Outline of FDP Exam Study Guide Contents

Introduction to FDP Program ................................................................................................................................................. 3

Data Science ......................................................................................................................................................................................... 4

FDP Program: Prerequisites ................................................................................................................................................................. 5
Datacamp .................................................................................................................................................................................................................. 6
Dataquest .................................................................................................................................................................................................................. 8
Metis ................................................................................................................................................................................................................... 10

FDP Examination ................................................................................................................................................................................. 12

The FDP Curriculum: Outline .......................................................................................................................................................... 13

Other Study Tools and Resources ................................................................................................................................................... 13

The FDP Curriculum: Reading List .................................................................................................................................................. 14
Introduction to Data Science & Big Data ........................................................................................................................................... 14
Data Mining & Machine Learning: Introduction ............................................................................................................................ 14
Data Mining & Machine Learning: Regression, LASSO, Predictive Models,
Time Series & Tree Models .......................................................................................................................................................... 14
Data Mining & Machine Learning: Classification & Clustering ...................................................................................................... 15
Data Mining & Machine Learning: Performance Evaluation, Backtesting & False Discoveries .................................................. 15
Data Mining & Machine Learning: Representing & Mining Text .................................................................................................. 15
Big Data, Data Mining & Machine Learning: Ethical & Privacy Issues .......................................................................................... 15
Big Data & Machine Learning in the Financial Industry ................................................................................................................ 16

Learning Objectives ............................................................................................................................................................................ 17

Topic 1. Introduction to Data Science & Big Data .......................................................................................................................... 17
Topic 2. Data Mining & Machine Learning: Introduction ............................................................................................................... 23
Topic 3. Data Mining & Machine Learning: Regression, LASSO, Predictive Models,
Times Series & Tree Models ......................................................................................................................................................... 28
Topic 4. Data Mining & Machine Learning: Classification & Clustering ........................................................................................ 37
Topic 5. Data Mining & Machine Learning: Performance Evaluation,
Backtesting & False Discoveries ................................................................................................................................................. 40
Topic 6. Data Mining & Machine Learning: Representing & Mining Text ..................................................................................... 44

Action Words .................................................................................................................................................................................. 62
Introduction to FDP Program

The FDP Institute® was founded by the Chartered Alternative Investment Association® to create the FDP® charter. It is the only globally recognized professional designation in the area of financial data science, an increasingly important part of the financial services industry.

The FDP is a self-study program designed to educate financial professionals about the emerging field of data science and its practical applications in the financial sector, and to award the FDP charter to candidates who complete its exam successfully. The FDP curriculum provides breadth and depth by first requiring candidates to complete the prerequisites of the FDP program, which consist of one or more short online classes covering the foundations of mathematics, statistics, and Python programming. Candidates who satisfy the prerequisites will be able to sit for the FDP exam using the readings and learning objectives discussed in this document.

The university faculty and industry practitioners who have helped create the FDP Charter program bring years of experience in the financial services industry. Consequently, the curriculum is consistent with recent advances in the applications of data science to the financial industry.

Passing the FPD examination is an important accomplishment and will require a significant amount of preparation. All candidates will need to study and become familiar with the FDP curriculum material in order to develop the knowledge and skills necessary to be successful on examination day.

This study guide is organized to facilitate quick learning and easy retention. Each topic is structured around learning objectives that define the content to be tested on the exam. The learning objectives are an important way for candidates to organize their study, as they form the basis for examination questions. All learning objectives reflect the content in the FDP curriculum and all exam questions are written to directly address the learning objectives. A candidate who is able to meet all learning objectives in the study guide should be well prepared for the exam. For these reasons, we believe that the FDP Institute has built a rigorous program with high standards, while also maintaining an awareness of the value of candidates’ time.

Candidates for the FDP exam are required to have completed the online prerequisite classes and therefore are assumed to have an understanding of the central concepts of quantitative analysis. In addition, since the FDP program is designed for finance professionals, it is assumed that candidates have an understanding of finance. This includes awareness of the roles and characteristics of various financial institutions and instruments as well as the financial models employed by these institutions to value the instruments and measure risk. These concepts are covered in CAIA®, CFA®, and FRM® exams, and in dedicated undergraduate or graduate courses covering financial markets, investments, and risk management.
Data Science

Understanding data science begins with three basic areas:

• **Mathematics, probability and statistics:** Data scientists use concepts from these disciplines to explore data, develop and test a hypothesis, and present the results to end users.

• **Computer programming:** Coding allows a data scientist to manipulate large data sets, calculate critical variables needed for statistical analysis and create output that is useful to decision makers.

• **Domain knowledge:** This distinguishes a data scientist from a statistician or a computer programmer. A data scientist must be able to formulate the questions that are critical to a data-driven decision-making process, identify the variables that will enter the algorithm, and present the results in a manner that is most useful to decision makers.
FDP Program: Prerequisites

Before candidates can enroll in the FDP program, they must demonstrate their knowledge of:

- Computer programming using Python
- Probability and statistical methods

FDP candidates can satisfy the prerequisite requirements by completing a set of short online classes offered by a select group of online educational organizations approved by the FDP Institute. Only one set of classes offered by these providers must be completed to meet the pre-requisite requirements of the FDP program. Depending on the candidate’s background, it is estimated to take 25-60 hours to complete the prerequisites. No programming background is required to complete the pre-requisite requirement.

