

In the multiple regression setting with p predictors, we need to ask whether all of the regression coefficients are zero. We use a hypothesis test to answer this question. We test the null hypothesis, $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$, versus the alternative H_a : at least one β_j is non-zero. This hypothesis test is performed by computing the F-statistic. Sometimes, we want to test that a particular subset of q of the coefficients are zero. In this case we fit a second model that uses all the variables except those last q, then compute residual sum of squares for that model and the appropriate F-statistic.

C. Calculate the F-statistic given TSS, RSS, n, and p

$F = \frac{(TSS-RSS)/p}{RSS/(n-p-1)}$

D. Describe how important variables can be decided in multiple regression.

This task is referred to as variable selection. Ideally, we can try out a lot of different models, each containing a different subset of the predictors, then se- lect the best model out of all of the models that we have considered. Various statistics can be used to judge the quality of a model, such as Mallow's C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R^2.

E. Define forward selection, backward selection, and mixed selection.

Trying out every possible subset of the predictors may be infeasible. Three classical automated and efficient approaches to choose a smaller set of models to consider are:

- Forward selection: We begin with the null model – a model that contains an intercept but no predictors. We then fit p simple linear regressions and add to the null model the variable that results in the lowest RSS. We then add to that model the variable that results in the lowest RSS for the new two-variable model,. This approach is continued until some stopping rule is satisfied.
- Backward selection: We start with all variables in the model, and remove the variable with the largest p-value – that is, the variable that is the least statistically significant. The new (p-1)-variable model is fit, and the variable with the largest p-value is removed. This procedure continues until a stopping rule is reached.
- Mixed selection: This is a combination of forward and backward selection. We start with no variables in the model, and as with forward selection, we add the variable that provides the best fit. We continue to add variables one-by-one. If at any point the p-value for one of the variables in the model rises above a certain threshold, then we remove that variable from the model. We continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

F. Describe the tools used to examine model fit for multiple regression.

Two of the most common numerical measures of multiple regression model fit are the RSE and R^2, the fraction of variance explained. These quantities are computed and interpreted in the same fashion as for simple linear regression.

3.2.3 Considerations in the regression model

A. Define dummy variables.

If a qualitative predictor (also known as a factor) only has two levels, or possible values, then incorporating it into a regression model is very simple. We simply create an indicator or dummy variable that takes on two possible numerical values: 0 or 1.

B. Describe how qualitative variables with more than two levels can be used in multiple regression.

When a qualitative predictor has more than two levels, a single dummy variable cannot represent all possible values. In this situation, we can create additional dummy variables. There will always be one fewer dummy variable than the number of levels.

C. Interpret the coefficients of a dummy variable.

The level with no dummy variable is known as the baseline. The estimated coefficients (of the dummy variables) can be intepreted as the difference in the average response between that level and the baseline category.

D. Describe additive and linear assumptions for linear regression model.

Two of the most important assumptions of the standard linear regression model state that the relationship between the predictors and response are additive and linear. The additive assumption means that the effect of changes in a predictor on the response is independent of the values of the other predictors. The linear assumption states that the change in the response due to a one-unit change in a predictor is constant, regardless of the value of the predictor.

E. Define interaction effect.

Suppose the assumption that the effect on the response of one predictor is independent of the value of another predictor is incorrect. One way of extending this model to allow for interaction effects is to include a third predictor, called an interaction term, which is constructed by computing the product of the values of the two variables.

F. Interpret the coefficients of an interaction term.

We can interpret coefficient as the increase in the effectiveness of one predictor for a one unit increase in the other predictor (or vice-versa).

G. Describe hierarchical principle for multiple regression.

If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

H. Define polynomial regression.

A very simple way to directly extend the linear model to accommodate non-linear relationships, using polynomial regression, is to include transformed versions of the predictors in the model. We can add a quadratic term or several polynomial functions of the predictors in the regression model, and use standard linear regression software to estimate coefficients in order to produce a non-linear fit.

I. Describe the potential problems, such as non-linearity, correlation of error terms, non–constant variance, outliers, high-leverage points, and collinearlity, for linear regression model.

1. **Non-linearity of the Data**: If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspect. In addition, the prediction accuracy of the model can be significantly reduced.
2. **Correlation of Error Terms**: If in fact there is correlation among the error terms, then the estimated standard errors will tend to underestimate the true standard errors. As a result, confidence and prediction intervals will be narrower than they should be.
3. **Non-constant Variance of Error Terms**: The standard errors, confidence intervals, and hypothesis tests associated with the linear model rely upon the assumption that the error terms have a constant variance. Unfortunately, it is often the case that the variances of the error terms are non-constant. For instance, the variances of the error terms may increase with the value of the response.

4. **Outliers**: An outlier is a point for which its response value is far from the value predicted by the model. Outliers can arise for a variety of reasons, such as incorrect recording of an observation during data collection. Removing the outlier may (or may not) have little effect on the least squares line. But can cause other problems such as the RSE, which is used to compute all confidence intervals and can have implications for the interpretation of the fit.
5. **High Leverage Points**: Observations with high leverage have an unusual value for a predictor value. Removing the high leverage observation may have substantial impact on the estimated least squares line.
6. **Collinearity**: This refers to the situation in which two or more predictor variables are closely related to one another. The presence of collinearity can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response. It reduces the accuracy of the estimates of the regression coefficients, and causes the standard errors to grow. As a result, in the presence of collinearity, we may fail to reject null hypothesis that the coefficients are zero. This means that the power of the hypothesis test – the probability of correctly detecting a non-zero coefficient – is reduced.

J. Describe the limits for high leverage for a simple regression

It is cause for concern if the least squares line is heavily affected by just a couple of observations with high leverage, because any problems with these points may invalidate the entire fit.

K. Define heteroscedasticity.

Non-constant variances in the error terms of a regression model.

L. Define power of a hypothesis test.

The probability of a test correctly detecting a non-zero coefficient

M. Define multicollinearity and variance inflation factor.

Multicollinearity is the situation when collinearity exists between three or more variables even if no pair of variables has a particularly high correlation. A better way to assess multicollinearity, than inspecting the correlation matrix of predictor variables, is to compute the variance inflation factor of each predictor variable from the $R^2_{X_j|X_{-j}}$ of the regression onto all of the other predictors.

N. Describe the range of values for variance inflation factors.

If $R^2_{X_j|X_{-j}}$ is close to one, then collinearity is present, and so the VIF will be large. The smallest possible value for VIF is 1, which indicates the complete absence of collinearity. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.

O. Calculate variance inflation factor.

$VIF(\hat{\beta}_j) = 1/(1 - R^2_{X_j|X_{-j}})$. where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.

3.2.4 Subset selection

A. Describe best subset selection.

To perform best subset selection, we fit a separate least squares regression for each possible combination of the p predictors. That is, we fit all p models that contain exactly one predictor, all $p^2 = p(p-1)/2$ models that contain exactly two predictors, and so forth. We then look at all of the resulting 2^p models, with the goal of identifying the one that is best.

B. List the steps used in best subset selection.

1. Let M_0 denote the null model , which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, ...p$:
   - Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   - Pick the best among these $\binom{p}{k}$ models, and call it M_k. Here best is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $M_0, ..., M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .

C. Define deviance.

Deviance is a measure that plays the role of RSS for a broader class of models. The deviance is negative two times the maximized log-likelihood; the smaller the deviance, the better the fit.

D. Describe forward stepwise selection and backward stepwise selection.

Forward and backward stepwise selection are computationally efficient alternatives to best subset selection. While the best subset selection procedure considers all 2^p possible models containing subsets of the p predictors, forward stepwise considers a much smaller set of models. It begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model. In particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model. Backward stepwise selection begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

E. List the steps used in forward stepwise selection and backward stepwise selection.

Forward stepwise selection:

1. Let M_0 denote the null model, which contains no predictors.

2. For $k = 1, 2, ...p - 1$:
    - Choose the best among these $p-k$ models that augment the predictors in M_k with one additional predictor.
    - Choose the best among these $p - k$ models, and call it M_{k+1}. Here best is defined as having the smallest RSS or highest R^2.
3. Select a single best model from among $M_0, ..., M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .

Backward stepwise selection:

1. Let M_p denote the full model, which contains all p predictors.
2. For $k = p, p - 1, ..., 1$:
    - Choose all k models that contain all but one of the predictors for a total of $k - 1$ predictors.
    - Choose the best among these k models, and call it M_{k-1}. Here best is defined as having the smallest RSS or highest R^2.
3. Select a single best model from among $M_0, ..., M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .

F. Recognize and apply the equations for C_p , Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and adjusted R^2.

Techniques for adjusting the training error for the model size to select among a set of models with different numbers of variables, d:

- $C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$, where σ^2 is an estimate of the variance of the error associated with each response measurement, typically estimated using the full model containing all predictors. Adds a penalty of $2d\hat{\sigma}^2$ to the training RSS in order to adjust for the fact that the training error tends to underestimate the test error. The penalty increases as the number of predictors in the model increases; this is intended to adjust for the corresponding decrease in training RSS.
- Akaike Information Criterion: The AIC criterion is defined for a large class of models fit by maximum likelihood. In the case of the standard linear regression model with Gaussian errors, maximum likelihood and least squares are the same thing, hence C_p and AIC are proportional to each other. $AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$
- Bayesian Information Criterion: BIC is derived from a Bayesian point of view, but ends up looking similar to C p (and AIC) as well. BIC replaces the $2d\hat{\sigma}^2$ used by C_p with a $log(n)d\hat{\sigma}^2$ term, where n is the number of observations. Since $\log n > 2$ for any $n > 7$, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models. $BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\hat{\sigma}^2)$
- adjusted $R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$. The usual R^2 always increases (since RSS always decreases) as more variables are added. The intuition behind the adjusted R^2 is that once all of the correct variables have been included in the model, adding noise variables will lead to an increase in $\frac{RSS}{n-d-1}$ (due to the presence of d in the denominator) and consequently a decrease in the statistic. Despite its popularity, and even though it is quite intuitive, the adjusted R^2 is not as well motivated in statistical theory as AIC, BIC, and C_p.

3.2.5 Shrinkage methods

A. Define ridge regression, tuning parameter, and shrinkage penalty.

- Ridge regression: Very similar to least squares, except that the coefficients are estimated by minimizing a slightly different quantity: $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \text{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$.
- Shrinkage penalty: The second term, $\lambda\sum_{j=1}^{p}\beta_j^2$, is small when the coefficients are close to zero, and so it has the effect of shrinking the estimates towards zero.
- Tuning parameter: λ serves to control the relative impact of these RSS and shrinkage penalty terms on the coefficient estimates in a ridge regression model. When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates. However, as $\lambda \to \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero. Selecting a good value for λ is critical, such as by using cross-validation.

B. Define l2 norm and scale equivalent.

- l_2 norm: $||\beta||_2 = \sqrt{\sum_{j=1}^{p}\beta_j^2}$, which measures the distance of β from zero.
- Scale equivalent: When multiplying a predictor variable X_j by a constant c simply leads to a scaling of the coefficient estimates $\hat{\beta}_j$ by a factor of $1/c$.

C. Describe the effect of scale equivalent on regression coefficient.

Regardless of how the jth predictor is scaled, $X_j\hat{\beta}_j$ will remain the same.

D. Define standardizing the predictors.

Using the formula $\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}$ so that the predictor variables are all on the same scale. The denominator is the estimated standard deviation of the jth predictor. Consequently, all of the standardized predictors will have a standard deviation of one. As a result the final fit of the ridge regression model will not depend on the scale on which the predictors are measured.

E. Describe bias-variance tradeoff.

The test mean squared error (MSE) is a function of the variance plus the squared bias. The shrinkage of the ridge coefficient estimates leads to a substantial reduction in the variance of the predictions, at the expense of a slight increase in bias.

F. Describe the ridge regression.

Ridge regression is very similar to least squares, except that the coefficients are estimated by minimizing a slightly different quantity. The ridge regression coefficient estimates are the values that minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \text{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

where $\lambda \geq 0$ is a tuning parameter, to be determined separately.

G. Describe the role of the tuning parameter and shrinkage penalty in determining the value of coefficients of the ridge regression.

As the tuning parameter λ increases, the shrinkage penalty $\lambda \sum_j \beta_j^2$ increases, and the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias. At the least squares coefficient estimates, which correspond to ridge regression with $\lambda = 0$, the variance is high but there is no bias. As the tuning parameter increases, the shrinkage of the ridge coefficient estimates leads to a substantial reduction in the variance of the predictions, at the expense of a slight increase in bias. For an intermediate value of λ, the MSE is considerably lower.

H. Describe how ridge regression improves upon least squares.

Ridge regression's advantage over least squares is rooted in the bias-variance trade-off. In situations where the relationship between the response and the predictors is close to linear, the least squares estimates will have low bias but may have high variance. This means that a small change in the training data can cause a large change in the least squares coefficient estimates. In particular, when the number of variables p is almost as large as the number of observations n, the least squares estimates will be extremely variable. And if $p > n$, then the least squares estimates do not even have a unique solution, whereas ridge regression can still perform well by trading off a small increase in bias for a large decrease in variance. Hence, ridge regression works best in situations where the least squares estimates have high variance.

I. Describe the advantage of Lasso over the ridge regression.

Ridge regression will shrink all of the coefficients towards zero, but will include all p predictors in the final model. Lasso forces some of the coefficient estimates to be exactly equal to zero when the tuning parameter is sufficiently large. Depending on the tuning parameter, lasso can produce a model involving any number of variables, whereas ridge regression will always include all of the variables in the model, although the magnitude of the coefficient estimates will depend on λ. As a result, models generated from the lasso are generally much easier to interpret than those produced by ridge regression.

J. Define a sparse model.

Sparse models involve only a subset of the variables.

K. Describe the variable selection property of the Lasso.

When the tuning parameter is sufficiently large, some of the coefficient estimates are forced to be exactly equal to zero. This can also be visualized by plotting the constraint functions for the lasso which are polytopes (or squares when $p = 2$ or polyhedrons when $p = 3$). Coefficient estimates are given by the first point at which an ellipse of the error function contacts the constraint region. The lasso constraint has corners at each of the axes, and so the ellipse will often intersect the constraint region at an axis. When this occurs, one of the coefficients will equal zero. In higher dimensions, many of the coefficient estimates may equal zero simultaneously. Note that since ridge regression has a circular constraint with no sharp points, this intersection will not generally occur on an axis, and so the ridge regression coefficient estimates will be exclusively non-zero. Hence, much like best subset selection, the lasso performs variable selection.

L. Describe the role of the tuning parameter and shrinkage penalty in determining the value of coefficients of the Lasso.

The lasso coefficients $\hat{\beta}_\lambda^L$ minimize $\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^p |\beta_j|$. The lasso shrinks the coefficient estimates towards zero. The l_1 shrinkage penalty $\lambda \sum_{j=1}^p |\beta_j|$ has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.

M. Compare the Lasso to the Ridge regression.

Compared to ridge regression, the only difference is that the β_j^2 term has been replaced by $|\beta_j|$ in the lasso penalty. Lasso uses an l_1 penalty instead of an l_2 penalty. The l_1 norm of a coefficient vector is given by $||\beta||_1 = \sum |\beta_j|$. Unlike ridge regression, the lasso performs variable selection, and hence results in models that are easier to interpret. In terms of MSE, neither ridge regression nor the lasso will universally dominate the other. In general, one might expect the lasso to perform better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero. Ridge regression will perform better when the response is a function of many predictors, all with coefficients of roughly equal size. Ridge regression and the lasso can also be viewed through a Bayesian lens. If the prior distribution of the coefficients is Gaussian distribution, then the posterior mode is given by the ridge regression solution. If the prior is is a double-exponential (Laplace), then the posterior mode for is the lasso solution.

N. Describe how to select the tuning parameter.

We choose a grid of λ values, and compute the cross-validation error for each value. We then select the tuning parameter value for which the cross-validation error is smallest. Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

3.2.6 Dimension reduction methods

A. Define dimension reduction and linear combination.

Dimension reduction methods are a class of techniques that transform the p predictors and then fit a least squares model using the M transformed variables, where the problem is reduced to estimating fewer $p+1 < M+1$ coefficients. Linear combinations of our original p predictors are represented by $Z_1, Z_2, ..., Z_M$, where $M < p$: $Z_m = \sum_{j=1}^p \phi_{jm} X_j$ for some constants $\phi_{1m}, \phi_{2m}, ..., \phi_{pm}$, $m = 1, ..., M$.

B. Describe principal component analysis.

Principal components analysis (PCA) is a popular approach for deriving a low-dimensional set of features from a large set of variables. The first principal component direction of the data is that along which the observations vary the most. The second principal component is a linear combination of the variables that is uncorrelated with the first, and has largest variance subject to this constraint. One can then construct up to p distinct principal components.

C. Describe variability of data along different principal components.

The first principal component line minimizes the sum of the squared perpendicular distances between each point and the line. Out of every possible linear combination of the variables, this particular linear combination yields the highest variance. The second principal component is a linear combination of the variables that is uncorrelated with the first component, and has largest variance subject to this constraint. Additional components are constructed that successively maximize variance, subject to the constraint of being uncorrelated with the preceding components.

D. Describe principal component regression.

This approach involves constructing the first M principal components, $Z_1, Z_2, ..., Z_M$, and then using these components as the predictors in a linear regression model that is fit using least squares. The key idea is that often a small number of principal components suffice to explain most of the variability in the data, and assumes that the directions in which $X_1, ..., X_p$ show the most variation are the directions that are associated with Y (though this assumption is not guaranteed)

E. Describe partial least squares.

PLS, a supervised alternative to PCR makes use of the response Y in order to identify new features that not only approximate the old features well, but also that are related to the response. After standardizing the p predictors, PLS computes the first direction Z_1 by setting each ϕ_{j1} in equal to the coefficient from the simple linear regression of Y onto X_j, which isis proportional to the correlation between Y and X_j. Hence, PLS places the highest weight on the variables that are most strongly related to the response. The residuals, can be interpreted as the remaining information that has not been explained by the first PLS direction, are then used in exactly the same fashion to compute the next direction Z_2. This iterative approach can be repeated M times to identify multiple PLS components. Finally, at the end of this procedure, we use least squares to fit a linear model to predict Y using $Z_1, ..., Z_M$ in exactly the same fashion as for PCR.

F. Describe other interpretations of PCA.

The first principal component can be interpreted as the line that is as close as possible to the data, so that the projections of the points onto the line are as close as possible to the original observations. Successive components could be constructed, with direction that must be perpendicular, or orthogonal, to the preceding components.

3.2.7 Considerations in high dimensions

A. Define low-dimensional and high-dimensional data.

In a low-dimensional setting, the number of observations is much greater than the number of features. In a high-dimensional setting, the number of features is larger than the number of observations.

B. Describe what goes wrong in high dimensions.

Regardless of whether or not there truly is a relationship between the features and the response, least squares will yield a set of coefficient estimates that result in a perfect fit to the data, such that the residuals are zero. This is problematic because this perfect fit, even though the features are completely unrelated to the response, will almost certainly lead to overfitting of the data, as the resulting model will perform extremely poorly on an independent test set,

C. Describe regression in high dimensions.

Methods such as forward stepwise selection, ridge regression, the lasso, and principal components regression, are particularly useful for performing regression in the high-dimensional setting. Essentially, these approaches avoid overfitting by using a less flexible fitting approach than least squares. Regularization or shrinkage plays a key role in high-dimensional problems, and appropriate tuning parameter selection is crucial for good predictive performance,

D. List the three important points of Lasso when it is applied to high-dimensional data.

1. regularization or shrinkage plays a key role in high-dimensional problems
2. appropriate tuning parameter selection is crucial for good predictive performance, and
3. the test error tends to increase as the dimensionality of the problem (i.e. the number of features or predictors) increases, unless the additional features are truly associated with the response.

E. Define curse of dimensionality.

The test error tends to increase as the dimensionality of the problem (i.e. the number of features or predictors) increases, unless the additional features are truly associated with the response. Noise features exacerbating the risk of overfitting (since noise features may be assigned nonzero coefficients due to chance associations with the response on the training set) without any potential upside in terms of improved test set error. Even if they are relevant, the variance incurred in fitting their coefficients may outweigh the reduction in bias that they bring.

Reading 3.3 *Statistical modeling of financial time series*

Aas, K. and X. K. Dimakos. (2004). Statistical modeling of financial time series: An introduction. Oslo Norway: Norwegian Computing Center.

Keywords

Arithmetic return (p. 3)
$r_t = (Y_t - Y_{t-1})/Y_{t-1}$, where Y_t is the price of the asset at day t

Geometric return (p. 3)
$d_t = \log(Y_t) - \log(Y_{t-1})$, where Y_t is the price of the asset at day t

Time resolution (p. 5)
How densely data are recorded. In applications in the finance industry, this might vary from seconds to years.

Time horizon (p. 5)
Whether we are forecasting over the short-term or long-term.

Random walk model (p. 7)
A commonly used model in finance is the random walk, defined through $Y_t = \mu + Y_{t-1} + \epsilon_t$, where μ is the drift of the process and the increments $\epsilon_1, \epsilon_2, ...$ are serially independent random variables.

Autoregressive model (p. 8)
Random walk models cannot be used for all financial time series, such as interest rates. The autoregressive process provides a simple description of the stochastic nature of interest rates that is consistent with the empirical observation that interest rates tend to be mean-reverting.

AR(1) model (p. 8)
The AR(1)-model: $Y_t = \mu + \alpha Y_{t-1} + \epsilon_t$, where $|\alpha| < 1$ is a parameter and $\epsilon_1, \epsilon_2, ...$ are serially independent random variables. We are back to the random walk model if $\alpha = 1$.

Stationarity (p. 9)
A sequence of random variables X_t is covariance stationary if there is no trend, and if the covariance does not change over time, that is $E[X_t] = \mu$ and $Cov(X_t, X_{t-k}) = E[(X_t - \mu)(X_{t-k} - \mu)] = \gamma_k$ for all t and any k.

Autocorrelation function (p. 10)
Assume that we have a stationary time series X_t with constant expectation and time independent covariance. The autocorrelation function (ACF) for this series is defined as $\rho_k = \frac{Cov(X_t, X_{t-k})}{\sqrt{Var(X_t)Var(X_{t-k})}} = \frac{\gamma_k}{\gamma_0}$ for $k \geq 0$ and $\rho_{-k} = \rho_k$. The value k denotes the lag.

GARCH (1,1) (p. 13)
Describes the evolution of the variance σ_t^2 as $\sigma_t^2 = a_0 + a\epsilon_{t-1}^2 + b\sigma_{t-t}^2$

Marginal distribution (p. 18)
The marginal distribution is the distribution of the errors ϵ_t.

Conditional distribution (p. 18)
The conditional distribution is the distribution of ϵ_t/σ_t.

qq-plot (p.19)
A scatter-plot of the empirical quantiles of the data against the quantiles of a standard normal random variable. If the data is normally distributed, then the quantiles will lie on a straight line.

Shapiro-Wilk test (p.19)
A formal statistical test for normality.

Scaled Student's t-distribution (p.20)
The scaled Students-t distribution allows for heavy tails, and is defined by three parameters: μ and σ representing a location and a dispersion parameter respectively, and ν is the degrees of freedom parameter controlling the heaviness of the tails.

Extreme value theory (p.22)
The methods of extreme value theory focus on modelling the tail behaviour using extreme values beyond a certain threshold rather than all the data.

3.3.1 Concepts of time series

A. Define arithmetic and geometric returns.

Arithmetic returns are: $r_t = (Y_t - Y_{t-1})/Y_{t-1}$. Geometric returns are $d_t = \log(Y_t) - \log(Y_{t-1})$, where Y_t is the price of the asset at day t.

B. Recognize and apply the relationship between arithmetic and geometric returns.

The relationship between geometric and arithmetic returns is given by $D = \log(1 + R)$.

C. Describe the shape of the plotted line when geometric returns are plotted against arithmetic returns.

Geometric returns, D, can be decomposed into a Taylor series of arithmetic returns: $D = R + R^2 + R^3 + ...$, which simplifies to R if R is small. Thus, when arithmetic returns are small, there will be little difference between geometric and arithmetic returns. In practice, this means that if the volatility of a price series is small, and the time resolution is high, geometric and arithmetic returns are quite similar.

D. Define time resolution and time horizon.

Time resolution defines how densely data are recorded: in applications in the finance industry, this might vary from seconds to years. Time horizon determines whether we are forecasting over the short-term or long-term.