As of now, the following organizations have been designated to satisfy a set of prerequisites to the core topics of data science and its applications in the financial services industry.

- **Datacamp**: [https://www.datacamp.com/](https://www.datacamp.com/)
- **Dataquest**: [https://www.dataquest.io/](https://www.dataquest.io/)
- **Metis**: [https://www.thisismetis.com/](https://www.thisismetis.com/)

The list of prerequisite classes for each approved online provider appears on FDP Institute’s website as well as in this document.

The approved online courses offered by Dataquest and Datacamp are available as soon as a candidate registers on their respective sites. The approved course offered through Metis is offered once a month throughout the year. All three providers offer limited free access to their classes. Candidates should take advantage of the limited free access to determine which platform’s approach is more suited to their needs. Candidates cannot mix and match classes from different providers. Finally, the approved classes offered by the three platforms assume no prior knowledge of Python language or any specific computer programming language.

The topics covered by these prerequisite classes:

- **Mathematical foundation**: e.g., vectors and matrices, summation and products, matrix manipulation, logarithms, and exponents.
- **Data and statistical analysis**: e.g., histograms, descriptive statistics, random sampling, distribution of statistics, tests of hypothesis, Bayesian analysis, simple and multiple regression, principal component analysis.
- **Basic coding skills Python**: e.g., reading data, preparing data, applying the statistical techniques described above to the data.

The following classes should be completed to satisfy the prerequisite requirement of FDP program. The Candidates’ Handbook, which can be found on FDP’s website, describes the procedure for sending proof of successful completion of the prerequisite programs to the FPD Institute.
Datacamp

Candidates can access Datacamp classes through their website at https://www.datacamp.com/. Candidates are responsible for the cost of classes offered at Datacamp. Candidates are encouraged to take advantage of limited free access offered by Datacamp to evaluate its method of teaching. The classes listed below are short and, depending on the candidate’s background, each one can be completed within 3-7 hours.

1. **Introduction to Python**
   https://www.datacamp.com/courses/intro-to-python-for-data-science
   An introduction to the basic concepts of Python. Candidates will learn how to use Python both interactively and through a script. Candidates will learn about NumPy, a very important Python package, variables, and Python’s basic data types.

2. **Intermediate Python for Data Science**
   https://www.datacamp.com/courses/intermediate-python-for-data-science
   The intermediate python course is crucial to applying the techniques that candidates will learn in the FDP program. Candidates will learn to visualize real data with Matplotlib’s functions and to work with different data structures.

3. **Pandas Foundations**
   https://www.datacamp.com/courses/pandas-foundations
   Pandas dataframes are the most widely used representation of complex data collections within Python. Whether in finance, scientific fields, or data science, familiarity with Pandas is essential. This class teaches candidates to work with real-world data sets containing both string and numeric data, often structured around time series.

4. **Manipulating DataFrames with Pandas**
   https://www.datacamp.com/courses/manipulating-dataframes-with-pandas
   In this class, candidates will learn how to leverage Pandas’ extremely powerful data manipulation engine to get the most out of their data. Candidates will learn how to extract, filter, and transform data from dataframes in order to gain further insights into the data.

5. **Importing and Managing Financial Data in Python**
   In this class, candidates will learn how to get data out of Excel into pandas and back. Candidates will also learn how to pull stock prices from various online APIs like Google or Yahoo! Finance, macro data from the Federal Reserve, and exchange rates from OANDA. Finally, candidates will learn how to calculate returns for various time horizons, analyze stock performance by sector for IPOs, and calculate and summarize correlations.
Datacamp continued

6. Statistical thinking with Python I
   In this class, candidates will start building the foundation they need to think statistically. This refresher class on statistics will use Python to demonstrate the applications of important statistical techniques.

7. Statistical thinking with Python II
   In this class, candidates will to perform the two key tasks in statistical inference, parameter estimation, and hypothesis testing. Candidates will work with real data sets as they learn.
Dataquest

Candidates can access Dataquest classes through their website at https://www.dataquest.io/. Candidates are responsible for the cost of classes offered at Dataquest. Candidates are encouraged to take advantage of limited free access offered by Dataquest to evaluate its method of teaching. The classes listed below are short and, depending on the candidate’s background, each one can be completed within 3-7 hours.

1. **Python for Data Science: Fundamentals**  
   This class covers the basics of Python, including the execution of codes, printing of a statement, commenting on codes, understanding of variables, employment of conditional statements and understanding of an assignment operator.

2. **Pandas and NumPy Fundamentals**  
   [https://www.dataquest.io/course/pandas-fundamentals/](https://www.dataquest.io/course/pandas-fundamentals/)  
   This class introduces two of the most important packages in Python – NumPy and Pandas. An understanding of both packages is essential for cleaning, manipulating and visualization of data.

3. **Exploratory Data Visualization**  
   This class introduces Matplotlib, the main package for graphing in Python. It shows how to plot line charts, use sub-plots to plot multiple charts, and plot bar charts, scatter plots, histograms and box plots.

4. **Data Cleaning and Analysis**  
   [https://www.dataquest.io/course/python-datacleaning/](https://www.dataquest.io/course/python-datacleaning/)  
   This class builds on the prior knowledge of Pandas to introduce candidates to data aggregation and working with dataframes.