E. Describe how time resolution and time horizon affect the distribution of financial data.

The finer the resolution, the heavier the tails of the return distribution are likely to be. For longer time periods, however, many smaller contributions would average out and approach the normal as the lag ahead expands. Hence, market risk analysis over short horizons should consider heavy-tailed distributions of market returns.

3.3.2 Statistical models

A. Describe a random walk model and an autoregressive model.

A commonly used model in finance is the random walk, defined through $Y_t = \mu + Y_{t-1} + \epsilon_t$, where μ is the drift of the process and the increments $\epsilon_1, \epsilon_2, ...$ are serially independent random variables. The assumption of serially independent increments of the series assumes that it at a given moment is impossible to estimate where in the business cycle the economy is, and utilise such knowledge for investment purposes.

Random walk models cannot be used for all financial time series. The autoregressive process provides a simple description of the stochastic nature of interest rates that is consistent with the empirical observation that interest rates tend to be mean-reverting.

B. Recognize and apply an AR(1) model.

The AR(1)-model is the simplest first-order autoregressive model: $Y_t = \mu + \alpha Y_{t-1} + \epsilon_t$, where $|\alpha| < 1$ is a parameter and $\epsilon_1, \epsilon_2, ...$ are serially independent random variables. We are back to the random walk model if $\alpha = 1$. The parameter α determines the speed of mean-reversion towards the stationary value $\frac{\mu}{(1-\alpha)}$.

C. Recognize and apply the variances of a random walk and an autoregressive model.

The variance of the random walk process at time t is given by $Var(Y_t) = t\sigma^2$, i.e. it increases linearly with time. The variance σ might be dependent of the time t.

The stationary variance of the AR(1) autoregressive process is given by: $Var(Y) = \frac{\sigma^2}{1-\alpha^2}$. It is possible to let the volatility depend on time. A very common assumption for interest rates is to parameterise the volatility as a function of interest rate level $\sigma_t = \kappa Y_{t-1}^{\gamma}$.

D. Define stationarity and autocorrelation function.

A sequence of random variables X_t is covariance stationary if there is no trend, and if the covariance does not change over time, that is $E[X_t] = \mu$ and $Cov(X_t, X_{t-k}) = E[(X_t - \mu)(X_{t-k} - \mu)] = \gamma_k$ for all t and any k.

Assume that we have a stationary time series X_t with constant expectation and time independent covariance. The autocorrelation function (ACF) for this series is defined as $\rho_k = \frac{Cov(X_t, X_{t-k})}{\sqrt{(Var(X_t)Var(X_{t-k}))}} = \frac{\gamma_k}{\gamma_0}$ for $k \geq 0$ and $\rho_{-k} = \rho_k$. The value k denotes the lag.

E. Recognize and apply the formula for the autocorrelation function.

By plotting the autocorrelation function as a function of k, we can determine if the autocorrelation decreases as the lag gets larger, or if there is any particular lag for which the autocorrelation is large. For a non-stationary time series, Y_t, the covariance is not independent of t, but given by $Cov(Y_t, Y_{t-k}) = (t-k)\sigma^2$. This means that the autocorrelation at time t and lag k is given by $\rho_{k,t} = \frac{(t-k)\sigma^2}{\sqrt{t\sigma^2(t-k)\sigma^2}} = \sqrt{\frac{t-k}{t}}$. We see that if t is large relative to k, then $\rho_{k,t} \approx 1$.

F. List the properties of an autocorrelation function for an AR(1) process.

1. For both the random walk model and the AR(1)-model, the autocorrelation function for the increments $\epsilon_1, \epsilon_2, ...$ should be zero for any lag $k > 0$.
2. For the AR(1)-model, the autocorrelation function for Y_t should be equal to α^k for lag k (a geometric decline). For a random walk model, the autocorrelation function for Y_t is likely to be close to 1 for all lags.

3.3.3 Modeling volatility

A. Describe a GARCH(1,1) model.

Tt has been observed that although the signs of successive price movements seem to be independent, their magnitude, as represented by the absolute value or square of the price increments, is correlated in time. This phenomena is denoted volatility clustering, and indicates that the volatility of the series is time varying. GARCH-models have been successful at capturing volatility clustering in financial markets and are widely used in industry.

B. List the conditions that must be satisfied by the parameters of a GARCH(1,1) model.

The parameters satisfy $0 \leq a \leq 1,\ 0 \leq b \leq 1,\ a + b \leq 1$.

C. Recognize and apply the variance equation of a GARCH(1,1) model.

The fundamental idea of the GARCH(1,1)-model is to describe the evolution of the variance as $\sigma_t^2 = a_0 + a\epsilon_{t-1}^2 + b\sigma_{t-1}^2$. The variance process is stationary if $a + b < 1$, and the stationary variance is given by $\frac{a_0}{1-a-b}$.

D. Describe the goodness-of-fit for a GARCH model.

The goodness-of-fit of a GARCH-model is evaluated by checking the significance of the parameter estimates and measuring how well it models the volatility of the process. If a GARCH-model adequately captures volatility clustering, the absolute values of the standardised returns, given by $\epsilon_t^* = \hat{\epsilon}_t/\hat{\sigma}_t$, where $\hat{\sigma}_t$ is the estimate of the volatility, should have no autocorrelation. Such tests may performed by graphical inspection of the autocorrelation function, or by more formal statistical tests, like the Ljung-Box statistic

E. Define persistence.

The parameter $\eta = a + b$ is known as persistence and defines how slowly a shock in the market is forgotten.

F. Describe marginal and conditional distributions.

The marginal distribution is the distribution of the errors ϵ_t. The conditional distribution is the distribution of ϵ_t/σ_t.

G. Describe the marginal distribution for a time-series model with variance that follows a GARCH process.

If the variance of the time-series is assumed to follow a GARCH model, this implies that the marginal distribution for the returns has fatter tails than the Gaussian distribution, even though the conditional return distribution is Gaussian.

H. Describe a qq-plot. The qq-plot is an informal graphical diagnostic. It is a scatter-plot of the empirical quantiles of the data against the quantiles of a standard normal random variable.

I. Describe how qq-plot can be used to test normality of data.

If the data is normally distributed, then the quantiles will lie on a straight line. If there are more extreme deviations in the tails, then the the distribution of the standardised returns is more heavy-tailed than the normal distribution.

J. Compare reliability of the Shapiro-Wilk test to the qq-plot.

Even though formal statistical tests for normality like the Shapiro-Wilk test provide numerical measures of goodness of fit which are seemingly more precise than inspecting qq- plots, the reliability of such tests is generally small.

K. Describe the scaled Student's t-distribution.

The scaled Students-t distribution allows for heavy tails, and is defined by three parameters: μ and σ representing a location (mean) and a dispersion (variance) parameter respectively, and ν is the degrees of freedom parameter controlling the heaviness of the tails.

L. Describe the extreme value theory.

The methods of extreme value theory focus on modelling the tail behaviour using extreme values beyond a certain threshold rather than all the data. These methods have two features, which make them attractive for tail estimation; they are based on sound statistical theory, and they offer a parametric form for the tail of a distribution. The idea is to model tails and the central part of the empirical distribution by different kinds of (parametric) distributions. One introduces a threshold u, and consider the probability distribution of the returns, given that they exceed u. It can be shown that when u is large enough, this distribution is well approximated by the generalised Pareto distribution (GPD).

Topic 4. Data Mining & Machine Learning: Classification & Clustering

Reading 4.1 *Data Science for Business (ch. 6)*

Provost, F. and T. Fawcett. (2019). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapter 6.

Ch. 6. Similarity, Neighbors, and Clusters.

Keywords

Euclidean distance (p. 144)
Probably the most common geometric distance measure. Based on the Pythagorean theorem which tells us that the distance between two points is given by the length of the hypotenuse, and is equal to the square root of the summed squares of the lengths of the other two sides of the triangle.

Nearest neighbors (p. 144)
The most-similar instances, using some distance measure, are called nearest neighbors.

Combining function (p. 147)
We predict the new example's target value, based on the nearest neighbors' (known) target values. A combining function (like voting or averaging) operating on the neighbors' known target values will give us a prediction.

Weighted voting (p. 150)
Weighted voting, or similarity moderated voting, scales each neighbor's contribution by its similarity.

Manhattan distance (p. 159)
The Manhattan distance or L1-norm is the sum of the (unsquared) pairwise distances. It is called Manhattan (or taxicab) distance because it represents the total street distance you would have to travel in a place like midtown Manhattan (which is arranged in a grid) to get between two points – the total east-west distance traveled plus the total north-south distance traveled.

Jaccard distance (p.159)
Given two objects, X and Y, the Jaccard distance is the proportion of all the characteristics (that either has) that are shared by the two.

Cosine distance (p. 160)
Cosine distance is often used in text classification to measure the similarity of two documents. It is particularly useful when you want to ignore differences in scale across instances, i.e. when you want to ignore the magnitude of the vectors.

Edit distance or Levenshtein metric (p. 161)
This metric counts the minimum number of edit operations required to convert one string into the other, where an edit operation consists of either inserting, deleting, or replacing a character (one could choose other edit operators).

Clustering (p. 164)
Finding natural groupings in the data may be called unsupervised segmentation, or more simply clustering.

Hierarchical clustering (p. 165)
Hierarchical clusterings generally are formed by starting with each node as its own cluster. Then clusters are merged iteratively until only a single cluster remains. The clusters are merged based on the similarity or distance function that is chosen.

Dendogram (p. 165)
A graph that shows explicitly the hierarchy of the clusters.

Linkage function (p. 167)
For hierarchical clustering, a distance function between any two clusters, considering individual instances to be the smallest clusters.

Cluster center or centroid (p. 170)
The average of the values for each feature of each example in the cluster.

k-means clustering (p. 170)
The most popular centroid-based clustering algorithm is called k-means clustering. The "means" are the centroids, represented by the arithmetic means (averages) of the values along each dimension for the instances in the cluster. The k in k-means is simply the number of clusters that one would like to find in the data, and the k-means clustering method would return (i) the k cluster centroids when cluster method terminates information on which of the data points belongs to each cluster.

Distortion (p. 173)
A numeric measure which is the sum of the squared differences between each data point and its corresponding centroid. The clustering with the lowest distortion value can be deemed the best.

4.1.1 Similarity and distance

A. Recognize and apply the formula for calculating the general Euclidian distance.

General Euclidean distance $= \sqrt{(d_{1,A} - d_{1,B})^2 + (d_{2,A} - d_{2,B})^2 + ... + (d_{n,A} - d_{n,B})^2}$

B. Define nearest neighbors and combining function.

The most-similar instances, using some distance measure, are called nearest neighbors. A combining function (like voting or averaging) operating on the neighbors' known target values will give us a prediction.

C. Explain how combining function can be used for classification.

The nearest neighbors to a new instance are retrieved and their known target variables (classes) are consulted. A simple combining function to give a prediction would be majority vote.

D. Define weighted voting or similarity moderated voting.

Majority voting ignores an important piece of information: how close each neighbor is to the instance. Weighted voting, or similarity moderated voting, scales each neighbor's contribution by its similarity.

E. Calculate contributions for weighted voting for classification.

The contribution to the prediction from each neighbor is scaled by its similarity to the new instance, using as the scaling weight the reciprocal of the square of the distance.

F. Explain how k in k-NN can be used to address overfitting.

In terms of overfitting and its avoidance, k is a complexity parameter. At one extreme, we can set $k = n$ and we do not allow much complexity at all in our model. As described previously, the n-NN model simply predicts the average value in the dataset for each case. At the other extreme, we can set $k = 1$, and we will get an extremely complex model, which places complicated boundaries such that every training example will be in a region labeled by its own class.

G. Discuss issues with nearest-neighbor methods with focus on

- **Intelligibility.** There are two aspects to this issue: the justification of a specific decision and the intelligibility of an entire model. It usually is easy to describe how a single instance is decided: the set of neighbors participating in the decision can be presented, along with their contributions. But it is not easy to explain more deeply what "knowledge" has been mined from the data as there is no explicit model.
- **Dimensionality and domain knowledge.** Nearest-neighbor methods typically take into account all features when calculating the distance between two instances, but many may be irrelevant to the similarity judgment. There are several ways to fix the problem of many, possibly irrelevant attributes. One is feature selection, the judicious determination of features that should be included in the data mining model. Feature selection can be done manually by the data miner or with automated feature selection methods. Another way of injecting domain knowledge into similarity calculations is to tune the similarity/distance function manually, by assigning different weights to the different attributes
- **Computational efficiency.** One benefit of nearest-neighbor methods is that training is very fast because it usually involves only storing the instances. No effort is expended in creating a model. The main computational cost of a nearest neighbor method is borne by the prediction/classification step, when the database must be queried to find nearest neighbors of a new instance.

H. Define curse of dimensionality.

The problems that high-dimensionality pose for nearest neighbor methods. Since all of the attributes (dimensions) contribute to the distance calculations, instance similarity can be confused and misled by the presence of too many irrelevant attributes.

4.1.2 Technical details related to similarities and neighbors

A. Calculate the Manhattan distance and Cosine similarity.

Manhattan distance (L1 norm) $d_{Manhattan}(X,Y) = ||X - Y||_1 = |x_1 - y_1| + |x_2 - y_2| + ...$

Cosine distance $d_{cosine}(X,Y) = 1 - \frac{X \cdot Y}{||X||_2 \cdot ||Y||_2}$ where $|| \cdot ||_2$ represents the L2 norm, or Euclidean length, of each feature vector (for a vector this is simply the distance from the origin).

B. Define the Jaccard distance

Jaccard distance treats the two objects as sets of characteristics. Thinking about the objects as sets allows one to think about the size of the union of all the characteristics of two objects X and Y, $|X \cup Y|$, and the size of the set of characteristics shared by the two objects (the intersection), $|X \cap Y|$. Given two objects, X and Y, the Jaccard distance is the proportion of all the characteristics (that either has) that are shared by the two.

C. Define edit distance or Levenshtein metric.

This metric counts the minimum number of edit operations required to convert one string into the other, where an edit operation consists of either inserting, deleting, or replacing a character.

D. Define clustering, hierarchical clustering, and dendogram.

Clustering is applies the notion of similarity to find groups of objects (consumers, businesses, whiskeys, etc.), where the objects within groups are similar, but the objects in different groups are not so similar. Hierarchical clusterings generally are formed by starting with each node as its own cluster, then clusters are merged iteratively based on the similarity or distance function that is chosen until only a single cluster remains. A dendrograph is graph that shows explicitly the hierarchy of the clusters.

E. Describe how dendogram can help decide the number of clusters.

The diagram can be cut across at any point to give any desired number of clusters. Once two clusters are joined at one level, they remain joined in all higher levels of the hierarchy.

F. List the advantages of hierarchical clustering.

Hierarchical clustering doesn't just create "a clustering," or a single set of groups of objects. It creates a collection of ways to group the points. It allows the data analyst to see the groupings – the "landscape" of data similarity – before deciding on the number of clusters to extract.

G. Define linkage functions.

For hierarchical clustering, we need a distance function between clusters, considering individual instances to be the smallest clusters. This is sometimes called the linkage function. So, for example, the linkage function could be "the Euclidean distance between the closest points in each of the clusters," which would apply to any two clusters.

H. Describe how distance measure can be used to decide the number of clusters in a dendrogram.

Because the y axis represents the distance between clusters, the dendrogram can give an idea of where natural clusters may occur.

I. Define "cluster center" or centroid and k-means clustering.

The "cluster center" or centroid is generally he average of the values for each feature of each example in the cluster. The most popular centroid-based clustering algorithm is called K-means clustering, which would return (i) the k cluster centroids when cluster method terminates information on which of the data points belongs to each cluster. The first step is find the points closest to the chosen centers (possibly chosen randomly). This results in the first set of clusters. The second step finds the actual center of the clusters found in the first step. The process simply iterates: since the cluster centers have shifted, we need to recalculate which points belong to each cluster. Once these are reassigned, we might have to shift the cluster centers again.

J. Compare and contrast k-means clustering with hierarchical clustering.

Unlike hierarchical clustering, k-means starts with a desired number of clusters k.

K. Describe the k-means algorithm.

The algorithm starts by creating k initial cluster centers, usually randomly, but sometimes by choosing k of the actual data points, or by being given specific initial starting points by the user, or via a pre-processing of the data to determine a good set of starting centers. The clusters corresponding to these cluster centers are formed, by determining which is the closest center to each point. Next, for each of these clusters, its center is recalculated by finding the actual centroid of the points in the cluster. Since the cluster centers have shifted, we need to recalculate which points belong to each cluster and reassign. The k-means procedure keeps iterating until there is no change in the clusters (or possibly until some other stopping criterion is met).

L. Describe the reason for running k-means algorithm many times.

There is no guarantee that a single run of the k-means algorithm will result in a good clustering. The result of a single clustering run will find a local optimum – a locally best clustering – but this will be dependent upon the initial centroid locations. For this reason, k-means is usually run many times, starting with different random centroids each time. The results can be compared by examining the clusters (more on that in a minute), or by a numeric measure to choose the best clustering.

M. Define a cluster's distortion.

The clusters' distortion is a numeric measure, which is the sum of the squared differences between each data point and its corresponding centroid. The clustering with the lowest distortion value can be deemed the best clustering.

N. Describe the method for selecting k in k-means algorithm.

One method is simply to experiment with different k values and see which ones generate good results. The value for k can be decreased if some clusters are too small and overly specific, and increased if some clusters are too broad and diffuse. For a more objective measure, the analyst can experiment with increasing values of k and graph various metrics of the quality of the resulting clusterings. As k increases the quality metrics should eventually stabilize or plateau, either bottoming out if the metric is to be minimized or topping out if maximized.

Reading 4.2 *Introduction to Statistical Learning (ch. 4 & 10)*

James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). An Introduction to Statistical Learning: with applications in R. New York, NY: Springer. Chapters 4 & 10.

Ch. 4. Classification

Ch. 10. Unsupervised Learning

Keywords

Logistic function (p. 132)
Example of a function that gives outputs between 0 and 1

Odds (p. 132)
A probability value divided by 1 minus the value

Log odds (p. 132)
Taking the logarithm of the odds value.

Likelihood function (p. 133)
The probability predicted by the fitted model of the class values observed.

Principal component analysis (p. 375)
The process by which principal components are computed, and the subsequent use of these components in understanding the data.

Loadings (p. 375)
Coefficients in the normalized linear combinations of the features to form principal components.

Bottom-up agglomerative clustering (p. 390)
This is the most common type of hierarchical clustering built starting from the leaves and combining clusters up to the trunk (of a dendrogram).

Linkage (p. 394)
Extends the concept of dissimilarity between a pair of observations to a pair of groups of observations. The four most common types of linkage are complete, average, single, and centroid.

Inversion (p. 395)
When two clusters are fused at a height below either of the individual clusters in the dendrogram.

4.2.1 Logistic regression

A. Describe the limitations of linear regression for categorical response variable.

In general there is no natural way to convert a qualitative response variable with more than two levels into a quantitative response that is ready for linear regression.

B. Calculate the value of a logistic function.

The logistic function gives outputs between 0 and 1 for all input values hence can represent probabilities: $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$.

C. Calculate odds, and log-odds.

The quantity $\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$ is called the odds, and can take on any value between 0 and ∞. Values of the odds close to 0 and ∞ indicate very low and very high probabilities of default, respectively. Taking the logarithms, the left-hand side is called the log-odds or logit: $\log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X$.

D. Recognize and apply the likelihood function.

The likelihood function $l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{j:y_j=0} p(x_j)$. is maximized to choose the estimates $\hat{\beta}_0, \hat{\beta}_1$. The basic intuition behind using maximum likelihood to fit a logistic regression model is to seek coefficient estimates such that the predicted probability $\hat{p}(x_i)$ for each data point corresponds as closely as possible to its observed class.

E. Interpret the logistic regression coefficients with single regressor.

Increase in the log odds of a one-unit increase in the regressor.

F. Recognize and apply predictions using the logistic regression.

Once the coefficients have been estimated, the predicted probability is $\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\beta_0+\beta_1 X}}$.

G. Interpret the multiple logistic regression coefficients

Increase in the log odds of a one-unit increase in the regressor, holding values of all other regressors fixed.

4.2.2 Principal component analysis

A. Define principal component analysis.

The process by which principal components are computed, and the subsequent use of these components in understanding the data. PCA is an unsupervised approach, since it involves only a set of features. Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization.

B. Describe how principal components are found.

PCA finds a low-dimensional representation of a data set that contains as much as possible of the variation. PCA seeks a small number of dimensions that are as interesting as possible, where the concept of interesting is measured by the amount that the observations vary along each dimension. Each of the dimensions found by PCA is a linear combination of the p features. The first principal component of a set of features $X_1, X_2, ..., X_p$ is the normalized linear combination of the features $Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + ... + \phi_{p1}X_p$ that has the largest variance. In other words, the first principal component loading vector solves the optimization problem: $\max_{\phi_{11}...,\phi_{p1}} \{\frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p}\phi_{j1}x_{ij})^2\}$ subject to $\sum_{j=1}^{p}\phi_{j1}^2 = 1$. We can then iteratively find the second principal component Z_2 (the linear combination of features that has maximal variance out of all linear combinations that are uncorrelated with previously found components), third component Z_3, and so on.

C. Define loadings of a principal component.

Coefficients $\phi_{11}, \phi_{21}, ..., \phi_{p1}$ in the normalized linear combination of the features to form principal components of the set of features.

D. Describe the constraints used in finding the principal components

We constrain the loadings so that their sum of squares is equal to one, since otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance. We can then iteratively find successive principal components as the linear combination of features that has maximal variance out of all linear combinations that are constrained to be uncorrelated with previously found components.

E. Describe another interpretation of the principal component.

Principal components provide low-dimensional linear surfaces that are closest to the observations. Using this interpretation, together the first M principal component score vectors and the first M principal component loading vectors provide the best M-dimensional approximation (in terms of Euclidean distance) to the ith observation: $x_{ij} \approx \sum_{m=1}^{M} z_{im}\phi_{jm}$

F. Describe the effect of scaling on the principal component.

The results obtained when we perform PCA will also depend on whether the variables have been individually scaled (each multiplied by a different constant). If we perform PCA on the unscaled variables, then the first principal component loading vector will have large loadings on the variables with the highest variance. However, this result is simply a consequence of the scales or units on which the variables were measured.

G. Describe uniqueness of the principal component.

Each principal component loading vector is unique, up to a sign flip. The signs may differ because each principal component loading vector specifies a direction in p-dimensional space: flipping the sign has no effect as the direction does not change.

H. Define proportion of variance explained and scree plot.

The proportion of variance explained (PVE) of the mth principal component is given by the ratio the variance explained by the mth principal component, and the total variance present in a data set (assuming that the variables have been centered to have mean zero): PVE $= \frac{\sum_{i=1}^{n}(\sum_{j=1}^{p}\phi_{jm}x_{ij})^2}{\sum_{j=1}^{p}\sum_{i=1}^{n}x_{ij}^2}$. A scree plot depicts the proportion of variance explained by each of the principal components.

I. Describe how number of principal components is decided.

We typically decide on the number of principal components required to visualize the data by examining a scree plot. We choose the smallest number of principal components that are required in order to explain a sizable amount of the variation in the data. This is done by eyeballing the scree plot, and looking for a point at which the proportion of variance explained by each subsequent principal component drops off. This is often referred to as an elbow in the scree plot. This type of visual analysis is inherently ad hoc. Unfortunately, there is no well-accepted objective way to decide how many are enough.

4.2.3 Clustering methods

A. Describe the algorithm for k-means clustering.

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   (a) For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   (b) Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

B. Describe the objective function of the k-means clustering.

The objective is to find a good clustering for which the within-cluster variation is as small as possible. The within-cluster variation of cluster c_k is a measure $W(C_k)$ of the amount by which the observations within a cluster differ from each other. Hence we want to solve the problem $min_{C_1,\ldots,C_K}\sum_{k=1}^{K}W(C_k)$. This formula says that we want to partition the observations into K clusters such that the total within-cluster variation, summed over all K clusters, is as small as possible. within-cluster variation. There are many possible ways to define this concept, but by far the most common choice involves squared Euclidean distance.

C. Define bottom-up agglomerative clustering.

The most common type of hierarchical clustering, which is an alternative approach (to K-means clustering) that does not require that we commit to a particular choice of K. bottom-up agglomerative clustering efers to the fact that a dendrogram is built starting from the leaves (at the bottom) and combining clusters up to the trunk.