5. **Statistics Fundamentals**  
   [https://www.dataquest.io/course/sql-fundamentals/](https://www.dataquest.io/course/sql-fundamentals/)  
   This class builds on the prior knowledge of data analysis and helps candidates to work with more advanced tools offered by Python. The course uses Python to provide an introductory refresher course in statistics.

6. **Statistics Intermediate: Averages and Variability**  
   In this class, candidates will learn how to use Python to summarize the distribution of a variable with a single value, such as the mean, the weighted mean, the median, the mode, variance, and standard deviation.
Dataquest continued

   https://www.dataquest.io/course/probability-statistics-intermediate/
   This class will complete the refresher classes in probability and statistics. Candidates will learn to use Python to perform basic statistical techniques that most candidates are likely to be familiar with already.

8. Linear Algebra for Machine Learning
   https://www.dataquest.io/course/linear-algebra-for-machine-learning/
   This refresher class covers basic concepts of linear algebra, which are important to understanding machine learning algorithms that are covered in the FDP program.
Metis

Candidates can access the Metis course through their website at https://www.thisismetis.com/. Candidates are responsible for the cost of the course offered at Metis. Candidates are encouraged to take advantage of free sample videos that are offered by Metis to evaluate its method of teaching.

Unlike the classes offered through Dataquest and Datacamp, which consist of pre-recorded videos and texts, Metis offers a live online classes with dedicated instructors who are ready to answer your questions during the live sessions as well as later during office hours. Further, while Dataquest and Datacamp classes can be taken at any time by a candidate, Metis’s live classes are offered monthly throughout the year. Enrolled candidates will be able to watch a video of the class should they miss a session. The approved prerequisite course offered by Metis lasts 6 weeks.

The single course offered by Metis and required as prerequisite for the FDP exam is titled *Beginner Python and Math for Data Science*, and it consists of the following 6 topics, which is usually covered over 12 live sessions. https://www.thisismetis.com/courses/beginner-python-and-math-for-data-science

1. **Python Basics**
   Candidates are introduced to programming in Python. Candidates will learn about Jupyter Notebooks – a popular platform for running Python programs. This part of the course will cover the basics of programming, including data structures, data operations, if else statements, for and while loops, and logical operations.

2. **Python Advanced**
   This segment of the course covers advanced functionality in Python, including functions, debugging, error handling, string manipulations, and writing efficient code.

3. **Python Mathematical Libraries**
   Candidates will learn about using libraries that are useful for data manipulation and visualization. Candidates will learn to use NumPy, Pandas, and Matplotlib. These libraries will allow candidates to load and save data, manipulate data such as aggregating, filtering, detecting outliers, and visualizing.

4. **Linear Algebra**
   This segment of the course is a refresher in linear algebra. It will cover the fundamentals of linear algebra, including vectors, and vector manipulations, matrices and matrix manipulations, linear equations and solutions, eigenvalues and eigenvectors.
Metis continued

5. Calculus and Probability
   This module is a refresher in the fundamentals of calculus. It reintroduces students to such central concepts of calculus such as derivatives, integrals, determining local maximum and minimum, and limits. In addition, the module provides a refresher on central concepts of probability such as random variables, mean, variance, probability mass and density functions, and cumulative distribution functions.

6. Statistics
   This final refresher module covers a few important statistical concepts such as ANOVA, hypothesis testing and p-value, and confidence intervals.
FDP Examination

The FDP examination, administered twice annually, is a four-hour computer-administered examination that is offered at test centers throughout the world. The FDP examination is comprised of 75 multiple choice questions weighted as 60% of the total points and two to three constructed response questions (multi-part essay type) weighted as 40% of the total points. The FDP exam will not contain any Python programming questions.

The FDP examination is based on its study guide, which is organized to facilitate quick learning and easy retention. Each topic is structured around learning objectives and keywords that define the content to be tested on the exam. The learning objectives and keywords are an important way for candidates to organize their study, as they form the basis for examination questions. All learning objectives reflect content in the FDP curriculum and all examination questions are written to directly address the learning objectives.

For additional information about the FDP examination please see the Candidate’s Handbook, which can be found on FDP Institute website.
The FDP Curriculum: Outline

Candidates for the FDP Charter will have to enroll in the self-study program created by the FDP Institute and follow its carefully designed study guide. To become an FDP Charterholder, candidates must pass the FDP exam and become a member of the FDP Institute. The rest of this document discusses the FDP curriculum. Below is the outline of the curriculum:

<table>
<thead>
<tr>
<th>Topics</th>
<th>Approximate Weight %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction to Data Science &amp; Big Data</td>
<td>5-10</td>
</tr>
<tr>
<td>2. Data Mining &amp; Machine Learning: Introduction</td>
<td>5-10</td>
</tr>
<tr>
<td>4. Data Mining &amp; Machine Learning: Classification &amp; Clustering</td>
<td>5-15</td>
</tr>
<tr>
<td>5. Data Mining &amp; Machine Learning: Performance Evaluation, Backtesting &amp; False Discoveries</td>
<td>5-10</td>
</tr>
<tr>
<td>6. Data Mining &amp; Machine Learning: Representing &amp; Mining Text</td>
<td>5-10</td>
</tr>
<tr>
<td>7. Big Data, Data Mining &amp; Machine Learning: Ethical &amp; Privacy Issues</td>
<td>5-10</td>
</tr>
<tr>
<td>8. Big Data and Machine Learning in the Financial Industry</td>
<td>30-50</td>
</tr>
</tbody>
</table>

Other Study Tools and Resources

In addition to this study guide and candidates handbook, the FDP Institute website directs you to the readings that the curriculum is based on. The readings are detailed below by topic area, and include textbooks, often used across topics, as well as several individual articles that are usually topic-specific. Both types of readings may either be available for purchase from Amazon or publishers, or whenever possible, they are posted on the FDP Institute and freely available to registered candidates.

Page Number References for Key Words

For candidate’s convenience, a set of eight articles published by IPR Journals is provided in one collection titled *Big Data & Machine Learning in the Financial Industry: Readings for the Financial Data Professional Exam* and is available at a discounted price of $99 for registered candidates. In this collection there are two sets of pages numbers: one corresponding to the collections table of contents, and one corresponding to the each article’s page number in he original journal. For the purposes of referencing the location of keywords in the study guide, the page numbers always refer to the article’s page number in the original journal.
The FDP Curriculum: Reading List

Introduction to Data Science & Big Data


Data Mining & Machine Learning: Introduction


Data Mining & Machine Learning: Regression, LASSO, Predictive Models, Time Series & Tree Models


The FDP Curriculum: Reading List continued

Data Mining & Machine Learning: Classification & Clustering

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

Data Mining & Machine Learning: Representing & Mining Text

Big Data, Data Mining & Machine Learning: Ethical & Privacy Issues
The FDP Curriculum: Reading List *continued*

**Big Data & Machine Learning in the Financial Industry**

  https://www.bryankellyacademic.org/
  http://dachxiu.chicagobooth.edu/download/ML.pdf
  The paper posted on the FDP Institute’s website will be the version used for exam questions.
Learning Objectives

**Topic 1. Introduction to Data Science & Big Data**

**Readings**


**Keywords**

<table>
<thead>
<tr>
<th>Data mining (p. 2)</th>
<th>Data science (p. 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Churn (p. 4)</td>
<td>Data-driven decision making (p. 5)</td>
</tr>
<tr>
<td>Data engineering (p. 5, 7)</td>
<td>Data-analytic thinking (p. 12)</td>
</tr>
<tr>
<td>Target (p. 24)</td>
<td>Label (p.24)</td>
</tr>
<tr>
<td>Unsupervised data mining (p. 24)</td>
<td>Supervised data mining (p. 25)</td>
</tr>
</tbody>
</table>

**Learning Objectives**

Demonstrate proficiency in the following areas:

1) **Data analytic thinking (Ch. 1)**

   **For example:**
   
   o Discuss the ubiquity of data opportunities.
   
   o Define data science, engineering, and data-driven decision making.
   
   o Explain data and data science capability as a strategic asset.
   
   o Describe data-analytic thinking.
   
   o Compare data science and the work of the data scientist.
Learning Objectives continued

2) Business problems and data science solutions (Ch. 2)  
   For example:
   o Describe how one transitions from business problems to data mining tasks.
   o Compare supervised methods to unsupervised methods.
   o Describe the difference between data mining and using the results of data mining.
   o Describe key aspects of the data mining process, including business understanding, data understanding, data preparation, modeling, and evaluation.


Keywords

<table>
<thead>
<tr>
<th>Alternative data (p. 5)</th>
<th>Discretionary (p. 9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Data (p. 5)</td>
<td>Quantitative (p. 9)</td>
</tr>
<tr>
<td>Data Science (p. 5)</td>
<td>Financial Data (p. 9)</td>
</tr>
<tr>
<td>Data Analytics (p. 5)</td>
<td>Market Data (p. 9)</td>
</tr>
<tr>
<td>Moore’s Law (p. 5)</td>
<td>Tagging (p. 11)</td>
</tr>
<tr>
<td>Bezos Law (p. 5)</td>
<td>Incrementalism (p.12)</td>
</tr>
<tr>
<td>Unstructured (p. 7)</td>
<td></td>
</tr>
</tbody>
</table>

Learning Objectives

Demonstrate proficiency in the areas of:

1) Defining big data  
   For example:
   o Define alternative data, big data, data science and data analytics.
   o Discuss the recent phenomenon of data proliferation in terms of:
     • Moore’s Law and Bezos law
     • Digitization of consumer lifestyles and business processes
     • Business intelligence tools
     • Big data intermediaries
   o Compare types and sources of alternative data.
   o List five characteristics used to assess the quality of a big data source.
Learning Objectives continued

2) The benefits and limitations of big data for investment decisions

For example:

- 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).

- Describe a process of harnessing data-driven insights including:
  - Sourcing
  - Backtesting
  - Tagging
  - Mapping, and
  - Visualization

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

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

For example:

- Describe barriers to entry and the learning curve.

- Explain the need to critically evaluate data sources.

- Discuss three major facets of refining data.

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

4) Implementation and costs involved in utilizing alternative data in the investment process

For example:

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

- Discuss resource requirements and methods of reducing their costs.


Keywords

Quantitative fundamental analysis (p. 60, 62)
Learning Objectives continued

Learning Objectives
Demonstrate proficiency in the following areas:

1) Facets of the big data phenomenon
   For example:
   o Describe three facets of the big data phenomenon.

2) Investment managers’ use big of data
   For example:
   o Describe the spectrum of big data adoption.
   o Identify the players driving the adoption of big data in investment firms.
   o Explain why there will be pressure to incorporate big data principles in investment research.
   o Describe third party providers that facilitate efforts to accelerate adoption of big data principles.