D. Describe how to interpret a dendogram.

A dendrogram is generally depicted as an upside-down tree. Each leaf of the dendrogram represents one of the observations. As we move up the tree, some leaves begin to fuse into branches. These correspond to observations that are similar to each other. As we move higher up the tree, branches themselves fuse, either with leaves or other branches. The earlier (lower in the tree) fusions occur, the more similar the groups of observations are to each other. On the other hand, observations that fuse later (near the top of the tree) can be quite different. The height of this fusion, as measured on the vertical axis, indicates how different the two observations are. We draw conclusions about the similarity of two observations based on the location on the vertical axis where branches containing those two observations first are fused.

E. Describe how to extract clusters from a dendrogram.

We begin by defining some sort of dissimilarity measure between each pair of observations. Most often, Euclidean distance is used. The algorithm proceeds iteratively. Starting out at the bottom of the dendrogram, each of the n observations is treated as its own cluster. The two clusters that are most similar to each other are then fused so that there now are $n-1$ clusters. Next the two clusters that are most similar to each other are fused again, so that there now are $n-2$ clusters. The algorithm proceeds in this fashion until all of the observations belong to one single cluster, and the dendrogram is complete

F. Define linkage as well as the four types of linkage.

Linkage extends the concept of dissimilarity between a pair of observations to a pair of groups of observations:

1. Complete: Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities.
2. Single: Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
3. Average: Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the average of these dissimilarities.
4. Centroid: Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable inversions.

G. Define inversion.

When two clusters are fused at a height below either of the individual clusters in the dendrogram. Centroid linkage, often used in genomics, suffers from a major drawback in that an inversion can occur. This can lead to difficulties in visualization as well as in interpretation of the dendrogram.

H. Describe the choice of dissimilarity measure for clustering.

The choice of dissimilarity measure is very important, as it has a strong effect on the resulting dendrogram. Correlation-based distance considers two observations to be similar if their features are highly correlated, even though the observed values may be far apart in terms of Euclidean distance. Correlation-based

distance focuses on the shapes of observation profiles rather than their magnitudes. One must also consider whether or not the variables should be scaled to have standard deviation one before the dissimilarity between the observations is computed.

I. Compare and contrast correlation-based distance measures to Euclidean distance measures.

Correlation-based distance considers two observations to be similar if their features are highly correlated, even though the observed values may be far apart in terms of Euclidean distance. Correlation-based distance focuses on the shapes of observation profiles rather than their magnitudes.

J. Describe the practical issues with clustering.

1. Small Decisions with Big Consequences: In order to perform clustering, some decisions must be made. In practice, we try several different choices, and look for the one with the most useful or interpretable solution:
    - Should the observations or features first be standardized in some way?
    - In the case of hierarchical clustering, what dissimilarity and type of linkage should be used; Where should we cut the dendrogram?
    - In the case of K-means clustering, how many clusters should we look for in the data?
2. Validating the Clusters Obtained: Whether the clusters that have been found represent true subgroups in the data, or whether they are simply a result of clustering the noise.
3. Other Considerations in Clustering: Clustering methods generally are not very robust to perturbations to the data, a set of clusters formed after removing a random subset of observations can be quite dis-similar. Clustering methods that force every observation into a cluster find clusters that may be heavily distorted due to the presence of outliers that do not belong to any cluster – other approaches like mixture models may be more attractive.
4. A Tempered Approach to Interpreting: Results should not be taken as the absolute truth about a data set but as a starting point for further study. We recommend performing clustering with different choices of parameters, and clustering subsets of the data in order to get a sense of robustness.

Topic 5. Data Mining & Machine Learning: Performance Evaluation, Backtesting & False Discoveries

Reading 5.1 *Data Science for Business (ch. 7 & 8)*

Provost, F. and T. Fawcett. (2019). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapters 7 & 8.

Ch. 7. Decision Analytic Thinking I: What Is a Good Model?

Ch. 8. Visualizing Model Performance.

Keywords

Accuracy (p. 189)
The proportion of correct decisions, as a measure of classifier performance.

Confusion matrix (p. 189)
A sort of contingency table, which separates out the decisions made by the classifier, making explicit how one class is being confused for another.

False positive (p. 190)
Negative instances classified as positive.

False negative (p. 190)
Positives instances classified as negative.

Expected value (p. 194)
In an expected value calculation the possible outcomes of a situation are enumerated. The expected value is then the weighted average of the values of the different possible outcomes, where the weight given to each value is its probability of occurrence.

Class prior (p. 201)
The probabilities of seeing each class.

Precision (p. 204)
Precision is the accuracy over the cases predicted to be positive.

Recall (p. 204)
Recall is the same as true positive rate, the frequency of being correct when the instance is actually positive.

F-measure (p. 204)
The F-measure is the harmonic mean of precision and recall at a given point.

Profit curve (p. 212)
Ranking the list of instances by score from highest to lowest and sweeping down through it, recording the expected profit after each instance.

Base rate (p.214)
Usually refers to the proportion of positives in the target population.

ROC graph (p. 215)
The Receiver Operating Characteristics (ROC) graph shows the entire space of performance possibilities. It is a two-dimensional plot of a classifier with false positive rate on the x axis against true positive rate on the y axis.

Hit rate (p. 216)
What percent of the actual positives does the classifier get right.

False alarm rate (p.216)
What percent of the actual negative examples does the classifier get wrong (i.e., predict to be positive).

AUC (p. 219)
Though a ROC curve provides more information, the area under a classifier's curve (AUC) expressed as a fraction of the unit square is useful when a single number is needed to summarize performance, or when nothing is known about the operating conditions.

Cumulative response curve (p. 219)
ROC curves are not the most intuitive visualization for many business stakeholders who really ought to understand the results. An alternate visualization is the use of the "cumulative response curve," rather than the ROC curve. They are closely related, but the cumulative response curve is more intuitive. Cumulative response curves plot the hit rate (tp rate; y axis), i.e., the percentage of positives correctly classified, as a function of the percentage of the population that is targeted (x axis).

5.1.1 Evaluating classifiers

A. Define accuracy.

accuracy = (Number of correct decisions made) / (Total number of decisions made)

B. Describe a confusion matrix.

A confusion matrix for a problem involving n classes is an $n \times n$ matrix with the columns labeled with actual classes and the rows labeled with predicted classes. Each example in a test set has an actual class label as well as the class predicted by the classifier (the predicted class), whose combination determines which matrix cell the instance counts into.

C. Define false positives and false negatives.

The errors of the classifier are the false positives (negative instances classified as positive) and false negatives (positives classified as negative).

D. Describe the problems with unbalanced data.

Because the unusual or interesting class is rare among the general population, the class distribution is unbalanced or skewed. As the class distribution becomes more skewed, evaluation based on accuracy breaks down. The "base rate" of a class, With such skewed domains the base rate for the majority class could be very high, which corresponds to how well a classifier would perform by simply choosing that class for every instance.

E. Identify false positive and false negative from a confusion matrix. The off-diagonal contains the counts of the errors of the classifier: the false positives (negative instances classified as positive) and false negatives (positives classified as negative).

F. Describe the problems with unequal costs and benefits.

Another problem with simple classification accuracy as a metric is that it makes no distinction between false positive and false negative errors. By counting them together, it makes the tacit assumption that both errors are equally important. With real-world domains this is rarely the case. These are typically very different kinds of errors with very different costs because the classifications have consequences of differing severity.

5.1.2 A key analytical framework: expected value

A. Calculate expected value and expected benefit.

The general form of an expected value calculation $EV = p(o_1) \cdot v(o_1) + p(o_2) \cdot v(o_2) + p(o_3) \cdot v(o_3)...$ Each o_i is a possible decision outcome; $p(o_i)$ is its probability and $v(o_i)$ is its value.

B. Describe how expected value can be used to frame classifier use.

In use, we have many individual cases for which we would like to predict a class, which may then lead to an action. The benefits of action and inaction need to be determined separately, as part of the Business Understanding step, while the respective probabilities come from the historical data as summarized in our predictive model. The expected value can then be calculated to determine if classifier use has expected value (profit) greater than zero.

C. Describe how expected value can be used to frame classifier evaluation.

Expected value can also be used to determine, in aggregate, how well does each model do. Instead of individual decisions, we evaluate the collection of decisions made by a model when applied to a set of examples. We can use the expected value framework just described to determine the best decisions for each particular model. Each outcome corresponds to one of the possible combinations of the class we predict, and the actual class. We then calculate the expected profit for a model in aggregate over all the different possible cases.

D. Define class priors.

The class priors, p(p) and p(n), specify the likelihood of seeing positive and negative instances, respectively.

E. Calculate expected profit using priors.

Expected profit equation with priors $\mathrm{p}(p)$ and $\mathrm{p}(n)$ factored is $\mathrm{p}(p) \cdot [\mathrm{p}(Y|p) \cdot b(Y,p) + \mathrm{p}(N|p) \cdot c(N,p) + \mathrm{p}(n) \cdot [\mathrm{p}(N|n) \cdot b(N,n) + \mathrm{p}(Y|n) \cdot c(Y,n)$, where correct classifications (true positives and true negatives) correspond to the benefits $b(Y,p)$ and $b(N,n)$, respectively. Incorrect classifications (false positives and false negatives) correspond to the "benefit" $b(Y,n)$ and $b(N,p)$, respectively, which may well actually be a cost (a negative benefit). From this equation, we now have one component (the first one) corresponding to the expected profit from the positive examples, and another (the second one) corresponding to the expected profit from the negative examples. Each of these is weighted by the probability that we see that sort of example.

F. Describe the two pitfalls common to formulating cost-benefit analysis.

1. It is important to make sure the signs of quantities in the cost-benefit matrix are consistent. Either take benefits to be positive and costs to be negative, or if the focus is on minimizing cost rather than maximizing profit, then the signs are reversed. Mathematically, there is no difference, but it is important to pick one view and be consistent.
2. An easy mistake in formulating cost-benefit matrices is to "double count" by putting a benefit in one cell and a negative cost for the same thing in another cell (or vice versa). A useful practical test is to compute the benefit improvement for changing the decision on an example test instance.

G. Define precision and recall.

Precision is TP/(TP + FP), while recall is TP/(TP + FN)

H. Calculate the value of the F-measure.

F-measure is $2 \cdot (\text{precision} \cdot \text{recall}) / (\text{precision} + \text{recall})$

5.1.3 Visualizing model performance

A. Describe a ranking classifier.

A different strategy for making decisions is to rank a set of cases by these scores, and then take actions on the cases at the top of the ranked list. Instead of deciding each case separately, we may decide to take the top n cases (or, equivalently, all cases that score above a given threshold). It may be that the model gives a score that ranks cases by their likelihood of belonging to the class of interest, but which is not a true probability. Or it also may be that costs and benefits cannot be specified precisely, but nevertheless we would like to take actions.

B. Define a profit curve.

With a ranking classifier, we can produce a list of instances and their predicted scores, ranked by decreasing score, and then measure the expected profit that would result from choosing each successive cut-point in the list. At each cut-point we record the percentage of the list predicted as positive and the corresponding estimated profit. Graphing these values gives us a profit curve.

C. Describe the properties of a profit curve.

Each curve shows the expected cumulative profit for that classifier as progressively larger proportions of the list are targeted. All curves begin and end at the same point.

D. Describe the ROC graph.

A Receiver Operating Characteristics (ROC) graph is a two-dimensional plot of a classifier with false positive rate on the x axis against true positive rate on the y axis. As such, a ROC graph depicts relative trade-offs that a classifier makes between benefits (true positives) and costs (false positives).

E. Define base rate.

The class prior, or proportion, of positive instances in the target population.

F. Describe the four corners and the diagonal of the ROC graph.

The lower left point (0, 0) represents the strategy of never issuing a positive classification; such a classifier commits no false positive errors but also gains no true positives. The opposite strategy, of unconditionally issuing positive classifications, is represented by the upper right point (1, 1). The point (0, 1) represents perfect classification, represented by a star. The diagonal line connecting (0, 0) to (1, 1) represents the policy of guessing a class. No classifier should be in the lower right triangle of a ROC graph. This represents performance that is worse than random guessing.

G. Define hit rate and false alarm rate.

The true positive (tp) rate is sometimes referred to as the hit rate – what percent of the actual positives does the classifier get right. The false positive (fp) rate is sometimes referred to as the false alarm rate – what percent of the actual negative examples does the classifier get wrong (i.e., predict to be positive).

H. Describe how the ROC space can be used to evaluate classifiers.

One point in ROC space is superior to another if it is to the northwest of the first (*tp* rate is higher and *fp* rate is no worse; *fp* rate is lower and *tp* rate is no worse, or both are better). Classifiers appearing on the lefthand side of a ROC graph, near the x-axis, may be thought of as "conservative": they raise alarms (make positive classifications) only with strong evidence so they make few false positive errors, but they often have low true positive rates as well. Classifiers on the upper righthand side of a ROC graph may be thought of as "permissive": they make positive classifications with weak evidence so they classify nearly all positives correctly, but they often have high false positive rates. An advantage of ROC graphs is that they decouple classifier performance from the conditions under which the classifiers will be used. Specifically, they are independent of the class proportions as well as the costs and benefits. The region(s) on the ROC graph that are of interest may change as costs, benefits, and class proportions change, but the curves themselves should not.

I. Define AUC.

This is simply the area under a classifier's curve expressed as a fraction of the unit square. Its value ranges from zero to one. The AUC is equivalent to the Mann-Whitney-Wilcoxon measure, a well-known ordering measure in Statistics. It is also equivalent to the Gini Coefficient, with a minor algebraic transformation. Both are equivalent to the probability that a randomly chosen positive instance will be ranked ahead of a randomly chosen negative instance.

J. Describe a cumulative response curve.

Cumulative response curves plot the hit rate (*tp* rate; y axis), i.e., the percentage of positives correctly classified, as a function of the percentage of the population that is targeted (x axis). Conceptually as we move down the list of instances ranked by the model, we target increasingly larger proportions of all the instances. Hopefully in the process, if the model is any good, when we are at the top of the list we will target a larger proportion of the actual positives than actual negatives. As with ROC curves, the diagonal line x=y represents random performance.

Reading 5.2 *A Backtesting Protocol in the Era of Machine Learning.*

Arnott, R., C. B. Harvey, and H. Markowitz. (2019). A Backtesting Protocol in the Era of Machine Learning. Journal of Financial Data Science, 1(1), 64-74.

Keywords

Exaggerated positive (p. 68)
An outcome that seems stronger, perhaps much stronger, than it is likely to be in the future.

5.2.1 Backtesting Protocol in the Era of Machine Learning

A. List the five attractive features of the simulated strategy.

1. It relies on a consistent methodology through time.
2. Performance in the most recent period does not trail off, indicating that the strategy is not crowded.
3. The strategy does well during the financial crisis, gaining nearly 50%.
4. The strategy has no statistically significant correlations with any of the well-known factors, such as value, size, and momentum, or with the market as a whole.
5. The turnover of the strategy is extremely low, less than 10% a year, so the trading costs should be negligible.

B. Describe the lessons learned from the data-mined strategy.

1. The strategy was discovered by brute force, not machine learning. Machine learning implementations would carefully cross-validate the data by training the algorithm on part of the data and then validating on another part of the data. In this case, the cross-validation is not randomized: the strategy was identified in the first quarter-century of the sample, then found to work in the second quarter-century.
2. The data are very limited. Today, we have about 55 years of high-quality equity data (or less than 700 monthly observations) for many of the metrics in each of the stocks we may wish to consider. This tiny sample is far too small for most machine learning applications and impossibly small for advanced approaches such as deep learning.
3. We have a strong prior that the strategy is false: If it works, it is only because of luck. Machine learning, and particularly unsupervised machine learning, does not impose economic principles. If it works, it works in retrospect but not necessarily in the future.

C. Describe the winner's curse.

In other areas of science, this phenomenon of an outcome that seems much stronger than it is likely to be in the future, is sometimes called the winner's curse.

D. Define exaggerated positive.

An outcome that seems stronger, perhaps much stronger, than it is likely to be in the future. Once the trial is replicated, the effect is far smaller than in the original finding (e.g., if microcap stocks are excluded or if the replication is out of sample); or there is no effect, and the research is eventually discredited. In investing, private gain and social loss pose a twist to the winner's curse: The investment manager pockets the fees until the flaw of the strategy becomes evident, and the investor bears the losses until the great reveal that it was a bad strategy all along.

E. Describe the seven protocols suggested for avoiding false positives.

1. RESEARCH MOTIVATION
   - Establish an Ex Ante Economic Foundation
   - Beware an Ex Post Economic Foundation
2. MULTIPLE TESTING AND STATISTICAL METHODS
   - Keep Track of What Is Tried
   - Keep Track of Combinations of Variables
   - Beware the Parallel Universe Problem: a discovery at two sigma for a lucky test oucome would be discarded because a two-sigma threshold is too low for, say, 20 different tests.
3. SAMPLE CHOICE AND DATA
   - Define the Test Sample Ex Ante
   - Ensure Data Quality
   - Document Choices in Data Transformations

- Do Not Arbitrarily Exclude Outlier
- Select Winsorization Level before Constructing the Model

4. CROSS-VALIDATION
   - Recognize That Iterated Out of Sample Is Not Out of Sample
   - Do Not Ignore Trading Costs and Fees
   - Acknowledge Out of Sample Is Not Really Out of Sample
5. MODEL DYNAMICS
   - Be Aware of Structural Changes
   - Acknowledge the Heisenberg Uncertainty Principle and Overcrowding
   - Refrain from Tweaking the Model
6. MODEL COMPLEXITY
   - Beware the Curse of Dimensionality
   - Pursue Simplicity and Regularization
   - Seek Interpretable Machine Learning
7. RESEARCH CULTURE
   - Establish a Research Culture
   - That Rewards Quality
   - Be Careful with Delegated Research

Reading 5.3 *An investigation of the false discovery rate and the misinterpretation of p-values.*

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science, London, U.K.: Royal Society Open Science.

Keywords

Specificity (p. 2)
In a test, the fraction of samples without the condition that will be correctly diagnosed as not having it.

Sensitivity (p. 2)
In a test, the fraction of samples with the condition that will be detected.

Power (p. 4)
The probability that the test will give the right result when there is a real effect.

5.3.1 An investigation of the false discovery rate and the misinterpretation of p-values

A. Define specificity and sensitivity.

Specificity is the fraction of samples without the condition that will be correctly diagnosed as not having it; Sensitivity is the fraction with the condition that will be detected by the test.

B. Describe the false discovery rate with the help of a tree diagram.

If there were actually no effect (if the true difference between means were zero) then the probability of observing a value for the difference equal to, or greater than, that actually observed would be the p-level. In other words, the p-level is the probability of seeing a difference at least as big as we have done, by chance alone.

To calculate the false discovery rate, we must take a guess at the fraction of tests that we do in which there is a real difference, and the probability that the test will give the right result when there is a real effect (power). Figure from page 4 of Colquhoun (2014):

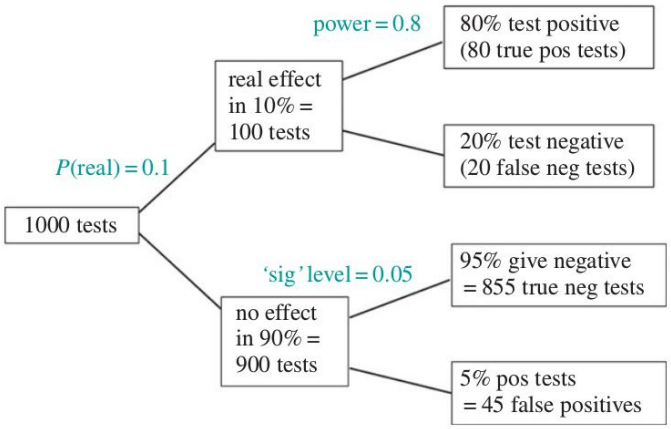


Figure 2. Tree diagram to illustrate the false discovery rate in significance tests. This example considers 1000 tests, in which the prevalence of real effects is 10%. The lower limb shows that with the conventional significance level, $p = 0.05$, there will be 45 false positives. The upper limb shows that there will be 80 true positive tests. The false discovery rate is therefore $45/(45 + 80) = 36\%$, far bigger than 5%.

C. Calculate the probability of real effect given a result is significant.

One minus the false discovery rate (which is the probability of no real effect given a result is significant).

D. Define power of a test.

The probability that the test will give the right result when there is a real effect. The power of the significance test is the same thing as the sensitivity of a screening test.

E. Describe the false discovery rate in simulated t-tests.

Combining the simulations that had no true effect with those simulations for which there was a true effect in the same proportion as that assumed in the tree diagram delivers the same false discovery rate (which is also greater than the assumed p-value)

F. Calculate false discovery rate.

The ratio of false discoveries to number of positive tests, which is the sum of false discoveries (i.e. false positives) and true positives. To calculate the false discovery rate, we must take a guess at the fraction of tests that we do in which there is a real difference, and the probability that the test will give the right result when there is a real effect (power).

G. Describe underpowered study.

Underpowered studies have low sensitivity, i.e. low probability that the test will give the right result when there is a real effect, which contributes to both high false discovery rates and effect size inflation.

H. Describe the inflation effect in the context of false discovery.

When the test correctly spots a difference of means, it get its size wrong – because the test is more likely to be positive in the small number of experiments that show a larger than average effect size. The inflation effect gets really serious when the power is low.

I. Describe what happens when we consider p=0.05 rather than p¡=0.05.

An observation of $p \sim 0.05$ tells you remarkably little about whether or not you have made a discovery, the simulations find the least a 26% chance of being wrong (that you have made a discovery), and often a much bigger chance.

J. Describe Berger's approach.

A Bayesian would refer to the prevalence as the prior probability that there is a real effect. Berger devised a result that applies regardless of what the shape of the prior distribution might be. For $p = 0.05$, the false discovery rate calculated in this way is 0.289, which is quite close to the simulated t-tests. Berger's calibration suggests that $p = 0.0027$ corresponds to a false discovery rate of 0.042, not far from the 0.05 level that is customarily abused. This procedure amounts roughly to adopting a three-sigma policy, rather than a two-sigma rule.

Reading 5.4 *A Data Science Solution to the Multiple-Testing Crisis in Financial Research.*

Lopez de Prado, M. (2019). A Data Science Solution to the Multiple-Testing Crisis in Financial Research. Journal of Financial Data Science, 1(1), 99-110.

Keywords

Selection bias under multiple testing (p. 99)
Due to SBuMT, the probability of obtaining a false positive would increase as a test is repeated multiple times over the same dataset and selecting the most favorable outcome (the one that rejects the null with the lowest false positive probability).

5.4.1 A Data Science Solution to the Multiple-Testing Crisis

A. Define selection bias under multiple testing.

Standard hypothesis test used in most scientific applications may not consider the possibility of performing multiple tests on the same dataset and selecting the most favor- able outcome (the one that rejects the null with the lowest false positive probability). But the probability of obtaining a false positive would increase as a test is repeated multiple times over the same dataset.

B. Describe the three properties satisfied by the heatmap of the correlation matrix of the returns of all trials.

1. Complete. The set includes every backtest computed by any of the authors for this or similar investment mandates. Researchers do not have the ability to delete trials, and they are not allowed to backtest outside the official research platform.
2. Coerced. Researchers do not choose what to log or present. Terabytes of intermediate research metadata are automatically recorded and curated by research surveillance systems.
3. Untainted. Every batch of backtests must be preapproved by the research committee to prevent that externally preselected trials contaminate the internal trials.

C. Describe clustering of trials.

To determine the number of effectively uncorrelated clusters of trials, which may depend on how much researchers searched predefined theoretical foundations or less mathematical (more arbitrary) configurations of strategies.

D. Describe cluster statistics and how it can reduce the probability of selecting a false positive.

Forming one time series per cluster further the bias caused by selecting outliers: We do not evaluate the strategy based on a single (potentially "lucky") trial, but based on a large collection of similar trials. In particular, we compute each cluster's returns applying the minimum variance allocation so that highly volatile trials do not dominate the returns time series

E. Describe the implications for authors, journals, and financial firms.

To reassert credibility, the publication of future discoveries could be accompanied with information regarding all the trials involved in those discoveries:

- authors could (1) explain why the purported discovery is not a false positive caused by SBuMT; (2) certify that they have logged and recorded all the trials that took place during their research; and (3) provide to journal referees the outcomes from all trials.
- Journals could publish the outcomes from all trials in their websites so that researchers can evaluate the totality of the evidence, not only the trials handpicked by the authors or referees. Journals could demand that authors (1) disclose all trials; (2) report the extent to which their findings are affected by SBuMT; and (3) evaluate the robustness of their findings to alternative scenarios of SBuMT.
- Financial firms could (1) avoid the practice of optimizing backtests (i.e., picking the winners while ignoring the losers); (2) implement research surveillance frameworks that record, store, and curate every single research trial that takes place within the organization; and (3) estimate the probability of a false positive objectively controlling for SBuMT.