3) The expansion and transformation of quantitative fundamental analysis
   For example:
   o Describe the evolution of systematic investment management in terms of the complexity of data used and the type of databases used.
   o Compare the traditional decision-making process of a fundamental analyst with the potential decision-making processes that incorporate big data fundamentals.

4) Emerging portfolio management models that utilize big data principles
   For example:
   o 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.
   o Describe a new type of systematic portfolio management model that bases trade selection on the likely ‘future value’ of a company.
   o Describe intermediate approaches to big data adoption that use a subset of alternative data to improve portfolio management.
   o Describe how fundamental discretionary managers can benefit from big data models.
Learning Objectives continued


Keywords

<table>
<thead>
<tr>
<th>Alternative data life cycle (p.145)</th>
<th>Quantamental investing (p. 127)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fundamental prediction (p.124)</td>
<td>Exhaust data (p.151)</td>
</tr>
<tr>
<td>Fundamental law of active management (p.127)</td>
<td>Nowcasts (p.151)</td>
</tr>
<tr>
<td>Quant quake (p.110)</td>
<td></td>
</tr>
</tbody>
</table>

Learning Objectives

Demonstrate proficiency in the following areas:

1) **Taming big data (Ch. 2)**
   
   *For example:*
   
   - Discriminate between alternative data and big data.
   - Contrast drivers of adoption of alternative data with its challenges in the investment community.
   - Identify the largest categories of alternative data types in use today.
   - Evaluate the usefulness of an alternative data set.
   - Describe the likely attributes that differentiate alternative data sets in terms of cost.
   - Discuss some of the biggest alternative data trends.

2) **Implementing alternative data in an investment process (Ch. 4)**
   
   *For example:*
   
   - Describe the “quant quake” and how it motivated the search for alternative data.
     
     • Note that Table 4.1 is cut off in the ebook version. The full version of the table appears at the end of this section.
   - 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.
   - Discuss methods for improving the efficiency of evaluating datasets for the purpose of finding alpha.
   - Describe issues involved with selecting a data source for evaluation within the context of a quant equity process.
   - Explain why and under what circumstances a fundamental prediction may be more appropriate than an asset price prediction when working with alternative data.
   - Apply the fundamental law of active management and describe how it applies differently to discretionary managers than quant managers.
Learning Objectives continued

- Describe the transition from fundamental analysis to “quantamental analysis.”
- 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.

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

For example:
- Define general concepts within big data and alternative data including “exhaust data.”
- Compare traditional model building approaches and machine learning.
- Discuss how big data and alternative data can be used to improve economic forecasts and “nowcasts.”
- 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.

<table>
<thead>
<tr>
<th>TABLE 4.1</th>
<th>Average annualized return of dollar-neutral, equally-weighted portfolios of liquid US equities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>More crowded factors</td>
</tr>
<tr>
<td></td>
<td>Earnings yield (%)</td>
</tr>
<tr>
<td>2001-2007</td>
<td>11.00 14.76 35.09 20.28 8.64 3.60 17.10 9.78</td>
</tr>
<tr>
<td>7 Aug. 2007</td>
<td>-1.06 -0.11 -0.34 -0.50 -0.06 0.33 -0.85 -0.19</td>
</tr>
<tr>
<td>8 Aug. 2007</td>
<td>-2.76 -4.19 0.23 -2.24 -0.21 -0.04 0.21 -0.01</td>
</tr>
<tr>
<td>9 Aug. 2007</td>
<td>-1.66 -3.36 -3.41 -2.81 -0.29 -1.27 -0.23 -0.60</td>
</tr>
<tr>
<td>10 Aug. 2007</td>
<td>3.91 4.09 12.45 6.82 0.71 -0.01 1.70 0.80</td>
</tr>
</tbody>
</table>
Learning Objectives continued

Topic 2. Data Mining & Machine Learning: Introduction

Readings
   http://www-bcf.usc.edu/~gareth/ISL/

   http://neuralnetworksanddeeplearning.com/chap1.html


Keywords
- Statistical learning (p. 1)
- Classification problems (p.28)
- Semi-supervised learning (p.28)
- Quantitative variables (p.28)
- Qualitative response (p.28)
- Binary response (p.28)
- Regression (p.28)
- Predictors (p.29)
- Mean squared error (MSE) (p.29)
- Training MSE (p.30)
- Test data (p.30)
- Test MSE (p.30)
- Degrees of freedom (p.32)
- Cross validation (p.33)
- Expected test MSE (p.34)
- Bias (p.35)
- Bias-variance trade-off (p.36)
- Error rate (p.37)
- Indicator variable (p.37)
- Training error (p.37)
- Test error (p.37)
- Bayes classifier (p.37)
- Conditional probability (p.37)
- Bayes decision boundary (p.38)
- Bayes error rate (p.38)
- K-nearest neighbors (p.39)

Learning Objectives
Demonstrate proficiency in the following areas:

1) Organization and resources of the book *An Introduction to Statistical Learning: with applications in R* (Ch. 1).
   This chapter is assigned to facilitate your studies but no exam questions will be drawn from this chapter.
Learning Objectives continued

2) Statistical learning (Ch. 2.1)
   *For example:*
   - 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.
   - Compare and contrast parametric and non-parametric learning methods.
   - Describe the trade-offs between prediction accuracy, flexibility and model interpretability, including the role of overfitting.
   - Determine when a supervised learning model is preferable to unsupervised or semi-supervised learning models.
   - Explain how the appropriateness of regression problems relative to classification problems may be related to whether responses are quantitative or qualitative.