Topic 6. Data Mining & Machine Learning: Representing & Mining Text

Reading 6.1 *Data Science for Business (ch. 10)*

Provost, F. and T. Fawcett. (2019). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapter 10.

Ch. 10. Representing and Mining Text.

Keywords

Linguistic structure (p. 252)
Text is often referred to as "unstructured" data that does not have the sort of structure that we normally expect for data: tables of records with fields having fixed meanings, as well as links between the tables. Text of course has plenty of structure, but it is linguistic structure – intended for human consumption, not for computers.

Dirty data (p. 252)
As data, text is relatively dirty. People write ungrammatically, they misspell words, they run words together, they abbreviate unpredictably, and punctuate randomly. Even when flawlessly expressed text may contain synonyms and homographs. Terminology and abbreviations in one domain might be meaningless in another domain. Context is important. Text must undergo a good amount of preprocessing before it can be used as input to a data mining algorithm.

Document (p. 253)
A document is one piece of text, no matter how large or small. A document could be a single sentence or a 100 page report, or anything in between.

Token (p. 253)
A document is composed of individual tokens or terms, which can be just words, or n-grams.

Terms (p. 253)
Tokens.

Corpus (p. 253)
A collection of documents.

Bag of words (p. 254)
The approach is to treat every document as just a collection of individual words, ignoring grammar, word order, sentence structure, and (usually) punctuation.

Term frequency (p. 254)
How many times a word is used in a document

Stemmed (p. 255)
Suffixes removed, so that similar verbs are all reduced to a common root term. Noun plurals are transformed to the singular forms.

Stopwords (p. 255)
Very common word.

Inverse document frequency (p. 256)
This sparseness of a term. The fewer documents in which a term occurs, the more significant it likely is to be to the documents is does occur in.

n-grams (p. 265)
Include sequences of adjacent words as terms, to preserve some information about word order in the representation.

Latent information model (p. 268)
Topic models are a type of latent information model. Latent information as a type of intermediate, unobserved layer of information inserted between the inputs and outputs. In the case of text, words map to topics (unobserved) and topics map to documents.

Information triage (p. 276)
Recommending news stories based on their future effects on stock prices.

6.1.1 Broad issues involved in mining text

A. Explain why text is "dirty" which makes mining text is difficult.

People write ungrammatically, they misspell words, they run words together, they abbreviate unpredictably, and punctuate randomly. Even when flawlessly expressed text may contain synonyms (multiple words with the same meaning) and homographs (one spelling shared among multiple words with different meanings). Terminology and abbreviations in one domain might be meaningless in another domain. Because text is intended for communication between people, context is important. Text must undergo a good amount of preprocessing before it can be used as input to a data mining algorithm.

6.1.2 Text representation

A. Understand the meaning of terms when used in the field of information retrieval.

A document is one piece of text, no matter how large or small. A document is composed of individual tokens or terms, which can just words or n-grams. Corpus is a collection of documents.

B. Describe the "bag of words"approach including the following steps:

- **Measuring term frequency (TF).** Number of times each term occurs in a document.
- **Measuring sparseness: inverse document frequency (IDF).** The boost a term gets for being rare. $\text{IDF}(t) = 1 + \log(\frac{\text{Total number of documents}}{\text{Number of documents containing t}})$.
- **Combining them: TFIDF**. A very popular representation for text. The TFIDF value of a term t in a given document d is: $\text{TFIDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$. Note that the TFIDF value is specific to a single document (d) whereas IDF depends on the entire corpus.

C. Apply appropriate methods to search an example set of documents.

1. First, basic stemming is applied on the set of documents. Next, stopwords are removed, and the words are normalized with respect to document length. These values would be used as the Term Frequency (TF) feature values, or to generate the full TFIDF representation by multiplying each term's TF value by its IDF value and renormalized. This is the feature vector representation of each "document"
2. Translate the search query to its TFIDF representation.
3. Compute the similarity, using a distance metric such as Cosine Similarity, of our query term to each document.

D. Express entropy in terms of the IDF.

We can express entropy as the expected value of IDF(t) and IDF(not t) based on the probability of its occurrence in the corpus: $\text{entropy}(t) = p \cdot \text{IDF}(t) + (1 - p)\text{IDF}(\text{not } t)$.

6.1.3 Additional text representation approaches beyond "bag of words"

A. Describe N-gram sequences.

Include sequences of adjacent words as terms, to preserve some information about word order in the representation. Adjacent pairs are commonly called bi-grams. N-grams are useful when particular phrases are significant but their component words may not be. The main disadvantage of n-grams is that they greatly increase the size of the feature set.

B. Describe named entity extraction.

Preprocessing component that knows when word sequences constitute proper names, to process raw text and extract phrases annotated with terms like person or organization .

C. Describe topic models.

The main idea of a topic layer is first to model the set of topics in a corpus separately. General methods for creating topic models include matrix factorization methods, such as Latent Semantic Indexing and Probabilistic Topic Models, such as Latent Dirichlet Allocation. In topic modeling, the terms associated with the topic, and any term weights, are learned by the topic modeling process. As with clusters, the

topics emerge from statistical regularities in the data. The final document classifier is defined in terms of these intermediate topics rather than words.

6.1.4 Mining news stories to predict stock price movement

A. Describe how a given task, such as recommending a news story that is likely to result in a significant change in a stock's price, must be formulated into a problem with simplifying assumptions.

1. It is difficult to predict the effect of news far in advance. With many stocks, news arrives fairly often and the market responds quickly.
2. It is difficult to predict exactly what the stock price will be. Instead, we will be satisfied with the direction of movement
3. It is difficult to predict small changes in stock price, so instead we'll predict relatively large changes.
4. It is difficult to associate a specific piece of news with a price change: how do you decide which of today's thousands of stories are relevant?

B. Describe required considerations for data preprocessing.

1. Many events occur outside of trading hours, and fluctuations near the opening of trading can be erratic. For this reason, instead of measuring the opening price at the opening bell (9:30 am EST) we measure it at 10:00 am, and track the difference between the day's prices at 4 pm and 10 am.
2. The stories are pre-tagged with stocks, but not perfectly. It tends to be overly permissive, such that stories are included in the news feed of stocks that were not actually referenced in the story. Because we want a fairly tight association of a story with the stock(s) it might affect, we reject any stories mentioning more than two stocks. This gets rid of many stories that are just summaries and news aggregations.
3. Reduce each story to a TFIDF representation. In particular, each word was case-normalized and stemmed, and stopwords were removed. Finally, we created n-grams up to two, such that every individual term and pair of adjacent terms were used to represent each story.
4. Each story is tagged with a label (change or no change) based on the associated stock(s) price movement. 25% of the stories were followed by a significant price change (surge or plunge) to the stocks involved, and 75% were not.

C. Choose and discuss appropriate methods for analyzing the results.

- The purpose of news recommendation (answering "which stories lead to substantial stock price changes?") is left open, for which exact costs and benefits of decisions were not specified. Expected value calculations and profit graphs aren't really appropriate here
- To get a sense of how well this problem of predictability can be solved, the ROC curves from ten-fold cross-validation of three sample classifiers are shown: Logistic Regression, Naive Bayes, and a Classification Tree.

- The lift curves of these three classifiers, again averaged from ten-fold cross-validation, shows the lift in precision we would get if we used the model to score and order the news stories.
- Look at the important terms found, with high information gains.

Reading 6.2 *Naive Bayes and Sentiment Classification*

Jurafsky, D. and J. Martihttps. (2018). Chapter 4. Naive Bayes and Sentiment Classification, In Speech and Language Processing.

Keywords

Sentiment analysis (p. 1)
The extraction of sentiment, the positive or negative orientation that a writer expresses toward some object.

Probabilistic classifier (p. 2)
A probabilistic classifier is capable of mapping from a new document to its correct class, and additionally will tell us the probability of the observation being in the class.

Generative classifier (p. 2)
Generative classifiers like naive Bayes build a model of how a class could generate some input data. Given an observation, they return the class most likely to have generated the observation.

Discriminative classifier (p. 2)
Discriminative classifiers like logistic regression instead learn what features from the input are most useful to discriminate between the different possible classes.

Linear classifier (p. 5)
Classifiers that use a linear combination of the inputs to make a classification decision – like naive Bayes and also logistic regression – are called linear classifiers.

Sentiment lexicon (p. 9)
lists of words that are preannotated with positive or negative sentiment. Four popular lexicons are the General Inquirer (1966), LIWC (2007), the opinion lexicon of Hu and Liu (2004) and the MPQA Subjectivity Lexicon (2005).

Gold labels (p.11)
The human-defined labels for each document that we are trying to match

Precision (p.12)
Precision measures the percentage of the items that the system detected (i.e., the system labeled as positive) that are in fact positive (i.e., are positive according to the human gold labels).

Recall (p.12)
Recall measures the percentage of items actually present in the input that were correctly identified by the system.

F-measure (p.13)
F-measure comes from a weighted harmonic mean of precision and recall. The harmonic mean of a set of numbers is the reciprocal of the arithmetic mean of reciprocals.

Macroaveraging (p.13)
Compute the performance for each class, and then average over classes.

Microaverageing (p.13)
Collect the decisions for all classes into a single contingency table, and then compute precision and recall from that table.

6.2.1 Classification

A. Describe common applications of classifying text.

- Sentiment analysis, the extraction of sentiment, the positive or negative orientation that a writer expresses toward some object.
- Spam detection, the binary classification task of assigning an email to one of the two classes spam or not-spam.
- The language it's written in.
- Determining a text's author (authorship attribution), or author characteristics like gender, age, and native language
- Assigning a library subject category or topic label to a text.

B. Describe tasks often involved in classifying text.

- Period disambiguation (deciding if a period is the end of a sentence or part of a word),
- word tokenization (deciding if a character should be a word boundary).
- Language modeling can be viewed as classification: each word can be thought of as a class, and so predicting the next word is classifying the context-so-far into a class for each next word.
- A part-of-speech tagger classifies each occurrence of a word in a sentence as, e.g., a noun or a verb.

C. Compare alternative methods of classification.

Classifiers are capable of mapping from a new document to its correct class. A probabilistic classifier additionally will tell us the probability of the observation being in the class. One method for classifying text is to use hand-written rules, but rules can be fragile as situations or data change over time, and for some tasks humans aren't necessarily good at coming up with the rules. Two ways machine learning algorithms of doing classification are: Generative classifiers like naive Bayes build a model of how a class could generate some input data – given an observation, they return the class most likely to have generated the observation; and Discriminative classifiers like logistic regression which instead learn what features from the input are most useful to discriminate between the different possible classes. Classifiers that use a linear combination of the inputs to make a classification decision – like naive Bayes and also logistic regression – are called linear classifiers.

6.2.2 Math behind Naive Bayes classifiers

A. Explain why in the context of classifying a document the denominator can be dropped from Bayes Rule.

The denominator $P(d)$, the probability of document d, doesn't change for each class. We are always asking about the most likely class for the same document d, which must have the same probability $P(d)$, which doesn't change by dropping the denominator.

B. Explain the bag of words and naive Bayes assumptions.

Bag of words assumes position doesn't matter, hence the features only encode word identity and not position. Naive Bayes assumption is the conditional independence assumption that the probabilities $\mathrm{P}(f_i|c)$ of each feature f_i are independent given the class c and hence can be 'naively' multiplied as follows: $P(f_1, f_2, ..., f_n|c) = P(f_1|c) \cdot P(f_2|c) \cdot ... \cdot P(f_n|c)$

C. Explain why Naive Bayes calculations are done in log space so that the predicted class is a linear function of input features.

To avoid underflow and increase speed

6.2.3 Training the Naive Bayes classifiers

A. Explain why Laplace smoothing is commonly used in Bayes text categorization.

Since naive Bayes naively multiplies all the feature likelihoods together, zero probabilities in the likelihood term for any class will cause the probability of the class to be zero, no matter the other evidence. The simplest solution is the add-one (Laplace) smoothing.

B. Explain how stop words and unknown words are treated during training.

The solution for such unknown words is to ignore them – remove them from the test document and not include any probability for them at all. some systems choose to completely ignore stop words: very frequent words like *the* and *a*. In most text classification applications, however, using a stop word list doesntt improve performance, and so it is more common to make use of the entire vocabulary

C. Calculate the prior probabilities of two classes given a training set categorized into two classes.

Let N_c be the number of documents in our training data with class c and N_{doc} be the total number of documents. Then: $P(c) = N_c/N_{doc}$

D. Determine the class that a test sentence belongs to using the Naive Bayes classifier.

$\hat{c} = \mathrm{argmax}_{c \in C} \log P(c) \sum_{i \in n} \log P(w_i|c)$, where the maximum likelihood estimate of the probability of frequency of word w_i is $P(w_i|c) = \frac{\mathrm{count}(w_i,c)}{\sum_{w \in V} \mathrm{count}(w,c)}$

6.2.4 Optimizing for sentiment analysis

A. Explain how binary multinomial Naive Bayes differs from Naive Bayes.

The binary multinomial naive Bayes (or binary NB) variant clips the word counts in each document at 1.

B. Explain why binary multinomial Naive Bayes (also called binary NB) might improve results.

For sentiment classification and a number of other text classification tasks, whether a word occurs or not seems to matter more than its frequency. Thus it often improves performance to clip the word counts in each document at 1.

C. Describe two other methods (besides binary NB) that can improve the results of sentiment analysis.

1. A very simple baseline commonly used to deal with negation is during text normalization to prepend the prefix $NOT\_$ to every word after a token of logical negation (n't, not, no, never) until the next punctuation mark.
2. In some situations we might have insufficient labeled training data to train accurate naive Bayes classifiers using all words in the training set to estimate positive and negative sentiment. In such cases we can instead derive the positive and negative word features from sentiment lexicons, lists of words that are pre-annotated with positive or negative sentiment

6.2.5 Evaluation of sentiment analysis results

A. Calculate precision and recall statistics given system output and gold standard label results.

Precision measures the percentage of the items that the system detected (i.e., the system labeled as positive) that are in fact positive (i.e., are positive according to the human gold labels): Precision = (true positives) / (true positives + false positives). Recall measures the percentage of items actually present in the input that were correctly identified by the system: Recall = (true positives) / (true positives + false negatives)

B. Describe the F-measure and various methods of weighting precision and recall.

$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P+R}$. The β parameter differentially weights the importance of recall and precision, based perhaps on the needs of an application. Values of $\beta > 1$ favor recall, while values of $\beta < 1$ favor precision. When $\beta = 1$, precision and recall are equally balanced; this is the most frequently used metric, and is called F_1.

C. Compare macroaveraging and microaverageing approaches to evaluating the categorization performance of multiple classes.

In order to derive a single metric that tells us how well the system is doing, we can combine these values in two ways. In macroaveraging, we compute the performance for each class, and then average over classes.

In microaveraging, we collect the decisions for all classes into a single contingency table, and then compute precision and recall from that table.

D. Compare 10-fold cross-validation with bootstrap tests.

Cross-validation uses all our data both for training and test We randomly choose a training and test set division of our data, train our classifier, and then compute the error rate on the test set. Then we repeat with a different randomly selected training set and test set. We do this sampling process 10 times and average these 10 runs to get an average error rate.

We may want to compare, with statistical significance testing, a new version of the system to the unaugmented version. or our algorithm to a previously published one to know which is better, or simply to compare the performance of two classifiers. The standard approach to computing $p - value(x)$ in natural language processing is to use non-parametric tests like the bootstrap test. Bootstrapping refers to repeatedly drawing large numbers of smaller samples with replacement (called bootstrap samples) from an original larger sample. The intuition of the bootstrap test is that we can create many virtual test sets from an observed test set by repeatedly sampling from it. The method only makes the assumption that the sample is representative of the population.

Topic 7. Big Data, Data Mining & Machine Learning: Ethical & Privacy Issues

Reading 7.1 *Business Ethics and Big Data*

Institute of Business Ethics. (2016, June). Business Ethics and Big Data (IBE Issue 52). London, U.K.

7.1.1 Keywords

Data trust deficit (p. 2)
The public trust in companies to use data appropriately is lower than trust generally

Veracity (p. 6)
Trustworthiness and integrity of data.

7.1.2 Big data for business

A. Discuss the potential and concerns of big data for business.

Potential positive impact of Big Data:

1. As a source of innovation – An example of innovation driven by 'Big Data sharing' is provided by the pharmaceutical company GlaxoSmithKline, which developed an online portal where researchers can request access to the underlying data of these trials help further their own experiments
2. As a source of development – The digital traces from an exponential increase in the use of digital communication technologies in low and middle-income countries have been identified by policymakers and researchers as a potential solution to the lack of reliable local statistical data and could help inform policy and interventions

On the other hand, the increasingly prominent role of the IoT in everyone's lives has made it possible for companies to collect data in ways that might not be fully understood by users. The public feel constantly under the scrutiny of a 'Big Brother' that serves the economic interests of businesses and over which they have little or no control. Public trust in companies to use data appropriately is lower than trust generally, which can negatively affect the reputation of companies or whole industries.

B. Explain how the new term "data trust deficit" developed.

The public has developed a greater awareness and sensitivity of the increasingly prominent role of the IoT in everyone's lives. The IoT, in particular, has made it possible for companies to collect data in ways that might not be fully understood by users (e.g. from mobile phone calls or public transport travel passes). As a result, some feel constantly under the scrutiny of a 'Big Brother' that serves the economic interests of businesses and over which they have little or no control. This perception has produced a 'data trust deficit': the public trust in companies to use data appropriately is lower than trust generally.

7.1.3 Ethical issues

A. List five methods of protecting human rights in the 'Era of Big Data.'

1. Stop High-Tech Profiling.
2. Ensure fairness in Automated Decisions.
3. Respect the Law.
4. Enhance Individual Control of Personal Information.
5. Protect People from Inaccurate Data.

B. Provide an example of a concern for each of three main areas of privacy issues: customer profiling, group privacy and data security.

1. Customer profiling. The information collected by some organisations allows them to create profiles of their customers. While this is predominantly used for marketing purposes, it can also be used in ways that determine personal attributes. One such example involved the retailing company Target in the US, which was able to predict specific events in the life of its consumers, such as the birth of a child, based on changing spending habits to target certain products. One person targeted in this way was a teenage girl, whose family were unaware of her pregnancy and who found out as a consequence of Target's approach.
2. Group Privacy. When used to analyse large groups of people, the information that Big Data can reveal may be hugely beneficial, such as tracking the spread of a disease more quickly, or bringing relief to a disaster zone more effectively. However, datasets could easily be acquired by companies with ethically questionable marketing strategies, or political groups wanting to use the information to target specific sets of people. These privacy issues can only be magnified by the spread of the IoT particularly in low and middle income countries.
3. Data security. Threats might be both external and internal, including the risk of misuse by employees of the company's information. For example, in July 2015, a group called 'The Impact Team' hacked the database of Ashley Madison, a dating website for extramarital affairs. The group copied personal information about the site's user base and threatened to release users' names and personally identifying information if Ashley Madison was not immediately shutdown.

C. Discuss what constitutes informed consent.

In the UK,the Data Protection Act 1998 states that consent must be obtained from individuals before their data can be used for research or commercial purposes. The primary method for obtaining consent, especially on social media platforms is by asking users to agree to terms and conditions when they register

to use the service. However, this does not necessarily correlate to informed consent, as research has shown that users sign these complicated documents without reading them in order to open their accounts

D. Provide an example of how to improve the veracity of data.

When veracity, which refers to the trustworthiness and integrity of of a dataset, can't be guaranteed, some individuals or groups might accidentally be accorded more visibility and thus be favoured, or discriminated against. For example, the city of Boston faced this type of issue when implementing a mobile application that used a smartphone accelerometer and GPS feed to collect data about road conditions such as potholes. However, as some groups of people (e.g. individuals on low income and the elderly) were less likely to own a smartphone or download the app, information from these groups was not being recorded and repair services were therefore being concentrated in wealthier neighbourhoods. The city solved the problem by completing the dataset using other sources not subject to this bias, such as reports from city-roads inspectors and other more traditional channels.

7.1.4 The Ethics Test

A. List six question which Ethics Professionals within an organization using big data can ask themselves.

Do we know how the company uses Big Data and to what extent it is integrated into strategic planning?

1. Do we send a privacy notice when we collect personal data? Is it written in a clear and accessible language which allows users to give a truly informed consent?
2. Does my organisation assess the risks linked to Big Data?
3. Does my organisation have any safeguard mechanisms in place to mitigate these risks?
4. Do we make sure that the tools to manage these risks are effective and measure outcome?
5. Do we conduct appropriate due diligence when sharing or acquiring data from third parties?

Reading 7.2 *Business Ethics and Artificial Intelligence*

Institute of Business Ethics. (2018, January). Business Ethics and Artificial Intelligence (IBE Issue 58). London, U.K.

Keywords

Artificial intelligence (p. 1)
Artificial Intelligence (AI) is a term generally used to describe the simulation of elements of human intelligence processes by machines and computer systems.

Code of ethics (p. 6)
Many organisations include in their code of ethics guidance to support individual decision-making through a decision-making model or guide. This often takes the form of 'questions to ask yourself'.

7.2.1 The nature of and business risks of artificial intelligence (AI)

A. List three main features characterizing artificial intelligence.

1. Learning – the ability to acquire relevant information and the rules for using it;
2. Reasoning – the ability to apply the rules acquired and use them to reach approximate or definite conclusions;
3. Iterative – the ability to change the process on the basis of new information acquired.

B. List three immediate risks of artificial intelligence.

1. Ethics risk: certain applications of the AI systems adopted might lead to ethical lapses;
2. Workforce risk: automation of jobs can lead to a deskilled labour force;
3. Technology risk: black box algorithms can make it difficult to identify tampering and cyber-attacks;
4. Legal risk: data privacy issues, including compliance with GDPR;
5. Algorithmic risk: biased algorithms can lead to a discriminatory impact.

7.2.2 Values that form the cornerstone of an ethical framework of artificial intelligence in business

A. Discuss each of the following as they impact the ethical nature of applications of artificial intelligence in business:

- **Accurate results**: Companies need to ensure that the AI systems they use produce correct, precise and reliable results. To do so algorithms need to be free from biases and systematic errors deriving, for example, from an unfair sampling of a population, or from an estimation process that does not give accurate results.
- **Respect of privacy**: Everyone has the right to the protection of personal data concerning him or her. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified.
- **Transparency and openness**: Traditionally, many organisations do not allow public scrutiny as the underlying programming (the source code) is proprietary. Opening sourcing material in computer science, when appropriate, is an important step. It helps the development community to understand better how AI works and therefore be able to explain it more accurately to the public and the media. This is particularly important as better information within the general public improves trust and prevents unjustified fears.
- **Interpretability of algorithms**: The use of 'black box' algorithms makes it difficult not only to identify when things go wrong, but also to determine who is responsible in the event on any damage or ethical lapse. Interpretable and explainable AI will be essential for business and the public to understand, trust and effectively manage 'intelligent' machines. Organisations that design and use algorithms need to take care in producing models that are as simple as possible, to explain how complex machines work.

- **Fairness to stakeholders**: As AI systems are able to perform tasks, previously undertaken by humans, in a more efficient and reliable way, the workplace is going to change. However, companies have a role to play in ensuring that this transition will be smooth. This means tackling issues such as long term unemployment, social inequality and lack of trust from customers in the way AI is utilised.
- **Integrity and due diligence**: We should ensure that it is used only for its intended purpose, even when there is no means to enforce this. Tt is also necessary to conduct appropriate due diligence on the clients as well to minimise the risk of a potentially dangerous misuse.
- **Control of humans relative to machines**: Much of the public scepticism around the future of AI is fuelled by the fear that humans might lose control over the machines, which would then prevail and possibly wipe out humanity altogether. To have full control over AI systems, it is important that both companies and algorithm designers only work with technology that they fully understand.
- **Impact of a new technology**: Measuring the potential impact that a new technology can have before adopting it, can identify undesired side-effects and consequent ethical risks. Through tests of their the algorithms and AI implementations, companies need to gauge a clear idea of unwanted outcomes and identify what are the ethical risks involved, who is going to be impacted positively or negatively, who is going to bear the costs, and whether there is a valuable and less risky alternative.
- **Accountability assignment**: There should always be a line of responsibility for business actions to establish who has to answer for the consequences. AI systems introduce an additional strand of complexity: who is responsible for the outcome of the decision-making process of an artificial agent? This is compounded by AI development being largely outsourced by companies rather than developed in-house.
- **Learning about how the AI technologies work**: Employees and other stakeholders need to beempowered to take personal responsibility for the consequences of their use of AI and they need to be provided with the skills to do so, both technical and also an understanding of the potential ethical implications that it can have. It is important that companies improve their communications around AI, so that people feel that they are part of its development and not its passive recipients, or even victims. Ensuring business leaders are informed about these technologies and how they work is essential to prevent that they are unintentionally misused. It is important that businesses engage with external stakeholders as well, including media reporters and the general public.

7.2.3 The role of business decision makers

A. List five measures organizations can take to minimize the risk of ethical lapses due to improper use of AI technologies.

1. Design new and more detailed decision-making tools for meta-decisions
2. Engage with third parties for the design of AI algorithms only if they commit to similar ethical standards
3. Establish a multi-disciplinary Ethics Research Unit to examine the implications of AI research and potential applications;
4. Introduce ‘ethics tests’ for AI machines, where they are presented with an ethical dilemma
5. Empower people through specific training courses and communication campaigns in order to enable them to use AI systems efficiently, effectively and ethically

B. List some questions addressing the use of AI that could be included in a code of ethics.

1. What is the purpose of our job and what AI do we need to achieve it?
2. Do we understand how these systems work? Are we in control of this technology?
3. Who benefits and who carries the risks related to the adoption of the new technology?
4. Who bears the costs for it? Would it be considered fair if it became widely known?
5. What are the ethical dimensions and what values are at stake?
6. What might be the unexpected consequences?
7. Do we have other options that are less risky?
8. What is the governance process for introducing AI?
9. Who is responsible for AI?
10. How is the impact of AI to be monitored?
11. Have the risks of its usage been considered?

Reading 7.3 *Beyond Law: Ethical Culture and GDPR*

Institute of Business Ethics. (2018, May). Beyond Law: Ethical Culture and GDPR (IBE Issue 62). London, U.K.

Keywords

General Data Protection Regulation (p. 1)
Replacing the existing Data Protection Act, the General Data Protection Regulation (GDPR) recognises the need to update legislation appropriate to the digital age, thereby setting out articles for the legal capture, use and transition of personal data through organisations, and seeks to give back control to the individual over how organisations use their personal data and to harmonise privacy laws across Europe.

People risk (p. 3)
GPDR threats may arrive internally from within an organisation: a poorly trained individual with a desire or capacity to act against the values; careless or unaware employees.

7.3.1 General Data Protection Regulation (GDPR)

A. Describe the primary purpose of the GDPR.

In the context of rapid technological developments, globalisation, and increased cross-border flows of data, the GDPR 'seeks to harmonise the protection of fundamental rights and freedoms of natural persons in respect of processing activities and to ensure the free flow of personal data between Member States'. The two key drivers for GDPR are (i) give control of personal data back to the data subjects themselves, and (ii) mandating that organisations demonstrate accountability through evidence.

B. Describe the key changes in data protection regulation including the meaning of

- **rights of the individual**. Individuals have the right to access, amend, restrict, withdraw consent and request that their personal data be erased.
- **informed consent**. Requests for consent to be explicit (not implied) and written in clear and easy to understand language, and as easy to withdraw as to give.
- **notification**. There is a mandatory breach notification period of 72 hours organisations to notify the national regulator (ICO for the UK).
- **data portability**. Customers can transfer personal data from one company to another, but only data provided by the customer themselves and in a machine readable format.
- **supervision and enforcement**. Introduction of the 'one stop shop' approach where any national regulator can take or lead action across all member states. Organisations outside the EU but processing data of EU citizens can face sanctions and be subject to individual claims. Similarly, this applies to citizens 'in' the EU but not necessarily an EU citizen. Higher penalties for breaching the regulations.
- **liability**. Data Processors, in additional to the Data Controller, are directly liable.

7.3.2 Separating ethics and compliance

A. Distinguish between two types of threats of personal data breaches.

Breaches do not arrive from external threats (such as hackers, or personal information being exposed to fraud) only. Internal 'people risk' within an organisation should be awarded equal attention. Securing physical systems fully still cannot mitigate a poorly trained individual with a desire or capacity to act against the values. A recent survey showed that the majority of respondence rated 'careless or unaware employees' as the highest or second highest vulnerability in cyber risk.

B. Discuss 'people risk.'

GPDR threats internally from within an organisation. Securing physical systems fully still cannot mitigate a poorly trained individual with a desire or capacity to act against the values. 'Careless or unaware employees' is a high vulnerability in cyber risk.

C. List key questions around the role an ethical culture plays in preventing data breaches.

1. Have we decided at Board level and throughout the organisation what constitutes an appropriate attitude and approach in choosing how to respond to threats and breaches?
2. Is our culture sufficiently built on our values such that each individual knows and understands why and how we approach confidentiality and privacy the way we do? Do they understand the communications standards that must support those?
3. Do we make our decisions and state outcomes with openness, transparency and honesty?
4. Do we have a strategy that will serve to maintain trust with all stakeholders: eg how quickly must we respond to questions or issues, what level of information would we choose to share?

5. What monitoring/testing processes are in place such that if the culture is threatened, it can be identified quickly?
6. Are all individuals within the organisation trained and monitored appropriately?
7. Are individuals reinforced and rewarded through the appraisal system for 'doing the right thing'?
8. Do we have aligned policies and cultural values with our third party stakeholders such that they behave and work in a similar way and will inform us appropriately of any issues?
9. Which employees demonstrate the responsibility to be granted privileged access?
10. If an employee speaks up about a concern and seeks to prevent a breach before it happens, do we have appropriate procedures in place to allow investigations to be carried out such that the organisation does not then commit a breach?
11. If there has been a breach, how do we share lessons learned to help support others in identifying concerns?

7.3.3 Maintaining privacy of personal data

A. Describe how an organization must build awareness regarding employees' roles in protecting data.

It is particularly important that organisations have in place a strategy for allowing transparency up front regarding their processes and procedures. It is crucial that all individuals are aware of what their role might be in protecting the value of this data. Establishing the boundaries and clarity of processes and procedures early on, to all parties concerned, is vital.

B. Discuss liability if the 72-hour notification deadline is missed.

If the 72 hour notification deadline isn't met, then data controller is responsble, even if the data processor is a third party where there is a breach.

7.3.4 The GDPR Embedding Wheel

A. Describe how the tone from the top can help foster an ethical culture and compliance with the GDPR.

Business leaders must have informed those with such responsibilities in the organisation regarding the appropriate level of openness. This calls for increased transparency in senior level decision making: decisions taken at this level and how they flow through the organisation and ensure that not only security, HR, legal and risk managers are aware, but all employees in all parts of the business. Personal data must be the business of everyone, and an organisational approach to attitude, training and communication is crucial in recognising its value and protecting it.

B. Describe how establishing the boundaries and standards can help foster an ethical culture and compliance with the GDPR.

The policies that govern the process must provide clarity on the organisation's approach: what should people expect within the process and what can they do if they are not satisfied. These documents should provide detail that allows individuals to feel confident that their personal data is managed with consistency and rigour.

C. Describe how communication and training can help foster an ethical culture and compliance with the GDPR.

In the investigations process, communication is required for all parties involved to build trust: The earlier people know what information will be captured and where from, how will it be stored and transferred, and what will be available to third parties at different times, the stronger the outcomes. Training makes the policies and guidance easily accessible. Cultural workshops can support understanding of those and address more specific areas that require awareness. Investigator training is important in order to (1) increase the efficacy of the process, and (2) ensure expectations of GDPR are met appropriately, reducing the opportunity for individuals to create risks by making decisions outside the parameters of GDPR.

D. Describe how choice of the individual can help foster or hinder an ethical culture and compliance with the GDPR.

Employees should be aware of how they make choices and their responsibilities around those, since individuals are not just the victims of data breaches only – they can be active participants in the process. Organisations in their ethics programme should discuss and explore with employees the justifications and ways that they think, as well as providing a scenario and arriving at the most appropriate choice together

E. Describe how monitoring outcomes can help foster an ethical culture and compliance with the GDPR.

Processes for monitoring outcomes and celebrating good practice can make employees feel supported to do the right thing.

Topic 8. Big Data & Machine Learning in the Financial Industry

Reading 8.1 *Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications*

Financial Stability Board. (2017) Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications.

Keywords

Big data (p. 4)
The storage and analysis of large and/or complicated data sets using a variety of techniques including AI.

Artificial intelligence (p. 4)
the theory and development of computer systems able to perform tasks that traditionally have required human intelligence

Machine learning (ML) (p. 4)
A method of designing a sequence of actions to solve a problem, known as algorithms, which optimise automatically through experience and with limited or no human intervention. A sub-category of AI.

Supervised learning (p. 5)
In 'supervised learning', the algorithm is fed a set of 'training' data that contains labels on some portion of the observations. The algorithm will 'learn' a general rule of classification that it will use to predict the labels for the remaining observations in the data set.

Unsupervised learning (p. 5)
'Unsupervised learning' refers to situations where the data provided to the algorithm does not contain labels. The algorithm is asked to detect patterns in the data by identifying clusters of observations that depend on similar underlying characteristics.

Reinforcement learning (p. 5)
'Reinforcement learning' falls in between supervised and unsupervised learning. In this case, the algorithm is fed an unlabelled set of data, chooses an action for each data point, and receives feedback (perhaps from a human) that helps the algorithm learn.

Deep learning (p. 5)
'Deep learning' is a form of machine learning that uses algorithms that work in 'layers' inspired by the structure and function of the brain. Deep learning algorithms, whose structure are called artificial neural networks, can be used for supervised, unsupervised, or reinforcement learning.

Natural language processing (p. 5)
NLP allows computers to 'read' and produce written t ext or, when combined with voice recognition, to read and produce spoken language.

Sentiment indicators (p. 10)
Social media data analytics companies use AI and machine learning techniques to provide 'sentiment indicators' – 'bullishness' or 'bearishness' – to a number of financial services players.

Trading signals (p. 11)
Machine learning can help trading firms in quickly scanning and making decisions based on more sources of information than a human can.

Fraud detection (p. 11)
Identifying those transactins that are fraudulent and those that are not fraudulent.

RegTech (p. 11)
Uses of AI and machine learning by financial institutions for regulatory compliance

InsurTech (p. 13)
Insurance-related technology, sometimes called 'InsurTech,' often relies on analysis of big data to drive pricing, and lower costs and improve profitability.

Chatbots (p. 14)
Virtual assistants that help customers transact or solve problems. These automated programmes use NLP to interact with clients in natural language (by text or voice), and use machine learning algorithms to improve over time.

Know your customer (KYC) (p. 20)
Knowing the identity of customers ('know your customer' or KYC) through, for example, performing identity and background pre-checks, evaluating whether images in identifying documents match one another, calculating risk scores on which firms determine which individuals or applications need to receive additional scrutiny, and ongoing periodic checks based on public and other data sources

SupTech (p. 21)
uses of AI and machine learning by public authorities for supervision

Auditability (p.33)
The interpretability of AI and machine learning methods which is often lacking.

Fintech (p. 35)
Technologically enabled financial innovation that could result in new business models, applications, processes, or products with an associated material effect on financial markets and institutions and the provision of financial services.

Rob-advisors (p.35)
Applications that combine digital interfaces and algorithms, and can also include machine learning, in order to provide services ranging from automated financial recommendations to contract brokering to portfolio management to their clients, without or with very limited human intervention.

Tonality analysis (p.36)
A method to gauge the negativity of a piece of text by counting terms with a negative connotation.

8.1.1 Regulatory and supervisory issues around FinTech

A. Identify factors that may contribute to making the markets more efficient.

The more efficient processing of information, for example in credit decisions, financial markets, insurance contracts, and customer interaction, may contribute to a more efficient financial system. The RegTech and SupTech applications of AI and machine learning can help improve regulatory compliance and increase supervisory effectiveness.

B. Identify factors that may contribute to increases in third party dependencies among financial institutions.

Network effects and scalability of new technologies may in the future give rise to third-party dependencies, leading to the emergence of new systemically important players that could fall outside the regulatory perimeter.

C. Explain why unexpected forms of interconnectedness among institutions could be created.

Applications of AI and machine learning could result in new and unexpected forms of interconnectedness between financial markets and institutions, for instance based on the use by various institutions of previously unrelated data sources.

D. Explain why new forms of macro-level risks could emerge.

The lack of interpretability or "auditability" of AI and machine learning methods could become a macro-level risk. Similarly, a widespread use of opaque models may result in unintended consequences.

E. Explain why new risk management tools and techniques may be required.

It will be important to assess uses of AI and machine learning in view of their risks, including adherence to relevant protocols on data privacy, conduct risks, and cybersecurity. Adequate testing and 'training' of tools with unbiased data and feedback mechanisms is important to ensure applications do what they are intended to do.

8.1.2 Relationship between AI, machine learning and big data, and algorithms

A. Describe the two recent developments that have contributed to increased interest in AI.

Recent increases in computing power coupled with increases in the availability and quantity of data have resulted in a resurgence of interest in potential applications of artificial intelligence.

B. Factors contributing to making the markets more efficient.

see above.

C. Describe the relationship between AI, machine learning, and three algorithms of Figure 1.

This report defines AI as the theory and development of computer systems able to perform tasks that traditionally have required human intelligence. AI is a broad field, of which 'machine learning' is a sub-category. Machine learning may be defined as a method of designing a sequence of actions to solve a problem, known as algorithms, which optimise automatically through experience and with limited or no human intervention. These techniques can be used to find patterns in large amounts of data (big data analytics) from increasingly diverse and innovative sources.

8.1.3 Categories of machine learning algorithms

A. Define four categories of machine learning algorithms based on the degree of human intervention.

1. In 'supervised learning', the algorithm is fed a set of 'training' data that contains labels on some portion of the observations. The algorithm will 'learn' a general rule of classification that it will use to predict the labels for the remaining observations in the data set.
2. 'Unsupervised learning' refers to situations where the data provided to the algorithm does not contain labels. The algorithm is asked to detect patterns in the data by identifying clusters of observations that depend on similar underlying characteristics.
3. 'Reinforcement learning' falls in between supervised and unsupervised learning. In this case, the algorithm is fed an unlabelled set of data, chooses an action for each data point, and receives feedback (perhaps from a human) that helps the algorithm learn.
4. 'Deep learning' is a form of machine learning that uses algorithms that work in 'layers' inspired by the structure and function of the brain. Deep learning algorithms, whose structure are called artificial neural networks, can be used for supervised, unsupervised, or reinforcement learning.

B. Describe the role of machine learning algorithms in determining causality vs correlation.

It is important to note what machine learning cannot do, such as determining causality. Generally speaking, machine learning algorithms are used to identify patterns that are correlated with other events or patterns. The patterns that machine learning identifies are merely correlations, some of which are unrecognisable to the human eye. However, AI and machine learning applications are being used increasingly by economists and others to help understand complex relationships, along with other tools and domain expertise.

C. Define 'augmented intelligence'.

Many applications tend more toward 'augmented intelligence,' or an augmentation of human capabilities, rather than a replacement of humans or an attempt to fully replicate human intelligence.

D. Explain the limitations of machine learning algorithms in determining causality and correlations.

It is important to note what machine learning cannot do, such as determining causality. Generally speaking, machine learning algorithms are used to identify patterns that are correlated with other events or patterns. The patterns that machine learning identifies are merely correlations, some of which are unrecognisable to the human eye.

8.1.4 Drivers of the growth in use of fintech and adaptation of artificial intelligence

A. Discuss the supply factors related to advances in computing technologies and changes in the financial sector.

- Financial market participants have benefitted from the availability of AI and machine learning tools developed for applications in other fields. These include availability of computing power owing to faster processor speeds, lower hardware costs, and better access to computing power via cloud services. Similarly, there is cheaper storage, parsing, and analysis of data through the availability of targeted databases, software, and algorithms. There has also been a rapid growth of datasets for learning and prediction owing to increased digitisation and the adoption of web-based services.
- A variety of technological developments in the financial sector have contributed to the creation of infrastructure and data sets. The proliferation of electronic trading platforms has been accompanied by an increase in the availability of high quality market data in structured formats. In some countries, market regulators allow publicly traded firms to use social media for public announcements. In addition to making digitised financial data available for machine learning, the computerisation of markets has made it possible for AI algorithms to interact directly with markets. Meanwhile, retail credit scoring systems have become more common. With the growth of data in financial markets as well as datasets – such as online search trends, viewership patterns and social media that contain financial information about markets and consumers – there are even more data sources that can be explored and mined in the financial sector.

B. Discuss the demand factors related to search for higher profits, increased competition and changes in the regulatory environment.

- Opportunities for cost reduction, risk management gains, and productivity improvements have encouraged adoption, as they all can contribute to greater profitability.
- In many cases these factors may also drive 'arms races' in which market participants increasingly find it necessary to keep up with their competitors' adoption of AI and machine learning, including for reputational reasons (hype).
- There is also demand due to regulatory compliance. New regulations have increased the need for efficient regulatory compliance, which has pushed banks to automate and adopt new analytical tools that can include use of AI and machine learning. Financial institutions are seeking cost effective means of complying with regulatory requirements. Correspondingly, supervisory agencies are faced with responsibility for evaluating larger, more complex and faster-growing datasets, necessitating more powerful analytical tools to better monitor the financial sector.

8.1.5 Use cases of artificial intelligence and machine learning in financial sector

A. Describe customer-focused uses, such as credit scoring, insurance and client-facing chatbots.

Large-scale client data are fed into new algorithms to assess credit quality and thus to price loan contracts. Similarly, such data can help assess risks for selling and pricing insurance policies. Finally, client interactions may increasingly be carried out by AI interfaces with so-called 'chatbots,' or virtual assistance programs that interact with users in natural language.

B. Describe operations-focused uses, such as optimal allocation of capital, risk management modeling, market impact analysis.

Financial institutions can use AI and machine learning tools for a number of operational (or back-office) applications: (i) capital optimisation by banks (maximisation of profits given scarce capital); (ii) model risk management (back-testing and model validation); and (iii) market impact analysis (modelling of trading out of big positions).

C. Describe portfolio management and trading uses.

- Trading firms (sell-side) are looking to AI and machine learning to use data to improve their ability to sell to clients, more pro-actively manage risk exposures and help compliance with trading regulations.
- In portfolio management, AI and machine learning tools are being used to identify new signals on price movements and to make more effective use of the vast amount of available data and market research than with current models. Among asset managers, machine learning is used most extensively by systematic ('quant') funds, most of which are hedge funds. Specialist firms are making available to asset managers machine learning tools to gain insight from the vast volume of news and market research available

D. Describe regulatory compliance and supervision uses by financial institutions, central banks macroprudential authorities, and market regulators.

- For analysing unstructured data, RegTech can use machine learning combined with NLP. Besides being applied to the monitoring of behaviour and communication of traders for transparency and market conduct, machine learning together with NLP can interpret data inputs such as e-mails, spoken word, instant messaging, documents, and metadata. The KYC process, where machine learning is increasingly used, has been costly, laborious, and highly duplicative across many services and institutions.
- AI and machine learning methods may help to improve macroprudential surveillance by automating macroprudential analysis and data quality assurance. Machine learning can be applied to systemic risk identification and risk propagation channels. Specifically, NLP tools may help authorities to detect, measure, predict, and anticipate, among other things, market volatility, liquidity risks, financial stress, housing prices, and unemployment
- Some regulators are using AI for fraud and AML/CFT detection. Market regulators can also use these techniques for disclosure and risk assessment.

8.1.6 The micro-financial analysis of artificial intelligence and machine learning uses.

A. Describe the uses of artificial intelligence and machine learning in information gathering and processing their potential impacts on financial markets.

- May enable certain market participants to collect and analyse information on a greater scale. In particular, these tools may help market participants to understand the relationship between the formulation of market prices and various factors, such as in sentiment analysis. This could reduce information asymmetries and thus contribute to the efficiency and stability of markets.
- May lower market participants' trading costs. Moreover, AI and machine learning may enable them to adjust their trading and investment strategies in accordance with a changing environment in a swift manner, thus improving price discovery and reducing overall transaction costs in the system.

B. Describe the uses of artificial intelligence and machine learning in improving efficiency of financial institutions.

- May enhance machine-based processing of various operations in financial institutions, thus increasing revenues and reducing costs. For example, if AI and machine learning help to identify customers' needs and better target or tailor products to profitable customers, financial institutions could more efficiently allocate resources toward serving those customers that account for substantial fees or have the potential for future growth. Automating routine business processes may allow for lower operating costs.
- Can be used for risk management through earlier and more accurate estimation of risks. Tools that mitigate tail risks could be especially beneficial for the overall system. Also, AI and machine learning could be used for anticipating and detecting fraud, suspicious transactions, default, and the risk of cyber-attacks. But AI and machine learning based tools might also miss new types of risks and events because they could potentially 'overtrain' on past events, and remain untested at addressing risk under shifting financial conditions.
- The data intensity and open-source character of research in AI and machine learning may encourage collaboration between financial institutions and other industries, such as e-commerce and sharing economy businesses

C. Describe the uses of artificial intelligence and machine learning by financial institutions and their potential impacts on customers and investors.

- Consumers and investors could enjoy lower fees and borrowing costs if AI and machine learning reduce the costs for various financial services.
- Consumers and investors could have wider access to financial services. Moreover, AI and machine learning, through advanced credit scoring for FinTech lending, might make wider sources of funds available to consumers and small and medium enterprises (SMEs).
- AI and machine learning could facilitate more 'customised' and 'personalised' financial services through big data analytics. Nonetheless, the use of consumers' data may entail issues of data privacy, information security and governance structure for consumer and investor protection.

8.1.7 The macro-financial analysis of artificial intelligence and machine learning uses.

A. Describe economic growth and enhanced economic efficiency that could result from the applications of artificial intelligence and machine learning to financial services.

- Enhancing the efficiency of financial services: more efficient risk management of individual banks' loan portfolio and insurers' liabilities may benefit the aggregate system. AI and machine learning could help process information on the fundamental value of assets, thus allocating funds to investors and projects more effectively. Moreover, if AI and machine learning increase the speed and reduce the costs of payment and settlement transactions, this may stimulate transactions for real economic activities.
- Were AI and machine learning to facilitate collaboration between financial services and various industries, such as e-commerce and 'sharing economy' industries, this could realise new economies of scope and foster greater economic growth.
- Stimulating investments in AI and machine learning related areas.

B. Describe the implications of uses of artificial intelligence and machine learning by financial institutions for market concentration and systemic importance of those institutions.

The emergence of a relatively small number of advanced third-party providers in AI and machine learning could increase concentration of some functions in the financial system. Similarly, access to big data could be a source of systemic importance, especially if firms are able to leverage their proprietary sources of big data to obtain substantial economies of scope. Finally, the most innovative technologies may be mainly affordable to large companies. But if AI and machine learning can 'unbundle' traditional banking services and entice new firms to offer financial services, this might reduce the systemic importance of individual large universal banks. These banks could focus on a more narrow set of activities, rather than continuing to offer universal services

C. Describe how the uses of artificial intelligence and machine learning by financial institutions could be sources of greater stability and vulnerability in financial markets.

The divergent development of trading applications by a wide range of market players could benefit financial stability. For example, machine learning-powered robo-advisors can give more customised advice to individuals, reducing the barriers to entry for retail consumers to invest could also expand the investor base in capital markets, the use of AI and machine learning for new and uncorrelated trading strategies could also result in greater diversity in market movements, and more efficient processing of information could help to reduce price misalignments earlier.

On the other hand, new trading algorithms based on machine learning may be less predictable than current rule-based applications and may interact in unexpected ways. If a similar investment strategy based on AI and machine learning is widely used, it might increase market volatility through large sales or purchases executed almost simultaneously.

AI and machine learning could increase liquidity in financial markets through enhanced speed and efficiency of trading activities, detect excessive risks and overly-complicated transactions and to design more effective hedging strategies for risk management. On the other hand, the use of AI and machine learning may increase risks by allowing for much tighter liquidity buffers, higher leverage, and faster maturity transformation.

D. Describe how the uses of artificial intelligence and machine learning by insurance industry could affect both moral hazard and adverse selection problems.

If AI and machine learning are used to continuously adjust insurance fees in accordance with changing behaviour of the policyholders, this may reduce moral hazard. If AI and machine learning are utilised to offer customised insurance policies reflecting detailed characteristics of each person, it may also decrease adverse selection. On the other hand, these uses may pose various new challenges. For example, the more accurate pricing of risk may lead to higher premiums for riskier consumers, or entail biases that can lead to non-desirable discrimination and even reinforce human prejudices.