3) Assessing Model Accuracy (Ch. 2.2)
   *For example:*
   - Recognize and explain the equation for mean squared error.
   - 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.
   - Explain the purpose of cross validation.
   - Explain the bias-variance trade-off with an MSE decomposition into three fundamental quantities.
   - Explain the salient features of a simple Bayes classifier (for two classes) including the Bayes decision boundary and Bayes error rate.
   - Explain how the K-nearest neighbors classifier is related to the Bayes classifier and how the choice of K impacts results.
Learning Objectives continued

In Neural Networks and Deep Learning, Determination Press.

Keywords
Perceptron neurons (p. 3)  Feedforward neural networks (p. 17)  
Weights (p. 4)  Recurrent networks (p. 17)  
Threshold value (p. 4)  Cost function (p. 24)  
Layer (p. 6)  Loss function (p. 24)  
Bias (p. 6)  Objective function (p. 24)  
NAND gate (p. 7)  Quadratic cost function (p. 25)  
Input layer (p. 9)  Mean Squared error (MSE) (p. 25)  
Learning algorithms (p. 10)  Gradient descent algorithm (p. 25)  
Sigmoid neuron (p. 11)  Learning rate (p. 30)  
Sigmoid function (p. 12)  Stochastic gradient descent (p. 34)  
Activation function (p. 14)  Mini-batch (p. 34)  
Input neurons (p. 16)  Epoch (p. 35)  
Output neurons (p. 16)  Validation set (p. 37)  
Hidden layer (p. 16)  Hyper-parameters (p. 37)  
Multilayer perceptrons (p. 16)  Deep neural networks (p. 55)

Learning Objectives
Demonstrate proficiency in the following areas:

1) Motivation for using neural nets to recognize handwritten digits
   For example:
   o Explain the use of a training set as an alternative to a rules-based program to recognize digits.

2) Perceptron neurons
   For example:
   o Identify a modern version of a perceptron that is in common use today.
   o Identify an equation that describes how perceptron neurons work in its simplest form.
   o Describe the intuition of a perceptron as a decision-making device.
   o Describe a perceptron as a NAND gate and what it implies for perceptron networks with respect to computing logical functions.
   o Explain how perceptron neurons are more than new types of NAND gates.
Learning Objectives continued

3) Sigmoid neurons
   For example:
   - Recognize a limitation of perceptron neurons that can be overcome by sigmoid neurons.
   - Recognize and differentiate perceptron neurons from sigmoid neurons.
   - Recognize the sigmoid function which is also referred to as a logistic function in its simplistic form as well as when explicitly accounting for inputs, weights and bias.
   - Explain why the smoothness of the sigmoid function is important.

4) The architecture of neural networks
   For example:
   - Identify components of a simple network with appropriate terminology.
   - Describe the central feature of a feed forward network.
   - Compare and contrast feedforward networks with recurrent networks.

5) A simple network to classify handwritten digits
   For example:
   - Argue for a natural order for solving the two problems of segmenting digits and classifying digits.
   - Calculate the required input neurons for classifying an individual digit in an image of a certain size in pixels.
   - Explain the choice to use ten output neurons instead of four for classifying an individual digit.

6) Learning with gradient descent
   For example:
   - Recognize a quadratic cost function of weights and biases and alternative terminology for the cost function.
   - Explain why minimizing a quadratic cost function is preferable to different types of cost functions.
   - Recognize an equation for an update rule that defines the gradient descent algorithm and explain the purpose of each component in the equation.
   - Recognize the equation for that minimizes if we constrain the size of the move so that for some small fixed, and define all notation in the equation.
   - Explain how quickly stochastic gradient descent can speed up learning given a training set size n and a mini-batch size, m.
Learning Objectives continued

7) Implementing a network to classify digits
   For example:
   o Understand the role of hyper-parameters and their impact output for each epoch.

8) Why deep learning matters
   For example:
   o Describe deep learning in terms of neural networks and their performance relative to networks that are not based on deep learning methods.
Learning Objectives continued