E. Describe challenges posed by the lack of interpretability or auditability in applications of artificial intelligence and machine learning in the financial industry.

The lack of interpretability or 'auditability' has the potential to contribute to macro-level risk if not appropriately supervised by microprudential supervisors. Many AI and machine learning developed models are being 'trained' in a period of low volatility. As such, the models may not suggest optimal actions in a significant economic downturn or in a financial crisis, or the models may not suggest appropriate management of long-term risks. Should there be widespread use of opaque models, it would likely result in unintended consequences. For example,it would be very difficult for both firms and supervisors to predict how actions directed by trading models will affect markets. Similar unintended consequences may occur in applications aimed at credit scoring, capital optimisation, or cyber threat detection, where the build-up of risks may occur slowly.

8.1.8 The terms listed in the glossary

Algorithm: a set of computational rules to be followed to solve a mathematical problem. More recently, the term has been adopted to refer to a process to be followed, often by a computer.

Artificial intelligence: the theory and development of computer systems able to perform tasks that traditionally have required human intelligence.

Augmented intelligence: augmentation of human capabilities with technology, for instance by providing a human user with additional information or analysis for decision-making.

Big data: a generic term that designates the massive volume of data that is generated by the increasing use of digital tools and information systems.

Chatbots: virtual assistance programmes that interact with users in natural language.

Cluster analysis: A statistical technique whereby data or objects are classified into groups (clusters) that are similar to one another but different from data or objects in other clusters.

Deep learning: a subset of machine learning, this refers to a method that uses algorithms inspired by the structure and function of the brain, called artificial neural networks.

FinTech: technologically enabled financial innovation that could result in new business models, applications, processes, or products with an associated material effect on financial markets and institutions and the provision of financial services.

InsurTech: the application of FinTech for insurance markets.

Internet of things: the inter-networking of physical devices, vehicles, buildings, and other items embedded with electronics, software, sensors, actuators, and network connectivity that enable these objects to collect and exchange data and send, receive, and execute commands.

Machine learning: a method of designing a sequence of actions to solve a problem that optimise automatically through experience and with limited or no human intervention.

Natural Language Processing (NLP): An interdisciplinary field of computer science, artificial intelligence, and computation linguistics that focuses on programming computers and algorithms to parse, process, and understand human language.

RegTech: any range of applications of FinTech for regulatory and compliance requirements and reporting by regulated financial institutions. This can also refer to firms that offer such applications, and in some cases can encompass SupTech (see below).

Reinforcement learning: a subset of machine learning in which an algorithm is fed an unlabelled set of data, chooses an action for each data point, and receives feedback (perhaps from a human) that helps the algorithm learn.

Robo-advisors: applications that combine digital interfaces and algorithms, and can also include machine learning, in order to provide services ranging from automated financial recommendations to contract brokering to portfolio management to their clients, without or with very limited human intervention. Such advisors may be standalone firms and platforms, or can be in-house applications of incumbent financial institutions.

Social trading: a range of trading platforms that allow users to compare trading strategies or copy the trading strategy of other investors. The latter is often referred to as 'copy trading' or 'mirror investing.'

SupTech: applications of FinTech by supervisory authorities.

Supervised learning: a subset of machine learning in which an algorithm is fed a set of 'training' data that contains labels on the observations.

Tonality analysis: a method to gauge the negativity of a piece of text by counting terms with a negative connotation.

Topic modelling: a method of unsupervised learning lets the data define key themes in text.

Unsupervised learning: a subset of machine learning in which the data provided to the algorithm does not contain labels.

Reading 8.2 *Robo-Advisors and Wealth Management.*

Phoon, K. and F. Koh. (2018). Robo-Advisors and Wealth Management. *Journal of Alternative Investments*, 20(3), 79-94. DOI: https://doi.org/10.3905/jai.2018.20.3.079

Keywords

Fintech (p. 79)
Innovations in financial technology

Robo-advisor (p. 80)
Digital platforms that provide automated, algorithm-driven financial planning services with little to no human supervision. A typical robo-advisor collects information from clients about their financial situation and future goals through an online survey, and then uses the data to offer advice and/or automatically invest client assets.

Work-flow (p. 83)
The approach robo-advisors employ to provide rudimentary financial advice, which may include reliance mainly on automated platforms, ETFs and other passive indexed investment, and mean-variance optimization to achieve a low-cost and tax-efficient portfolio. The desired outcome of such advice is to offer clients the flexibility to invest in their own portfolios, which are consistent with their objectives.

D2C platforms (p. 86)
Online platforms that provide automated algorithm-based portfolio management without human intervention.

Hybrid (p. 86)
Providing personalized services and actively managed portfolios blended with computerized portfolio recommendations.

B2B platforms (p. 86)
Digital platform providers that support traditional advisors to provide a digital wealth management solution.

8.2.1 Robo-advisors: key questions and definitions

A. What are the three key questions that arise in analyzing the robo-advisor space?

1. Are robo-advisors adequately meeting the needs of clients?
2. What are the gaps in client services?
3. Which services are needed but not served?

B. What is the definition of a robo-advisor?

The Sovereign Wealth Fund Institute (SWFI) defined robo-advisors as a type of financial advisor that provides web-based portfolio management with almost zero human intervention, typically using algorithms and formulas. On the other hand, Investopedia provided a narrower definition by specifying the services provided as follows: Robo-advisors are digital platforms that provide automated, algorithm-driven financial planning services with little to no human supervision. A typical robo-advisor collects information from clients about their financial situation and future goals through an online survey, and then uses the data to offer advice and/or automatically invest client assets.

8.2.2 Robo-advisors, their characteristics and services offered by them

A. List the three areas of passive strategies.

1. Asset allocation and implementation

2. Portfolio monitoring
3. Portfolio rebalancing

B. List the two type of robo-advisors.

1. Independent start-ups like Betterment and Wealthfront.
2. Robo-advisory platforms of established investment companies like Vanguard and BlackRock.

C. Compare and contrast North American and Asian robo-advisors.

Asian robo-advisors are still very much in the nascent stage: the AUM managed are smaller compared with those in the United States. In addition, the relative scarcity of Asian exchange-traded funds (ETFs) has resulted in slower development of the passive management approach compared to the U.S. robo-advisors.

D. Describe the generic workflow employed by robo-advisors to provide financial advice.

1. The availability of a set of investment products either produced internally as in Vanguard or sourced from a third party;
2. Analyze and/or combine the products tailored to the client based on his or her investment objectives, risk tolerance, and other factors that determine the suitability of the recommendation. The recommendation may use algorithms and/or heuristic judgement;
3. Timely communication with the client;
4. Some interactions with clients and allow for some level of discretion; and
5. The execution and transactions of decisions.

E. List the key characteristics of robo-advisors, and describe the three models (D2C, BsB and Hybrid)

1. Service Model:
    - D2C: Online platforms that provide automated algorithm-based portfolio management without human intervention.
    - B2B: Digital platform providers that support traditional advisors to provide a digital wealth management solution.
    - Hybrid: Providing personalized services and actively managed portfolios blended with computerized portfolio recommendations.
2. Minimum Investment Amount
3. Asset Management Fee
4. Access to Investment Products in Recommended Portfolio
5. Capabilities (tax planning and goal-based approach)

F. List the three main areas of private wealth management.

1. Investment advice, including tax advice,
2. Retirement and legacy advice, including estate planning and risk management,
3. Asset management, including asset gathering and allocation.

G. Describe areas of private wealth management that robo-advisors could be more effective.

1. Investment Advisory and Asset Management: Product suitability, expenses and liquidity.
2. Retirement Planning: Robo-advisors may help by using data analytics and simulations.
3. Estate Planning Services: simulation and data analytics may help the client to decide on appropriate intergeneration asset transfers
4. Tax Planning: Rules and algorithms can be customized to recommend combinations of assets and securities from different source countries that are tax beneficial to clients
5. Insurance Management: A robo-advisor may also add insurance products as a solution for clients who need wealth and/or income protection.
6. Client Education: Robo-advisors would do well to provide more client education and online training to help individual investors make more informed investment choices.
7. Mobile Platform: use of the mobile phone as a communication and personal planning tool, and device for payments as well as investment execution.
8. Data-Mining and Artificial Intelligence Software: Use of data analytics and artificial intelligence can lead to a better matching of investment opportunities with the needs of clients and lead to improved investment outcomes, including satisfying behavioral preferences
9. Enhanced Client Servicing: Chatbots can provide intuitive answers to generic customer questions, thus freeing up relationship managers to focus on complex requests and products. Chatbots would allow robo-advisors to enlarge the client base, catering to another market segment that embraces technology and social media

Reading 8.3 *Rethinking Alternative Data in Institutional Investment.*

Monk, A., M. Prins, and D. Rook. (2019). Rethinking Alternative Data in Institutional Investment. *Journal of Financial Data Science*, 1(1), 14-31. DOI: https://doi.org/10.3905/jfds.2019.1.1.014

Keywords

Alternative data (p. 14)
Datasets that are not conventionally used in investment decision making: e.g. satellite imagery of commercial or economic activity; social-media streams; microdata about consumers' shopping activities; and data exhaust created by people's online browsing.

Social media (p. 14)
Social-media streams, from which consumer, political, or other sentiment may be gauged;

Microdata (p. 14)
Microdata about consumers' shopping activities (e.g., credit card transactions or in-app purchases on smartphones)

Data exhaust (p.14)
The assortment of log files, cookies, and other digital footprints created by people's online browsing (including geolocation data from searches on mobile devices).

Rivalry (p. 16)
The extent to which one entity's use of a resource diminishes its value for another entity.

Excludability (p.16)
The degree to which one entity can prevent another from using a resource.

Defensive strategies (p. 17)
Defensive strategies prioritize capital preservation and prudent risk-taking over speedily exploiting opportunities

Defensible strategies (p. 18)
Defensible alt-data strategies can help investors increase the excludability of an alt-dataset by either restricting access to it or by developing execution capabilities around it that are not replicable by other market participants.

Operational alpha (p. 19)
To better align operating resources with investment strategies by eliminating internal inefficiencies in how investment processes are executed. Tts chief aim is to improve net returns by reducing unneeded operating costs.

Aggregation (p. 19)
The inventive collation and synthesis of documents to uncover precious metadata that is able to provide insights for enhancing communication, culture, negotiation, time allocation, benchmarking, and diligence.

Disaggregation (p. 19)
The disaggregation of collective processes into individual contributions to give a clearer picture of where latent organizational resources – and opportunities to improve them – reside.

Volume (p. 21)
The size of a dataset.

Velocity (p. 21)
The rate at which new data arrive.

Variety (p. 21)
The types of data.

Veracity (p. 21)
The degree of uncertainty around a dataset.

Granularity (p. 21)
The scale covered by specific data points or entries (e.g., continental, industry-wide).

Relationality (p. 21)
How many fields a dataset shares with other datasets of interest.

Flexibility (p. 21)
How easily new fields can be added to a dataset.

Actionability (p. 22)
Degree to which significant actions or decisions can be made based on the data

Excludable (p. 28)
Those who create or acquire some alt-datasets first can prevent all others from possessing and transacting on them, either permanently or limitedly: can only exclude others from acquiring them (or replicating them, to some approximation) for a limited time or else can only restrict the number of others who obtain them to a limited extent.

Data hoarding (p. 29)
Whereby entities leap before looking and obtain alt-datasets that promise high scarcity and excludability but only minimally consider the actionability of such alt-datasets upfront.

8.3.1 Alternative data and institutional investors

A. Provide a definition of alternative data and list examples of alternative data.

Datasets that are not conventionally used in investment decision making.

B. List the most commonly used types of alternative data.

1. satellite imagery of commercial or economic activity (e.g., the number of cars in parking lots of major retailers, ships passing through ports, and agricultural or mining operations);
2. social-media streams, from which consumer, political, or other sentiment may be gauged;
3. microdata about consumers' shopping activities (e.g., credit card transactions or in-app purchases on smartphones);
4. data scraped from the internet (e.g., job postings to track corporate hiring patterns); and
5. data exhaust – the assortment of log files, cookies, and other digital footprints created by people's online browsing (including geolocation data from searches on mobile devices).

C. Explain why the alternative data's core value proposition is different for institutional investors.

Institutional investors' patience is more aligned with defensive and defensible approaches to alt-data than it is with the exploitative strategies that short-horizon investors tend to pursue. They will likely be better off using alt-data in ways that are unharmed by competition over alt-data (i.e., nonrivalrous) or for activities others cannot easily replicate (i.e., excludable).

D. Discuss advantages and disadvantages that institutional investors may have in using alternative data.

Advantages:

- Because of their long operating horizons, institutional investors can pursue investment strategies unavailable to other market players.
- Building capacity around alt-data is strategically valuable in its own right, doing so has the added benefit of promoting innovation in all aspects of an Investor's business (e.g., creative improvements in processes, people's skill sets, and technology).

Disadvantages:

- Speed is, in general, not a comparative advan- tage for institutional investors: When an alt-dataset's value is premised on it improving a market participant's ability to speedily seize trading opportunities, this value proposition implies that the alt-datasets should be more useful for other financial organizations with comparative advantages in rapid execution.
- Institutional investors are also comparatively disadvantaged in terms of agility. Rising rivalry and declining excludability of many alt-datasets means that market participants who attempt to use alt-data to exploit opportunities must be somewhat flexible to succeed

E. Discuss why the deepest value proposition alternative data has for institutional investors entails defensive and defensible strategies.

The most powerful comparative strength that institutional investors have is patience. Their long horizons of operation mean that they can reap greater gains than other market participants by being more methodical and disciplined in their investment activities. Accordingly, the deepest value proposition alt-data has for institutional investors entails defensive and defensible strategies.

Defensive strategies prioritize capital preservation and prudent risk-taking over speedily exploiting oppor tunities, hence should be centered on pursuits such as advanced risk analysis and management or improving operating efficiencies. Such alt-data strategies can substantially decrease the degree of rivalry over an alt-dataset (i.e., one institutional investor building a defensive strategy around an alt-dataset need not reduce the value to another of doing likewise).

Defensible alt-data strategies, meanwhile, can help institutional investors increase the excludability of an alt-dataset by either restricting access to it (e.g., via making it proprietary) or by developing execution capabilities around it that are not replicable by other market participants (e.g., through having privileged access to infrastructure deals via special relationships with local governments).

F. List examples of how alternative data may be used defensively for understanding risk

- harvesting dynamic pricing information from online sources to garner a clearer, more real-time picture of inflation (and draw on wider or more targeted sources of pricing information than are usual in generic consumer-price indexes);
- aggregating label information (e.g., nutrition facts, ingredients lists) from food-product companies' offerings to see how they may be vulnerable to shifting dietary trends or new warnings by health agencies (Investors may then be able to compel company managers to alter their offerings – e.g., through shareholder activism for publicly traded companies);

- assembling online price and ratings histories of possible competitors (e.g., from Airbnb, TripAdvisor, or Yelp) or price series of airfares to that locale when doing due diligence on candidate direct investments in leisure-related properties (e.g., hotels or casinos);
- using microsensors (or other remote sensors) to track fluctuations in soil moisture for determining what plants are best suited to intercropping in a plantation-forestry investment; and
- controlling reputational risk from investee companies by monitoring controversies about them that arise in social-media posts (or other localized or unconventional news outlets).

G. Discuss the applications of alternative data to risk measurement and management for institutional investors.

- Alt-data can supply more context about how events in the wider world drive downside moves in markets. A less rivalrous (and more durable) benefit of early detection is that it allows more time for to respond to downside events once they are flagged as likely. Moreover, added context can help warn about unprecedented downside events. When more variables are tracked, there is a higher likelihood of catching anomalous behavior that heralds highly atypical events, even if the precise impacts of such events might not be immediately apparent.
- A suitable supply of alt-data could allow Investors to design index-construction methods for public (or private) assets that create tailored, controlled risk exposures.
- Alt-data have applications in other areas of risk management, such as in asset oversight and due-diligence processes, especially in private markets.

H. Describe the operational alpha gains by institutional investors through the use of alternative datasets.

The chief idea behind operational alpha is to better align operating resources with investment strategies by eliminating internal inefficiencies in how investment processes are executed. hence improve net returns by reducing unneeded operating costs. Aggregation and disaggregation are key to converting conventional internal data into alt-data. For instance, inventive collation and synthesis of documents (e.g., e-mails, investment memos, and contracts) can uncover precious metadata that is able to provide insights for enhancing communication, culture, negotiation, time allocation, benchmarking, and diligence. Likewise, the disaggregation of collective processes into individual contributions can give a clearer picture of where latent organizational resources – and opportunities to improve them – reside.

I. Describe types of alternative datasets in terms of the origins of dataset.

1. Individual Processes: Social media, news and reviews, web searches, personal data;
2. Business Processes: Transaction data, corporate data, government agency data;
3. Sensors: Satellites, geolocation, other sensors.

J. Discuss why the volume, veracity and velocity of big data may not determine value of alternative data for institutional investors.

1. Velocity may be not so important for assets without value-determining properties which change frequently;
2. A dataset may contain many items (i.e., have high volume) from only a narrow number of categories of interest. In such a case, a dataset that has smaller volume, but encompasses more categories (i.e., is more comprehensive), would likely have higher value
3. Reliability (which covers the accuracy, precision, and verifiability of a dataset) seems to us a more fitting concept than veracity.

K. Describe the six-dimensional characterization of alternative data.

1. Reliability: How accurate, precise, and verifiable the data are (e.g., error-free, unbiased, checkable);
2. Granularity: The scale covered by specific data points or entries (e.g., continental, industry-wide);
3. Freshness: Age of the data (i.e., when collected/generated) relative to the phenomena they reflect;
4. Comprehensiveness: What portion of a given domain the data cover (e.g., 25% of households in Canada);
5. Actionability: Degree to which significant actions or decisions can be made based on the data;
6. Scarcity: How widely or readily available the data are to other (especially competing) organizations.

L. Discuss external asset managers and alternative data providers as methods of accessing alternative data.

Some external asset managers (e.g., some hedge funds) have enjoyed relatively lengthy experience in working with alt-data, though trusting external asset managers to provide indirect access comes at the cost of forfeiting some experience with learning to innovate. Furthermore,there are at least three additional problems. First external managers are less concerned about capital preservation and are more motivated to fixate upon investment alpha, which predisposes them to becoming engulfed in an escalating arms race around alt-data. Second institutional investors lose the ability to inspect, verify, and otherwise work with the data on which those managers are basing decisions. Third, the institutional investor is effectively subsidizing the external manager in improving its capacity for alt-data which increases both the manager's comparative advantage and the institutional investor's reliance on external parties for alt-data capacity.

Partnering with alternative data providers mitigates many of the problems with relying on external managers. First, vendor-supplied alt-data are not necessarily exposed to problems connected with opportunistic usage of alt-data. Second, concerns about transparency are partly lessened since institutional nvestors are actually able to examine the alt-datasets. Third, for defensive applications of alt-data, subsidization of the vendor's provision of additional alt-datasets would actually tend to be helpful. Nonetheless, the foremost downside is the low degree of excludability for vendor-supplied alt-data.

M. Discuss the consequences of the increased use of alternative data on risk for institutional investors.

In not using alt-data, market participants handicap themselves by limiting any informational edge. As more market participants embrace alt-datasets, markets (especially public markets) will be more affected by them, until they affect even passive investing.

Pressures toward short-termism bias decisions toward action rather than inaction: More market activity means greater volatility. Furthermore, intensified competition over alt-data means that there is pressure not only to act fast but also to act big because of fleeting actionability. More extensive activity also increases volatility. Finally, the increasing use of algorithmic methods for trading based on alt-data will likely contribute to higher market volatility. Increased volatility will probably raise costs of passive investing through a combination of higher transaction costs (because of faster turnover), hedging costs, liquidity threats, and cash drag.

N. Compare accessing alternative data through external asset managers versus the alternative data vendors.

see above.

O. Describe rivalry and excludability as determinants of alternative dataset's value.

Practically all data in finance are rivalrous in the sense that any use of data for transacting reduces (or even eliminates) the value in executing similar transactions thereafter, regardless of who conducts them. This property means any (profitable) actionability of data is eventually self-eliminating so that the value of a dataset decreases by acting on it.

Scarcity is a crucial reason why alt-datasets can be so precious. Excludability of many alt-datasets means substantial value can be realized by being first to capture a dataset

Reading 8.4 *A Machine Learning Approach to Risk Factors: A Case Study Using the Fama-French-Carhart Model.*

Simonian, J., C. Wu, D. Itano and V. Narayanan. (2019). A Machine Learning Approach to Risk Factors: A Case Study Using the Fama-French-Carhart Model. *Journal of Financial Data Science*, 1(1), 32-44. DOI: https://doi.org/10.3905/jfds.2019.1.032

Keywords

Factors (p. 32)
A restricted set of explanatory variables that adequately explain asset and portfolio behavior to a sufficient degree over time.

Linear (p. 32)
Variables combined only with a linear formula.

Nonlinear (p. 32)
Variables combined with interaction terms and complex functional forms.

Random forest (p. 33)
A machine learning algorithm able to account for the nonlinear relationships, discontinuities (e.g., threshold correlations), and interactions among the variables, while dispensing with the need for complex functional forms or additional interaction terms.

Supervised (p. 34)
Supervised learning (including reinforcement learning) algorithms use input variables that are clearly demarcated. The goal is to produce rules and/or inferences that can be reliably applied to new data, whether for classification or regression-type problems.

Unsupervised (p. 34)
Unsupervised learning algorithms include those encompassing clustering and dimension reduction, in which the goal is to draw inferences and define hidden structures from input data. Unsupervised algorithms are distinguished by the fact that the input data are not categorized or classified. Rather, the algorithm is expected to provide a structure for the data.

Root node (p. 34)
Top-most node of decision trees.

Decision node (p. 34)
One of series of decisions of a decision tree.

Terminal node (p. 34)
Final leaf node whose value is a predicted value for a target variable.

CART (p. 34)
CART uses an algorithm called binary recursive partitioning to construct decision trees.

Binary recursive partitioning (p. 34)
Features are evaluated using all sample values, and the feature that minimizes the cost function at a specific value is chosen as the best split. Recursive partitioning takes place a each level down the tree, and the value at each leaf of the tree is the average of all the resulting observations.

Bagging (p. 34)
An ensemble of decision trees is constructed via bootstrapping, which involves resampling from the data with replacement to build a unique dataset for each tree in the ensemble. The trees in the ensemble are then averaged (in the case of regression), resulting in a final model. The bootstrap aggregation of a large number of trees is called bagging.

Out-of-bag data (p. 34)
When a bootstrap is conducted, some observations are left out to be used for measuring estimation error and variable importance.

Feature importance (p. 35)
An output of the analsis that indicates the importance of each explanatory variable in contributing to the predicted value of the dependent variable in question.

Mean decrease accuracy (p. 34)
Measures the degree to which the predictive power of the model would be diluted if the values for the explanatory variable in question were randomly changed.

Fama-French-Carhart (p. 37)
A multifactor extension of the capital asset pricing model (CAPM) by introducing three new factors in addition to the market factor: the size factor (small-cap stock returns minus large-cap stock returns), value (high book-to-price stock returns minus low book-to-price stock returns), and momentum (high-returning stocks minus low-returning stocks).

Probabilistic Sharpe ratio (p. 42)
The PSharpe measure is designed to show the probability of a strategy achieving a given Sharpe ratio threshold given a specific track record or backtest length and the presence of non-normal returns.

8.4.1 Applications of random forest regression algorithm to factor models

A. Discuss two shortcomings of nonlinear factor models that are developed to address shortcomings of linear models.

1. The structure of nonlinear latter models is often heavily dependent on the sample data. As the sample expands or contracts, we at times find that the function specified by the model changes, sometimes dramatically.
2. Unlike linear models, parameter estimates cannot always be derived analytically but are often found using iterative methods, which may ultimately fail to converge if the initial values are too distant from possible solution values. Initial values that are remote from optimal values can also cause convergence to a local solution rather than a global solution.

B. Discuss the ability of random forest algorithm to overcome one shortcoming of linear models.

The random forest algorithm is able to account for the nonlinear relationships, discontinuities (e.g., threshold correlations), and interactions among the variables.

C. Discuss the ability of random forest algorithm to overcome one shortcoming of nonlinear models.

It dispenses with the need for complex functional forms or additional interaction terms (thus remaining in harmony with the principle of parsimony)

D. List 4 components of the decision tree when applied to the regression problem of factor models.

For regression-type problems, decision trees

1. start from a topmost or *root* node,
2. and proceed to generate *branches*, with each branch containing a condition, and a prediction in the form of a real-valued number, given the condition in question;
3. trees are composed of a series of conditions attached to *decision nodes*,
4. which ultimately arrive at a *leaf* or *terminal node* whose value is a real number: the latter value represents a predicted value for a target variable given a set of predictor values.

E. Describe how bagging is used in an ensemble of decision trees (random forest).

Bagging is useful because it generally reduces overfitting and has a lower variance when compared to processes that only use individual decision trees. An individual tree may end up learning highly idiosyncratic relationships among the data and hence may end up overfitting the model. Averaging ensembles of trees provides a better opportunity to uncover more general patterns and relationships between variables. Overfitting can also be addressed by using simpler trees (i.e., those with a lower number of splits).

F. Calculate the predicted value of an independent (response) variable given a set of predictor values and the outputs of a binary regression decision tree algorithm.

Predicted values of the response variable for regression at a given point x are given by $\hat{f}(x) = \frac{1}{J}\sum_{j=1}^{J}\hat{h}_j(x)$, where $\hat{h}_j(x)$ is the prediction of the response variable at x using the jth tree

G. Describe the role of out-of-bag observations in a random forest algorithm.

Some observations are left out of the bootstrap, are called out-of-bag (OOB) data, and are used for measuring estimation error and variable importance. Using all the trees may produce a false level of conf idence in the predictions of the response variable for observations in the training set. To remedy this risk, the prediction of the response variable for training set observations is done exclusively with trees for which the observation is OOB.

H. Discuss mean decrease accuracy approach to estimating feature importance in a random forest algorithm.

Once the jth tree is generated, the values for the predictor variables are randomly permuted in the bootstrapped sample, and the prediction accuracy is recalculated. For regression, the FI for the observation is calculated as the difference between the MSE of the predictions using the permuted data and the MSE of the predictions using the original data. Next, a normalization is generally conducted to allow an assignment of a relative FI (RFI) value to each feature, by adding the FI values for each factor in a single tree and dividing that value into the FI value for each factor. This will yield a cross section of FI values that sum to unity. This operation is repeated for each tree, and the normalized FI (NFI) values are then averaged across all the generated trees to produce an RFI value for a given feature.

I. Recognize and apply the probabilistic Sharpe ratio.

The PSharpe measure is designed to show the probability of a strategy achieving a given Sharpe ratio threshold given a specific track record or backtest length and the presence of non-normal returns: $\hat{PSR}(SR^*) = Z[\frac{\hat{SR}-SR^*)\sqrt{n-1}}{\sqrt{1-\hat{\gamma}_3 SR^* + \frac{\hat{\gamma}_4-1}{4}\hat{SR}^2}}]$ where $Z[\cdot]$ is the cumulative distribution function of a standard normal distribution, and $\hat{SR}$ is the observed Sharpe ratio. SR^* is the predefined benchmark Sharpe ratio (ex ante Sharpe ratio), n is the number of periods over which the strategy's performance is tested, and $\hat{\gamma}_3, \hat{\gamma}_4$ are the respective observed skewness and kurtosis values of the strategy.

Reading 8.5 *Machine Learning for Stock Selection.*

Rasekhschaffe, K. and R. Jones. Machine Learning for Stock Selection. Financial Analyst Journal, Forthcoming.

8.5.1 Keywords

Machine learning (p. 3)
Machine learning is an umbrella term for methods and algorithms that allow machines to uncover patterns without explicit programming instructions

Overfitting (p. 5)
Overfitting occurs when a model picks up noise instead of signals. Overfit models have very good in-sample performance, but little predictability on unseen data.

Signal-to-noise ratio (p. 4)
With lower signal-to-noise ratios, out of sample results diverge much faster from in-sample results.

Ensemble algorithms (p. 6)
These algorithms generate many forecasts from weak learners and then combine these forecasts to form a stronger learner

Feature engineering (p. 8)
Feature engineering uses domain knowledge to structure a problem so that it is more amenable to machine learning solutions.

Forecast horizon (p. 10)
For most stock-selection applications, an appropriate forecast horizon runs from daily to quarterly. Short forecast horizons are more suited to low-capacity, high-turnover strategies, while long horizons are more suited to high-capacity, lower-turnover strategies. The forecast horizon should also reflect the frequency of the underlying predictive data

Bagging (p. 11)
Bootstrap aggregation (bagging) independently fits estimators such as decision trees (weak learners) on random subsets of the training set. Each of the weak learners is overfit, but errors due to overfitting tend to be reduced when combining forecasts of weak learners into a strong learner.

Boosting (p. 11)
Boosting sequentially fits estimators on the training set and gives more weight to misclassified observations in successive boosting rounds. The strong learner is an accuracy-weighted average of the weak learners. By giving more weight to more successful learners, boosting can address bias.

Fundamental factors (p. 13)
Factors related to future economic success; are usually more important over longer horizons.

Technical factors (p. 11)
Factors related to future supply and demand; tend to be more predictive for shorter horizons.

8.5.2 The applications of machine learning algorithms to stock selection

A. Describe the role of signal to noise ratio in creating overfitted models.

Because of the low signal-to-noise ratios in forecasting stock returns, it is particularly important to avoid overfitting. With lower signal-to-noise ratios, out of sample results diverge much faster from in-sample results. Overfitting can make the results look much better than they are likely to be in any real-world application.

B. Discuss the implications of the paper's findings with regard to in-sample versus out--of-sample errors as the number of boosting iterations increase.

As we increase the number of boosting iterations, the errors always decline in-sample and become negligible after around 400 boosting iterations. In sharp contrast, the error rate in the hold-out sample first decreases, but then increases after approximately 50 boosting iterations. This is where the algorithm begins to overfit the data.

C. Describe the four different approaches to bagging and boosting employed by the paper to avoid overfitting.

1. Combining forecasts from different classes of algorithms
2. Combining forecasts based on different training windows
3. Combining forecasts that use different factor libraries
4. Combining forecasts for different horizons

D. Explain the importance of feature engineering in mitigating the overfitting problem.

Feature engineering uses domain knowledge to structure a problem so that it is more amenable to machine learning solutions. Feature engineering is where domain knowledge flows into the process. In the context of stock selection, this can cover decisions such as: what are we trying to forecast; which algorithms are likely to be most effective; which training windows are likely to be most informative; how should we standardize factors and returns; and which factors are likely to provide valuable information. It is one of the most effective ways to overcome overfitting because it allows us to increase the signal-to-noise ratio before training the algorithms.

E. Describe the three decisions that must be made with regard to the forecasting goals of the machine learning algorithms.

1. What are we forecasting
2. how to define these categories (of discrete variables that MLA's usually predict) – e.g. outperformer versus underperformer
3. Selecting a forecast horizon

F. Describe the bias versus variance tradeoff.

Bias is caused when the estimation method does not effectively capture fundamental relationships in the data (underfitting). Variance is an error arising from small changes in the training set, which means the estimator does not learn relationships that generalize out-of-sample (overfitting).

G. Explain the role of bagging and boosting in affecting the bias versus variance tradeoff.

Bootstrap aggregation (bagging) independently fits estimators such as decision trees (weak learners) on random subsets of the training set. Each of the weak learners is overfit, but errors due to overfitting tend to be reduced when combining forecasts of weak learners into a strong learner. Boosting sequentially fits

estimators on the training set and gives more weight to misclassified observations in successive boosting rounds. The strong learner is an accuracy-weighted average of the weak learners. By giving more weight to more successful learners, boosting can address bias. However, if we allow the boosting algorithm to overweight successful weak learners too aggressively, this benefit will be more than offset by increasing variance.

H. Understand the terms appearing in the glossary

Activation functions in neural networks determine the output of nodes given inputs. Non-linear activation functions allow neural networks to learn non-linear patterns. Popular activation functions include Rectified Linear Units or ReLU, tanh and sigmoid. Sigmoid activations are often used in the output layer of binary classification problems because they can map inputs to probabilities between 0 and 1.

Artificial neural networks and deep learning are algorithms loosely modeled on the human brain. Neural units are organized in layers and connected with each other making it possible to learn many interactive relationships. However, artificial neural networks can be easy to over-fit, and finding the correct architecture for a given problem is often difficult. Some recent innovations related to Deep Learning, such as dropout, make it possible to learn deeper architectures without overfitting.

Bagging (Bootstrap Aggregating) algorithms combine forecasts generated from base algorithms on randomly sampled learning sets. Random forests are an example of the application of bagging to CART models. In contrast to boosting, forecasts of the base algorithms are equally weighted. Bagging tends to increase stability of algorithms and helps prevent overfitting

Boosting algorithms combine forecasts from many base algorithms such as CART. In contrast to random forests, boosting gives more weight to more successful models. Boosting algorithms have the potential to learn more efficiently from data than random forests but require greater care when tuning parameters since they are easier to overfit. They also take longer to run than bagging because they require sequential processing.

Dropout in neural networks is a technique to limit overfitting. Similar to bagging, dropout effectively is a model-averaging technique. When training a neural network the algorithm drops out elements of layers leading to models that often generalize better on unseen data.

Gradient boosted trees use decision trees as base learners. Subsequent trees are trained on residuals from earlier iterations. The learning rate and number of boosting iterations are key parameters that influence how aggressively the model can learn and also overfit. The depth of the base learners also is an important parameter.

Random forest combine many CART models (or trees) by averaging their forecasts. This can often diversify away errors due to overfitting, and therefore random forests usually have more signal and less noise than the individual trees. Random forests are quite robust to overfitting, and often work well out of sample.

Reading 8.6 *Empirical Asset Pricing via Machine Learning.*

Gu, S., B. Kelly, and D. Xiu. (2018). Empirical Asset Pricing via Machine Learning.

8.6.1 Keywords

Machine Learning (p. 2)
Describes (i) a diverse collection of high-dimensional models for statistical prediction, combined with (ii) so-called "regularization" methods for model selection and mitigation of overfit, and (iii) efficient algorithms for searching among a vast number of potential model specifications.

Regularization (p. 2)
Refinements in implementation that emphasize stable out-of-sample performance to explicitly guard against overfit.

Mean squared error (p. 9)
An objective function for estimating model parameters

Ordinary least squares (p. 9)
Estimation of the simple linear model useing a standard least squares, or l_2 objective function

Heavy tails (p. 11)
A well known attribute of financial returns and stock-level predictor variables. Convexity of the least squares objective places extreme emphasis on large errors, thus outliers can undermine the stability of OLS-based predictions.

Huber loss function (p. 12)
A hybrid of squared loss for relatively small errors and absolute loss for relatively large errors, where the combination is controlled by a tuning parameter that can be optimized adaptively from the data.

Penalized linear models (p. 11)
The simple linear model begins to overfit noise rather than extracting signal in the presence of many predictors. The most common machine learning device for imposing parameter parsimony is to append a penalty to the objective function in order to favor more parsimonious specifications

Loss function (p. 11)
Objective function measuring model's in-sample performance.

Penalty function (p. 11)
A term incorporated in the objective function in order to favor more parsimonious specifications.

Elastic net(p. 11)
A popular penalty function that involves two non-negative hyperparameters

Hyperparameters (p. 12)
Determines penalty to the objective function in order to favor more parsimonious specifications.

Tuning parameters (p. 12)
aka Hyperparameters

LASSO (p. 13)
Uses an absolute value, or l_1, parameter penalization, which sets coefficients on a subset of covariates to exactly zero. In this sense, the LASSO imposes sparsity on the specification and can thus be thought of as a variable selection method.

Principal components (p. 13)
A classic dimension reduction technique that regularizes the prediction problem by zeroing out coefficients on higher order components. In the first step, principal components analysis combines regressors into a small set of linear combinations that best preserve the covariance structure among the predictors. In the second step, a few leading components are used in standard predictive regression.

Partial least squares (p. 13)
Performs dimension reduction by directly exploiting covariation of predictors with the forecast target. All predictors are averaged into a single aggregate component with weights proportional to univariate return prediction coefficient via OLS (i.e. the "partial" sensitivity of returns to each predictor j., placing the highest weight on the strongest univariate predictors, and the least weight on the weakest.

Approximation error (p. 15)
When the "true" model is complex and nonlinear, restricting the functional form to be linear introduces approximation error due to model misspecification.

Estimation error (p. 15)
Component of a model's forecast error which arises due to sampling variation.

Intrinsic error (p. 15)
Intrinsic error is irreducible; it is the genuinely unpredictable component of returns associated with news arrival and other sources of randomness in financial markets.

Terminal nodes (p. 16)
The "leaves" (terminal nodes) of a regression tree which gives a prediction based on the values of the characteristics.

Impurity (p. 16)
The loss associated with the forecast error for a branch in a regression tree.

Weak learners (p. 17)
Over-simplified trees which on their own are "weak learners" with minuscule predictive power.

Boosting (p. 18)
Many weak learners, as an ensemble, comprise a single "strong learner" with greater stability than a single complex tree.

Random forest (p. 18)
An ensemble method that is a variation on a more general procedure known as bootstrap aggregation, or "bagging". It draws different bootstrap samples of the data, fits a separate regression tree to each, then averages their forecasts.

Gradient boosted regression trees (p. 18)
A boosting procedure which starts by fitting a shallow tree that a weak predictor with large bias. Next, a second simple tree with the same shallow depth L is used to fit the prediction residuals from the first tree. Forecasts from these two trees are added together to form an ensemble prediction of the outcome, but the forecast component from the second tree is shrunken by a factor to help prevent the model from overfitting the residuals. At each new step, a shallow tree is fitted to the residuals from the model, and its residual forecast is added to the total with a shrinkage weight. This is iterated until the final output is an additive model of shallow trees

Neural network (p. 19)
Anonlinear method that is arguably the most powerful modeling device in machine learning. Has theoretical underpinnings as "universal approximators" for any smooth predictive association.

Feed-forward networks (p. 19)
A traditional type of neural network. Consists of an "input layer" of raw predictors, one or more "hidden layers" that interact and nonlinearly transform the predictors, and an "output layer" that aggregates hidden layers into an ultimate outcome prediction.

Input layer (p. 20)
First layer of raw predictors in a feed-forward network

Hidden layers (p. 19)
One or more layers in a feed-forward network that interact and nonlinearly transform the predictors in the input layer.

Output layer (p. 20)
Aggregates hidden layers into an ultimate outcome prediction in a feed-forward network

ReLU function (p. 21)
Rectified linear unit: a popular nonlinear activation function for nodes in a neural network

Stochastic gradient descent (p. 21)
Unlike standard descent that uses the entire training sample to evaluate the gradient at each iteration of the optimization, SGD evaluates the gradient from a small random subset of the data. This approximation sacrifices accuracy for enormous acceleration of the optimization routine.

Early stopping (p. 22)
A general machine learning regularization tool. In each step of the optimization algorithm, the parameter guesses are gradually updated to reduce prediction errors in the training sample. At each new guess, predictions are also constructed for the validation sample, and the optimization is terminated when the validation sample errors begin to increase. This typically occurs before the prediction errors are minimized in the training sample, hence the name early stopping.

Sharpe ratio (p. 36)
Evidence for economic gains from machine learning forecasts based on utility gains for a mean-variance investor

8.6.2 Applications of machine learning algorithms to empirical asset pricing

A. Describe the three components of the definition of machine learning.

1. a diverse collection of high-dimensional models for statistical prediction, combined with
2. so-called "regularization" methods for model selection and mitigation of overfit, and
3. efficient algorithms for searching among a vast number of potential model specifications.

B. Describe the three aspects of empirical asset pricing model that makes it attractive for the applications of machine learning algorithms.

1. Modern empirical asset pricing research (i) seeks to describe and understand differences in expected returns across assets and (ii) focuses on dynamics of the aggregate market equity risk premium. Machine learning, whose methods are largely specialized for prediction tasks, is thus ideally suited to the problem of risk premium measurement.
2. The collection of candidate conditioning variables for the risk premium is large. With an emphasis on variable selection and dimension reduction techniques, machine learning is well suited for such challenging prediction problems by reducing degrees of freedom and condensing redundant variation among predictors.
3. Further complicating the problem is ambiguity regarding functional forms through which the high-dimensional predictor set enter into risk premia. Machine learning is well suited for problems of ambiguous functional form.

C. Compare and contrast the overall performance of linear versus nonlinear models in predicting individual stock returns and portfolio returns.

Allowing for nonlinearities substantially improves predictions. Trees and neural nets unambiguously improve return prediction. But the generalized linear model, which introduces nonlinearity via spline functions of each individual baseline predictor (but with no predictor interactions), fails to robustly outperform the linear specification. This suggests that allowing for (potentially complex) interactions among the baseline predictors is a crucial aspect of nonlinearities in the expected return function. Predictive power at the portfolio level is more pronounced than at the stock level.

D. Explain one potential short coming of machine learning algorithms when used to predict asset returns.

Improved predictions are only measurements. The measurements do not tell us about economic mechanisms or equilibria. Machine learning methods on their own do not identify deep fundamental associations among asset prices and conditioning variables.

E. Describe the roles of "training" set, "validation" set and "testing" set in using machine learning algorithms in to predict stock returns.

The sample into three disjoint time periods that maintain the temporal ordering of the data. The first, or "training," subsample is used to estimate the model subject to a specific set of tuning parameter values. The second, or "validation," sample is used for tuning the hyperparameters. Tuning parameters are chosen from the validation sample taking into account estimated parameters, but the parameters are estimated from the training data alone. The idea of validation is to simulate an out-of-sample test of the model. the third, or "testing," subsample, which is used for neither estimation nor tuning, is truly out-of-sample and thus is used to evaluate a method's predictive performance.

F. Recognize the Huber loss function.

If $|x| \leq \xi$: $H(x;\xi) = x^2$; otherwise $H(x;\xi) = 2\xi|x| - \xi^2$. The Huber loss, $H(\cdot)$, is a hybrid of squared loss for relatively small errors and absolute loss for relatively large errors, where the combination is controlled by a tuning parameter, ξ, that can be optimized adaptively from the data

G. Describe the benefit of using the Huber loss function as opposed to standard least squares method to the estimation of linear models.

Heavy tails are a well known attribute of financial returns and stock-level predictor variables. Convexity of the least squares objective places extreme emphasis on large errors, thus outliers can undermine the stability of OLS-based predictions. In the machine learning literature, a common choice for counteracting the deleterious effect of heavy-tailed observations is the Huber robust objective function,

H. Recognize the "elastic net" approach for modeling penalized linear models.

A popular penalty function to incorporate in the loss function of the penalized linear model, which takes the form: $\phi(\theta; \lambda, \rho) = \lambda(1-\rho)\sum_{j=1}^{P}|\theta| + \frac{1}{2}\lambda_\rho \sum_{j=1}^{P}\theta_j^2$. The elastic net involves two non-negative hyper-parameters, λ, ρ.

I. Compare and contrast "elastic net" penalty versus LASSO and Ridge Regression.

The elastic net includes two well known regularizers as special cases. The $\rho = 0$ case corresponds to the LASSO and uses an absolute value, or l_1, parameter penalization, which sets coefficients on a subset of covariates to exactly zero. In this sense, the LASSO imposes sparsity on the specification and can thus be thought of as a variable selection method. The $\rho = 1$ case corresponds to ridge regression, which uses an l_2 parameter penalization, that draws all coefficient estimates closer to zero but does not impose exact zeros anywhere. In this sense, ridge is a shrinkage method that helps prevent coefficients from becoming unduly large in magnitude. For intermediate values of ρ, the elastic net encourages simple models through both shrinkage and selection.

J. Compare and contrast the principle components regression versus partial regression.

A drawback of PCR is that it fails to incorporate the ultimate statistical objective – forecasting returns – in the dimension reduction step. PCA condenses data into components based on the covariation among the predictors. This happens prior to the forecasting step and without consideration of how predictors associate with future returns.

In contrast, partial least squares performs dimension reduction by directly exploiting covariation of predictors with the forecast target. All predictors are averaged into a single aggregate component with weights proportional to its univariate return prediction coefficient via OLS ('partial" sensitivity of returns to each predictor), placing the highest weight on the strongest univariate predictors, and the least weight on the weakest.

K. Recognize and describe the three sources of forecast error (decomposition of forecast errors).

We can decompose a model's forecast error = approximation error + estimation error + instrinsic error, as: $r_{i,t+1} - \hat{r}_{i,t+1} = [g * (z_{i,t}) - g(z_{i,t}; \theta)] + [g(z_{i,t}; \theta) - g(z_{i,t}; \hat{\theta})] + \epsilon_{i,i+t}$

Intrinsic error is irreducible; it is the genuinely unpredictable component of returns associated with news arrival and other sources of randomness in financial markets. Estimation error, which arises due to sampling variation, is determined by the data. It is potentially reducible by adding new observations, though this may not be under the econometrician's control. Approximation error is directly controlled by the econometrician, and is potentially reducible by incorporating more flexible specifications that improve the

model's ability to approximate the true model. But additional flexibility raises the risk of overfitting and destabilizing the model out-of-sample.

L. Describe the boosting regularization method in the context of regression trees.

Recursively combines forecasts from many over-simplified trees. It starts by fitting a shallow tree; then, a second simple tree is used to fit the prediction residuals. Forecasts from these two trees are added together to form an ensemble prediction of the outcome, but the forecast component from the second tree is shrunken by a factor to help prevent the model from overfitting the residuals. At each new step, a shallow tree is fitted to the residuals, and its residual forecast is added to the total with a shrinkage weight. This is iterated until the final output, which is therefore an additive model of shallow trees. The tuning parameters are adaptively choose in the validation step.

M. Describe the random forest regularization method in the context of regression trees.

Random forest is an ensemble method that combines forecasts from many different trees, based on a more general procedure known as bootstrap aggregation, or "bagging". The procedure draws different bootstrap samples of the data, fits a separate regression tree to each, then averages their forecasts. Trees for individual bootstrap samples tend to be deep and overfit, making their individual predictions inefficiently variable. Averaging over multiple predictions reduces this variation, thus stabilizing the trees' predictive performance.

N. Describe the dropout method in the context of random forest regression trees.

A variation on bagging designed to reduce the correlation among trees in different bootstrap samples. It de-correlates trees by considering only a randomly drawn subset of predictors for splitting at each potential branch, so that early branches for at least a few trees will split on characteristics other than the most dominant predictor. This lowers the average correlation among predictions to further improve the variance reduction relative to standard bagging.

O. Recognize rectified linear unit (ReLU) activation function in the context of neural networks.

A nonlinear activation function for nodes. If $x < 0$: $\text{ReLU}(x) = 0$; otherwise: $\text{ReLU}(x) = x$.

Reading 8.7 *The 10 Reasons Most Machine Learning Funds Fail.*

Lopez de Prado, M. (2018). The 10 Reasons Most Machine Learning Funds Fail. The Journal of Portfolio Management, 44 (6) 120-133.

Keywords

Backtesting (p. 122)
Historical simulation of how the strategy would have performed in the past.

Volume clock (p. 123)
Forming bars, each row of a table, as a subordinated process of trading activity.

Dollar bars (p. 123)
Formed by sampling an observation every time a predefined market value is exchanged.

Stationary (p. 123)
The null hypothesis of a unit root is rejected for stationary series .

Integer differentiation (p. 123)
Integer 1 differentiation is like the one used for computing returns on log-prices

Fractional differentiation (p. 124)
Fractional differentiation allows us to generalize the notion of returns to noninteger (positive real) differences d. When $d = 0$, that is the case where the differentiated series coincide with the original one. When $d = 1$, that is the standard first-order integer differentiation, which is used to derive log-price returns

Triple barrier method (p. 127)
A method to label observations according to the condition that triggers an exit of a position. The two horizontal barriers are defined by profit-taking and stop-loss limits, which are a dynamic function of estimated volatility (whether realized or implied). The third barrier is defined in terms of the number of bars elapsed since the position was taken.

Precision (p. 128)
The number of true positives relative to the total number of predicted positives

Recall (p. 128)
The number of true positives relative to the total number of positives

F1-score (p. 128)
The harmonic average between precision and recall

Walk-forward approach (p. 129)
The most common backtest method in the literature is the walk-forward (WF) approach, where each strategy decision is based on observations that predate that decision.

Leakage (p. 129)
Leakage takes place when the training set contains information that also appears in the testing set

Probabilistic Sharpe ratio (p. 132)
The probabilistic Sharpe ratio (PSR) provides an adjusted estimate of the Sharpe ratio by removing the inflationary effect caused by short series with skewed and/or fat-tailed returns.