Readings


Keywords

<table>
<thead>
<tr>
<th>Information (p. 43)</th>
<th>Linear discriminant (p. 86)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree induction (p. 44)</td>
<td>Support vector machine (p. 92)</td>
</tr>
<tr>
<td>Predictive model (p. 45)</td>
<td>Hinge-loss (p. 94)</td>
</tr>
<tr>
<td>Prediction (p. 45)</td>
<td>Zero-one loss (p. 95)</td>
</tr>
<tr>
<td>Descriptive modeling (p. 46)</td>
<td>Squared error (p. 95)</td>
</tr>
<tr>
<td>Target variable (p. 46)</td>
<td>Odds ratio (p. 98)</td>
</tr>
<tr>
<td>Attributes or features (p. 46)</td>
<td>Log-odds (p. 99)</td>
</tr>
<tr>
<td>Model induction (p. 47)</td>
<td>Logistic function (p. 101)</td>
</tr>
<tr>
<td>Deduction (p. 47)</td>
<td>Generalization (p. 112)</td>
</tr>
<tr>
<td>Training data (p. 47)</td>
<td>Overfitting (p. 113)</td>
</tr>
<tr>
<td>Labeled data (p. 47)</td>
<td>Fitting graph (p. 113)</td>
</tr>
<tr>
<td>Supervised segmentation (p. 48)</td>
<td>Holdout data (p. 113)</td>
</tr>
<tr>
<td>Information gain (p. 51)</td>
<td>Base rate (p. 115)</td>
</tr>
<tr>
<td>Entropy (p. 51)</td>
<td>Sweet spot (p. 117)</td>
</tr>
<tr>
<td>Parent set (p. 52)</td>
<td>Cross-validation (p. 126)</td>
</tr>
<tr>
<td>Child set (p. 52)</td>
<td>Folds (p. 127)</td>
</tr>
<tr>
<td>Variance (p. 56)</td>
<td>Learning curve (p. 131)</td>
</tr>
<tr>
<td>Entropy graph/chart (p. 58)</td>
<td>Sub-training set (p. 134)</td>
</tr>
<tr>
<td>Classification tree (p. 63)</td>
<td>Validation set (p. 134)</td>
</tr>
<tr>
<td>Decision nodes (p. 63)</td>
<td>Nested holdout testing (p. 134)</td>
</tr>
<tr>
<td>Probability estimation tree (p. 64)</td>
<td>Nested cross validation (p. 135)</td>
</tr>
<tr>
<td>Tree induction (p. 64)</td>
<td>Sequential forward selection (p. 135)</td>
</tr>
<tr>
<td>Decision surface or boundary (p. 69)</td>
<td>Sequential backward elimination (p. 135)</td>
</tr>
<tr>
<td>Frequency-based estimation of class membership probability (p. 72)</td>
<td>Independent events (p. 236)</td>
</tr>
<tr>
<td></td>
<td>Joint probability using conditional probability (p. 237)</td>
</tr>
</tbody>
</table>
Learning Objectives continued

Keywords continued
Laplace correction (p. 73)   Bayes’ Rule (p. 237)
Linear classifier (p. 85)   Naive Bayes Classifier (p. 242)
Posterior probability (p. 238)   Generative Model (p. 244)
Prior (p. 238)   Lift (p. 244)
Likelihood (p. 240)   Naive-Naive Bayes (p. 245)
Conditional independence (p. 241)

Learning Objectives

Demonstrate proficiency in the following areas:

1) Models, Induction and Prediction
   
   For example:
   
   o Define information and tree induction.
   o Define prediction in the context of data science.
   o Compare and contrast predictive modeling with descriptive modeling.
   o Define attributes or features.
   o Describe model induction.
   o Compare and contrast induction with deduction.
   o Define training data and labeled data.

2) Supervised Segmentation
   
   For example:
   
   o Describe supervised segmentation.
   o List the complications arising from selecting informative attributes.
   o Define entropy and information gain.
   o Recognize and apply entropy with a set containing two distinct groups.
   o Recognize and apply entropy with maximum and minimum disorder.
   o Contrast parent set with child set.
   o Calculate information gain for a child relative to a parent.
   o Discuss the issues with the numerical variables for supervised segmentation.
   o Define variance and discuss its application to numeric variables for supervised segmentation.
   o Define an entropy graph/chart.
   o Describe how an entropy chart can be used to select informative variable.
   o Define a classification tree, decision nodes, probability estimation tree, and tree induction.
Learning Objectives continued

3) Visualizing Segmentations
   For example:
   - Define decision surface or decision boundaries.
   - Describe how decision trees can be described as a set of rules.
   - Define frequency-based estimation of class membership probability.
   - Describe how Laplace correction is used to modify probability of a leaf node with few members.

4) Classification via Mathematical Functions
   For example:
   - Define a linear classifier.
   - Recognize and apply the equation of a straight line using slope and intercept.
   - Define a linear discriminant.
   - Describe decision boundaries in 2-dimension, 3-dimension, and higher dimensions.
   - Interpret magnitude of a feature’s weight in a general linear model.
   - Describe how linear discriminant functions can be used for scoring and ranking instances.
   - Describe the objective function of the Support Vector Machine (SVM).
   - Describe the important ideas behind the Support Vector Machine (SVM).
   - Define the hinge-loss function and zero-one loss function, and squared error.
   - Describe the reason for not using squared loss function in classification problems.

5) Regression via Mathematical Functions
   - Describe the major drawback of the least square regression.
   - Define odds and log odds.
   - List the important points of the logistic regression.
   - Recognize and apply the logistic function.
   - Describe the shape of the logistic function.
   - Describe how objective function is formed in logistic regression.
   - Compare and contrast classification trees with linear classifiers.

6) Overfitting and Its Avoidance
   - Define generalization, overfitting, fitting graph, holdout data, and base rate.
   - Apply the concept of fitting graph to find the optimal tree induction model.
   - Define sweet spot for a typical fitting graph.
Learning Objectives continued

- Apply the concept of overfitting in mathematical functions.
- Analyze overfitting for logistic regression and support vector machine.
- Explain why overfitting is bad.
- Define cross-validation and folds.
- Define a learning curve.
- Compare and contrast a learning curve with a fitting graph.
- Describe shape of learning curves for logistic regression and tree induction.
- List strategies that can be used to avoid overfitting in tree induction.
- Describe how minimum number of instances in a tree leaf can be used to limit tree size.
- Explain how hypothesis testing can be used to limit tree induction.
- Define sub-training set, validation set, and nested holdout testing.
- Explain nested cross-validation.
- Describe sequential forward selection and sequential backward elimination.