Deflated Sharpe ratio (p. 132)
The deflated Sharpe ratio (DSR) computes the probability that the true Sharpe ratio exceeds a rejection threshold, where that rejection threshold is adjusted to reflect the multiplicity of trials.

8.7.1 The most common errors made when machine learning techniques are applied to financial data sets

A. Compare and contrast the silo approach in discretionary strategies versus meta-strategy in machine learning strategies.

Investment firms ask discretionary managers to work independently from one another, in silos, to ensure diversification. On the other hand, successful quantitative firms apply the meta-strategy paradigm: Tasks of the assembly line are clearly divided into subtasks. Quality is independently measured and monitored for each subtask. The role of each quant is to specialize in a particular task, to become the best there is at it, while having a holistic view of the entire process.

B. Compare and contrast repeated backtesting using machine learning versus examining feature importance of the results from a machine learning application.

Backtesting where we are repeating a test over and over on the same data will likely lead to a false discovery. Instead, the next natural step after fitting a good machine learning classifer should be to understand what features contributed to that performance. Maybe we could add some features that strengthen the signal responsible for the classifier's predictive power. Maybe we could eliminate some of the features that are only adding noise to the system. Most critically, understanding feature importance opens up the proverbial black box.

C. Describe the two problems with data samples generated using time bars.

1. Markets do not receive information at a constant time interval. This means that time bars oversample information during low-activity periods and undersample information during high-activity periods.
2. Time-sampled series often exhibit poor statistical properties, such as serial correlation, heteroskedasticity, and non-normality of returns

D. Describe the advantages of dollar bars over time bars in creating data for machine learning algorithms.

1. By sampling bars in terms of dollar value exchanged rather than ticks or volume, particularly when the analysis involves significant price fluctuations, the number of bars per day will not vary as wildly over the years.
2. Dollar bars tend be robust to changes in number of outstanding shares over the course of a security's life as a result of corporate actions; Even after adjusting for splits and reverse splits, there are other actions that will affect the amount of ticks and volumes, such as issuing new shares or buying back existing shares.

E. Describe the benefit of using fractional differentiation in generating stationary series and preserving memory.

Achieves stationarity without wiping out memory in and correlation with the original series.

F. Explain the triple-barrier method for labeling observed returns.

Label observations according to the condition that triggers an exit of a position. The two horizontal barriers are defined by profit-taking and stop-loss limits, which are a dynamic function of estimated volatility. The third barrier is defined in terms of the number of bars elapsed since the position was taken (an activity-based, If the upper horizontal barrier is touched first, we label the observation as a 1. If the lower horizontal

barrier is touched first, we label the observation as a -1. If the vertical barrier is touched first, we have two choices: the sign of the return, or a 0.

G. Describe the definitions of precision, recall and F1-score as features of machine learning algorithms.

Precision is the number of true positives relative to the total number of predicted positives. Recall is the number of true positives relative to the total number of positives. F1-score is the harmonic average between precision and recall.

H. Explain the role of non-independent identically distributed in returns in the failure of k-fold cross-validation in finance.

Placing serially correlated features in different sets leaks information from the training set to the testing set.

I. Describe walk forward (WF) approach to backtesting of trading strategies.

The most common backtest method in the literature is the walk-forward (WF) approach, where each strategy decision is based on observations that predate that decision.

J. Describe advantages and disadvantages of walk forward approach.

WF enjoys two key advantages:

1. WF has a clear historical interpretation; its performance can be reconciled with paper trading
2. History is a filtration; hence, using trailing data guarantees that the testing set is out-of-sample (no leakage), as long as purging has been properly implemented.

WF suffers from three major disadvantages:

1. First, a single scenario is tested (the historical path), which can be easily overfit
2. WF is not necessarily representative of future performance because results can be biased by the particular sequence of data points.
3. The initial decisions are made on a smaller portion of the total sample. Even if a warm-up period is set, most of the information is used by only a small portion of the decisions.

K. Explain the relationship between the maximum Sharpe ratio obtained from several backtested strategies and the return volatility of those strategies.

WF backtests exhibit high variance because a large portion of the decisions are based on a small portion of the dataset: a few observations will have a large weight on the Sharpe ratio. WF's high variance leads to false discoveries because researchers will select the backtest with the maximum estimated Sharpe ratio, even if the true Sharpe ratio is zero.

L. Describe the concept of probabilistic Sharpe ratio.

The probabilistic Sharpe ratio (PSR) provides an adjusted estimate of the Sharpe ratio by removing the inflationary effect caused by short series with skewed and/or fat-tailed returns.

M. List the impacts of nonnormalized Sharpe ratio, track record, skewness and kurtosis on probabilistic Sharpe ratio.

PSR increases with greater nonnormalized SR, longer track records, or positively skewed returns, but it decreases with fatter tails.

Reading 8.8 *Evaluating Trading Strategies.*

Harvey, C. R. and Y. Liu (2014). Evaluating Trading Strategies. [Special 40th Anniversary Issue]. The Journal of Portfolio Management, 40(5), 108-118.

Keywords

T-statistics (p. 110)
Number of standard deviations from the null hypothesis of zero.

Family-wise error rate (p. 111)
In the family-wise error rate, it is unacceptable to make a single false discovery. This is a very severe rule.

False discovery rate (p. 111)
The false discovery rate views unacceptable in terms of a proportion. For example, if one false discovery were unacceptable for 100 tests, then 10 are unacceptable for 1,000 tests.

Bonferroni test (p. 112)
The best-known FWER test is called the Bonferroni test which adjusts for the multiple tests. Given the chance that one test could randomly show up as significant, the Bonferroni requires the confidence level to increase.

Holm test (p. 112)
The Holm test captures the information in the distribution of the test statistics. The Holm test is less stringent than the Bonferroni because the hurdles are relaxed after the first test. However, the Holm still fits into the category of the FWER.

BHY hurdle (p. 112)
BHY allows for an expected proportion of false discoveries, which is less demanding than the absolute occurrence of false discoveries under the FWER approaches. Similar to the Holm test, BHY also relies on the distribution of test statistics. However, in contrast to the Holm test that begins with the most significant test, the BHY approach starts with the least significant test. Also, the hurdles are larger and thus more lenient that the Bonferroni implied hurdle.

Type I error (p. 113)
The Type I error is the false discovery (investing in an unprofitable trading strategy).

Type II error (p. 113)
The Type II error is missing a truly profitable trading strategy.

8.8.1 Using statistical techniques to evaluate trading strategies in the presence of multiple tests

A. Describe why standard statistical tools, such as p-values and t-statistics can lead to false discoveries in the presence of multiple tests.

The statement about the false discovery percentage is conditional on an independent test: this means there is a single test. With multiple tests, we need to adjust our hurdles for establishing statistical significance.

B. Calculate the t-statistic based on reported Sharpe ratio for testing a single trading strategy.

$T - \text{statistic} = \text{Sharpe Ratio} \times \sqrt{\text{Number of years}}$

C. Describe the Bonferroni tests in the context of the family-wise error rate (FWER) approach to adjusting p-values for multiple tests.

The best-known FWER test is called the Bonferroni test which adjusts for the multiple tests. Given the chance that one test could randomly show up as significant, the Bonferroni requires the confidence level to increase. Instead of 5%, you take the 5% and divide by the number of tests, that is, $5\%/10 = 0.5\%$. Again equivalently, you need to be 99.5% confident with 10 tests that you are not making a single false discovery. In terms of the t-statistic, the Bonferroni requires a statistic of at least 2.8 for 10 tests. For 1,000 tests, the statistic must exceed 4.1.

D. Describe the Holm method in the context of the family-wise error rate (FWER) approach to adjusting p-values for multiple tests.

The Holm test captures the information in the distribution of the test statistics. The Holm test is less stringent than the Bonferroni because the hurdles are relaxed after the first test. However, the Holm still fits into the category of the FWER. The Holm method begins by sorting the tests from the lowest p-value (most significant) to the highest (least significant), and comparing a threshold computed with the Holm function. In contrast to the Bonferroni, which has a single threshold for all tests, the other tests will have a different hurdle under Holm. Starting from the first test, we sequentially compare the p-values with their hurdles. When we first come across the test such that its p-value fails to meet the hurdle, we reject this test and all others with higher p-values.

E. Recognize and apply the Holm function to calculate adjusted p-values.

The Holm method begins by sorting the tests from the lowest p-value (most significant) to the highest (least significant). Starting from the first test ($k = 1$), the Holm function is evaluated: $p_k = \frac{\alpha}{M+1-k}$ where α is the level of significance (say, 0.05) and M is the total number of tests

F. Understand the process of accepting and rejecting tests using Holm method.

The Holm method begins by sorting the tests from the lowest p-value (most significant) to the highest (least significant). Starting from the first test ($k = 1$), the Holm function is evaluated, which gives the hurdle (observed p-value must be lower than the hurdle). Notice the hurdle for the first test is identical to the Bonferroni. tarting from the first test, we sequentially compare the p-values with their hurdles. When we first come across the test such that its p-value fails to meet the hurdle, we reject this test and all others with higher p-values.

G. Describe the false discovery approach to adjusting p-values in the presence of multiple tests.

The false discovery rate approach allows an expected proportional error rate. As such, it is less stringent than both the Bonferroni and the Holm test. Again, we sort the tests. Similar to the Holm test, BHY also relies on the distribution of test statistics. However, in contrast to the Holm test that begins with the most significant test, the BHY approach starts with the least significant.

H. Recognize and apply the BHY formula to calculate adjusted p-values.

The BHY formula is: $p_k = \frac{k \times \alpha}{M \times c(M)}$, where $c(M)$ is a simple function that is increasing in M and equals 2.93 when M = 10.

I. Understand the process of accepting and rejecting tests using BHY method.

We sort the tests from the lowest p-value (most significant) to the highest (least significant). Starting from the last test, we sequentially compare the p-values with their thresholds. When we first come across the test such that its p-value falls below its threshold, we declare this test significant and all tests that have a lower p-value.

J. Explain the relationship between avoiding false discoveries and missing on profitable opportunities.

Multiple testing adjusts the hurdle for significance because some tests will appear significant by chance. The downside of doing this is that some truly significant strategies might be overlooked because they did not pass the more stringent hurdle. This is the classic tension between Type I errors and Type II errors. The Type I error is the false discovery (investing in an unprofitable trading strategy). The Type II error is missing a truly profitable trading strategy.

Reading 8.9 *Big Data and Machine Learning in Quantitative Investments (ch. 10).*

Guida, T. (2019). Big Data and Machine Learning in Quantitative Investments. West Sussex, UK: John Wiley & Sons Ltd. Chapter 10.

Keywords

Mainstream (p. 336)
The news articles produced by the mainstream news providers like Thomson Reuters, Bloomberg and Factset which are usually accessed via news feeds services provided by the vendors.

Primary source (p. 336)
Primary information sources that journalists research before they write articles include Securities and Exchange Commission (SEC) filings, product prospectuses, court documents and merger deals. We can further categorize primary source news as scheduled or unscheduled.

Social media (p. 337)
Social media sources can include tweets, blogs and personal posts. The use of social media is also gaining popularity for disseminating company information as an alternative to primary information providers.

Sentiment analysis (p. 339)
Sentiment analysis aims to analyze the opinion that a body of text conveys on a particular subject or entity

Natural language processing (p. 347)
NLP is a subfield of artificial intelligence concerned with programming computers to process textual data in order to gain useful insights. It transcends many disciplines in various guises and names, such as textual analysis, text mining, computational linguistics and content analysis.

Tokenization (p. 348)
The first step in most NLP applications is usually to tokenize the raw text data, which breaks it up into units called tokens by locating word bound- aries.

Word filter (p. 348)
Vocabulary in the context of NLP refers to the set of distinct words that appear in the corpus to be processed. A common way to limit the vocabulary is by term frequency, where only the words occurring more frequently are retained. Due to ambiguity, an excessively large vocabulary is usually more prone to error

Part of Speech Tagging (p. 349)
The process of assigning a token to its grammatical category, e.g. verb, noun, etc., in order to understand its role within the sentence.

Stemming (p. 350)
Stemming usually operates on a single word without knowledge of the context and uses a crude heuristic process that removes derivational affixes in the hope of reducing the word to its stem.

Lemmatization (p. 350)
Lemmatization aims to achieve this in a more principled manner with the use of a vocabulary and morphological analysis of words to return the base or dictionary form of a word, also known as its lemma. Unlike stemming, lemmatization handles not only basic word variations like singular vs plural but also synonyms.

Naive Bayes (p. 355)
One of the most commonly used supervised methodologies in NLP is the Naive Bayes model, which assumes that all word features are independent of each other given the class labels. Due to this simplifying but largely false assumption, Naive Bayes is very compatible with a bag-of-words word representation. Despite

its simplifying assumptions, it often comes head to head and at times even outperforms more complicated classifiers.

FNN (p. 363)
A feed-forward neural network (FNN) contains a (possibly large) number of simple neuron-like nodes, organized in layers. Data enters the network at the input layer and, as the name suggests, is fed forward through the network, layer by layer, until it arrives at the output layer. Nodes in a layer alone never have connections and in general two adjacent layers are fully connectedT. here are no cycles or feedback loops between layers. FNNs were the first and the simplest network structure to be devised.

RNN (p. 363)
A recurrent neural networks (RNN) contains recursive loops that allow it to exhibit dynamic temporal behaviour and capture long-term dependencies in sequential inputs. As a result, RNNs are suitable for NLP since they can evaluate each word/token input in context of the words that appear before it. However, the training of such architectures can be problematic due to the recursive nature of the information and gradient flow. In order to alleviate these, different gating mechanisms have been proposed, resulting in various RNN architectures such as long short-term memory (LSTM).

CNN (p. 363)
Convolutional neural networks (CNN) consist of a sequence of convolutional blocks in between the input and output layers. For NLP applications, a single convolutional block usually consists of a convolution kernel that convolves the previous layer's input over a single spatial dimension, followed by a max pooling layer for down-sampling the convolutional output to produce a tensor of outputs. A CNN can combine basic features whereby words form n-grams, phrases and sentences.

8.9.1 Natural language processing of financial news

A. Describe the three categories of sources of news data.

1. Mainstream: The news articles produced by the mainstream news providers like Thomson Reuters, Bloomberg and Factset which are usually accessed via news feeds services provided by the vendors.
2. Primary source: Primary information sources that journalists research before they write articles include Securities and Exchange Commission (SEC) filings, product prospectuses, court documents and merger deals. We can further categorize primary source news as scheduled or unscheduled.
3. Social media: With news from social media services, the barrier for entry and consequentially the signal to noise ratio is low. Social media sources can include tweets, blogs and personal posts. Despite the high level of noise and the lack of verification and editorial, social media can still serve as a valuable information source due to the blisteringly fast speeds that news is made available online. The use of social media is also gaining popularity for disseminating company information as an alternative to primary information providers.

B. Explain the advantages and disadvantages of using the new category of social media.

- Despite the high level of noise and the lack of verification and editorial, social media can still serve as a valuable information source due to the blisteringly fast speeds that news is made available online.
- One of the arguments in favour is that blog posts or tweets allow one to tap into the 'Wisdom of Crowds', which refers to the phenomenon that the aggregation of information provided by many

individuals often results in predictions that are better than those made by any single member of the group.

- However, social media posts may lack credibility as most providers have no mechanism to fact-check the information shared or to incentivize high-quality information. Anecdotal evidence from contemporaneous political elections in developed countries demonstrates that the information in social media posts may be intentionally misleading to serve the posters' own agenda.

C. Describe sentiment analysis.

Sentiment analysis aims to analyze the opinion that a body of text conveys on a par- ticular subject or entity. In the financial domain, the primary motivation behind most sentiment analysis tasks is to relate these opinions to the directionality of future security returns. As a supervised learning exercise, sentiment analysis may involve manually labelling a training dataset with different sentiment categories/scores before feeding these into a classification or regression algorithm. An alternative to manual labelling is to compile a 'word list' that associates words with distinct sentiments.

D. Describe the word list approach to sentiment analysis.

An alternative to manual labelling is to compile a 'word list' that associates words with distinct sentiments. Using such a list, a researcher can count the number of words associated with a particular sentiment, whereby a higher proportion of pessimistic words in a news article indicates a negative sentiment. While an NLP practitioner may choose to compile and use their proprietary word lists, there also exist publicly available ones

E. Describe the three challenges associated with sentiment analysis.

The common challenges associated with sentiment analysis in finance include

1. the difficulty in extracting a consistent sentiment
2. the need to determine which securities a particular news item refers to (and to what extent), and
3. filtering novel articles from those that have been recycled

F. Describe the four steps – pre-processing, feature representation, inference and evaluation – in applying NLP to texts.

1. Pre-processing: In order to feed news data into the computer, we need to transform a collection of characters into a format that captures the information conveyed in an unambiguous and precise manner. These first steps in most NLP applications include:
    - tokenizing the raw text data, which breaks it up into units called tokens by locating word boundaries.
    - limiting the size of the vocabulary
    - assigning a token to its grammatical category, e.g. verb, noun, etc., in order to understand its role within the sentence.

- stemming and lemmatization are used to reduce words from their derived grammatical forms to their base forms.

2. Feature representation: The preprocessed tokens need to be translated into predictive features. The most commonly used feature representation technique in NLP is the bag-of-words model, according to which a document is encoded as an (unordered) set of its words, disregarding grammar and word order but retaining multiplicity. Representation of a word's meaning based on its neighbouring words is one of the most common extensions beyond the simple bag-of-words approach, such as N-gram models

3. Inference: Inference in ML falls under three broad categories, namely supervised, unsupervised and reinforcement learning. Inference from an NLP application can be used to aid the decision making by humans, where a utility function is applied to convert the inference into a decision. This utility function can be as simple as a probability threshold, or an implicit weighing down of pros and cons in a domain expert's brain. Alternatively, inference can directly be translated into an action by the computer as part of an automated quantitative strategy.

4. Evaluation: In general, inference in NLP tasks is assessed in a similar way to any other machine learning analysis. For regression models that try to predict a continuous dependent variable, such as return or volatility, evaluation metrics are usually various error terms including, but not limited to, root mean square error (RMSE), mean absolute error (MEA) and mean squared error (MSE). For classification exercises, where the output is categorical, there exist numerous confusion matrix-based metrics, such as accuracy, precision and recall.

G. Understand aspects of pre-processing

1. tokenization: breaks it raw text data into units called tokens by locating word boundaries. However, tokens do not have to be just words; they can be numbers or punctuation marks. In fact, tokenization may be bundled with removing punctuation and stop-words, which are extremely common words that may be of little value for the NLP task.

2. vocabulary: the set of distinct words that appear in the corpus to be processed. A common way to limit the vocabulary is by term frequency, where only the words occurring more frequently are retained. Due to ambiguity, an excessively large vocabulary is usually more prone to error when compared with tests focusing on fewer unambiguous words or phrases.

3. part of speech: the process of assigning a token to its grammatical category, e.g. verb, noun, etc., in order to understand its role within the sentence.

4. stemming and lemmatization: reduce words from their derived grammatical forms to their base forms. Stemming usually operates on a single word without knowledge of the context and uses a crude heuristic process that removes derivational affixes in the hope of reducing the word to its stem. In contrast, lemmatization aims to achieve this in a more principled manner with the use of a vocabulary and morphological analysis of words to return the base or dictionary form of a word, also known as its lemma

H. Understand aspects of representation of words as features:

1. bag of words: The most commonly used feature representation technique in NLP, according to which a document is encoded as an (unordered) set of its words, disregarding grammar and word order but

retaining multiplicity.

2. N-gram: partially addresses the lack of context by storing sequences of words that occur next to each other. So, for example, a two-word n-gram model, i.e. a bigram model, parses the text into a set of consecutive pairs. This clearly helps with capturing the co-occurrences of words. Theoretically, with larger n, a model can store more contextual information.

3. distributed representation: also known as a vector space model, or vector embedding, represents words in a continuous vector space where semantically similar words are grouped together. The different approaches that leverage distributed representations can be divided into two categories. The first is coined count-based methods (e.g. latent semantic analysis (LSA)), which quantify the co-occurrence frequencies of words with other words in a large text and map these statistics down to a dense vector for each distinct word. The second category is the so-called predictive methods that are trained by iteratively updating the vector coordinates of words in order to more accurately predict a word from its neighbours: The end result from both models is the same as that of count-based models, a set of dense embedding vectors for each distinct word in the vocabulary.

Reading 8.10 *Zero-Revelation RegTech: Detecting Risk through Linguistic Analysis of Corporate Emails and News.*

Das, S., S. Kim and B. Kothari. (2019). Zero-Revelation RegTech: Detecting Risk through Linguistic Analysis of Corporate Emails and News. The Journal of Financial Data Science, 1(2), 8-34.

Keywords

Natural language processing (p.8)
A fast-growing area of data science for systematically and efficiently drawing appropriate inferences from large bodies of unstructured text.

RegTech (p. 8)
RegTech is a new concept in the application of data science to financial risk management. Investopedia defines RegTech as a portmanteau of "regulatory technology" that was created to address regulatory challenges in the financial services sector through innovative technology.

Fintech (p. 11)
Address challenges in the financial services sector through innovative technology

Fabula (p. 11)
The narrative story in text

Syuzhet (p. 11)
Organization of text

Topic modeling (p. 11)
Topic modeling operates at higher level, such as at the paragraph or document level in its entirety, referred to as the 'episode' or 'topic unit'.

Episode (p. 11)
A topic unit obtained from topic modeling operating at the paragraph or document level

Topic Net sentiment (p. 16)
Net sentiment is the difference between the number of positive and negative words. This metric is often scaled by the sum of positive and negative words;

Polarity (p. 16)
Net sentiment

Disagreement (p. 16)
Disagreement is defined as one minus the absolute difference between the number of positive and negative words, scaled by the sum of positive and negative words

Term document matrix(TDM) (p.29)
Matrix representing term-frequency by document

Document term matrix (DTM) (p. 29)
Matrix representing document by term-frequency

8.10.1 Using linguistic analysis to perform risk analysis of investments.

A. Explain the difficulties associated with manual parsing of unstructured text.

Manual parsing is not only inefficient, but also raises substantial privacy concerns.

B. Describe the concept of RegTech.

RegTech is a new concept in the application of data science to financial risk management.

C. Describe how content and structure of emails could be used for risk analysis.

Indicators within employees' sent email content and sender/recipient networks can effectively predict changes in risk and subsequent performance. We examine not only sentiment-based indicators contained in the message bodies, but also non-textual structural characteristics such as the number of emails sent, the average length per email, and the shifting sender/recipient networks within the company over time. Of particular value are the non-content-based structural/network indicators of potential trouble, since email content may be easier to control or manipulate than the connectivity of a network. Network changes evidenced by shifting email clusters can indicate sources ripe for investigation, as fraudulent activity among corporate employees tends to occur in social-network clusters

D. Explain the effectiveness of textual versus hard numbers in corporate risk analysis.

Unlike numerical data, textual data may be able to extricate corporate risk better than hard numbers, as it contains richer characteristics and additional nuances such as sentiment, word length, the readability of text, size of the text file. Analyzing text from communications within the firm offers a unique and new approach to detecting escalations in operations risk. Furthermore, it also offers a way to assess or predict the impact of the first three risks on the firm's financial condition. For instance, by looking at the change in the frequency of risk words over time or by analyzing shifting email network metrics, one may be able to detect the sudden emergence of any of these forms of risk that may otherwise remain undetected for a longer period of time.

E. Define RegTech.

Investopedia defines RegTech as a portmanteau of "regulatory technology" that was created to address regulatory challenges in the financial services sector through innovative technology.

F. Apply the net sentiment metric to calculate the polarity of a text using its Pos and Neg figures.

Net sentiment is the difference between the number of positive and negative words, scaled by the sum of positive and negative words: $\frac{Pos-Neg}{Pos+Neg}$

G. Calculate the disagreement measure of a text using Pos and Neg figures of a text.

Disagreement is defined as one minus the absolute difference between the number of positive and negative words, scaled by the sum of positive and negative words: $1 - \frac{|Pos-Neg|}{Pos+Neg}$

H. Describe the findings of the paper regarding the effectiveness of email length as predictor of risk analysis of Enron.

A strong predictive association is found between email length and subsequent stock performance. Email length is a potentially powerful indicator of corporate malaise, with email length trending downward along with stock returns and stock prices as we approach Enron's demise. The fact that email length is the best predictor of poor performance is interesting in reflecting that email size matters more than email content, and more than news content. This suggests that simple quantification of email traffic may be useful and more complicated metrics are not needed. How much people talk is less likely to be manipulated, and is more important than what they say. The sentiment conveyed directly in email content itself is much easier to suppress than email length