7) Evidence and Probabilities

- Define independent events.
- Recognize and apply joint probability using conditional probability.
- Calculate joint probability for independent and dependent events.
- Explain the Bayes’ Rule with the help of an example.
- Define posterior probability, prior, likelihood, and conditional independence.
- Explain the naïve Bayes classifier.
- Explain why we do not need to calculate the denominator of the Bayes’ rule for naïve Bayes classifier.
- List the advantages and disadvantages of the naïve Bayes classifier.
- Define generative model, lift, and Naïve-Naïve Bayes.
Learning Objectives continued


Keywords

- Residual (p. 62)
- Residual sum of squares (p. 63)
- Population regression line (p. 63)
- Least squares line (p. 63)
- Bias (p. 65)
- Unbiased (p. 65)
- Standard error (p. 65)
- Residual standard error (p. 66)
- Confidence interval (p. 66)
- Null hypothesis (p. 67)
- Alternative hypothesis (p. 67)
- t-statistic (p. 67)
- $R^2$ statistic (p. 70)
- Total sum of squares (p. 70)
- F-statistic (p. 75)
- Forward selection (p. 78)
- Backward selection (p. 79)
- Mixed selection (p. 79)
- Dummy variable (p. 84)
- Additive linear (p. 86)
- Hierarchical principle (p. 89)
- Polynomial regression (p. 90)
- Heteroscedasticity (p. 95)
- Multicollinearity (p. 101)

- Power (p. 101)
- Variance inflation factor (p. 101)
- Best subset selection (p. 205)
- Deviance (p. 206)
- Forward stepwise selection (p. 207)
- Backward stepwise selection (p. 208)
- $C^p$ (p. 211)
- Akaike information criterion (AIC) (p. 211)
- Bayesian information criterion (BIC) (p. 211)
- Adjusted $R^2$ (p. 211)
- Ridge regression (p. 215)
- Tuning parameter (p. 215)
- Shrinkage penalty (p. 215)
- $l_2$ norm (p. 216)
- Scale equivalent (p. 217)
- Sparse (p. 219)
- Dimension reduction (p. 229)
- Linear combination (p. 229)
- Principal component analysis (p. 230)
- Principal component regression (p. 233)
- Partial least squares (p. 237)
- Low-dimensional (p. 238)
- High dimensional (p. 239)
- Curse of dimensionality (p. 242)

Learning Objectives

Demonstrate proficiency in the following areas:

1) Simple Linear Regression

For example:

- Define residual and RSS.
- Recognize and apply the least squares coefficient estimates.
- Interpret the least squares coefficients.
- Define population regression line and least squares line.
Learning Objectives continued

- Define the concept of bias and unbiased estimators.
- Define standard error and residual standard error.
- Calculate standard error of a statistic.
- Calculate the 95% confidence interval.
- Describe null and alternative hypothesis.
- Calculate the t-statistic.
- Describe how accuracy of a linear regression can be assessed.
- Recognize and apply $R^2$ statistic.
- Define total sum of squares.

2) Multiple linear regression

*For example:*

- Interpret the coefficients of a multiple linear regression.
- Describe how relationship between response and predictors is tested in a multiple linear regression.
- Recognize and apply the F-statistic.
- Describe how important variables can be decided in multiple regression.
- Define forward selection, backward selection, and mixed selection.
- Describe the tools used to examine model fit for multiple regression.

3) Considerations in the regression model

*For example:*

- Define dummy variables.
- Describe how qualitative variables with more than two levels can be used in multiple regression.
- Describe additive and linear assumptions for linear regression model.
- Define interaction effect.
- Interpret the coefficients of an interaction term.
- Describe hierarchical principle for multiple regression.
- Define polynomial regression.
- Describe the potential problems, such as non-linearity, correlation of error terms, non-constant variance, outliers, high-leverage points, and collinearility, for linear regression model.
- Define heteroscedasticity.
- Define power of a hypothesis test.
- Define multicollinearity and variance inflation factor.
Learning Objectives continued

4) Subset selection
   For example:
   o Define best subset selection.
   o List the steps used in best subset selection.
   o Define deviance.
   o Describe forward stepwise selection and backward stepwise selection.
   o List the steps used in forward stepwise selection and backward stepwise selection.
   o Recognize and apply the equations for CP, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and adjusted $R^2$.

4) Shrinkage methods
   For example:
   o Define ridge regression, tuning parameter, and shrinkage penalty.
   o Define l2 norm and scale equivalent.
   o Define standardizing the predictors.
   o Describe bias-variance tradeoff.
   o Describe the ridge regression.
   o Describe how ridge regression improves upon least squares.
   o Describe the advantage of Lasso over the ridge regression.
   o Define a sparse model.
   o Describe the variable selection property of the Lasso.
   o Compare the Lasso to the Ridge regression.
   o Describe how to select the tuning parameter.

5) Dimension reduction methods
   For example:
   o Define dimension reduction and linear combination.
   o Describe principal component analysis.
   o Describe principal component regression.
   o Describe partial least squares.
Learning Objectives continued

6) Considerations in high dimensions
   
   For example:
   - Define low-dimensional and high-dimensional data.
   - Describe what goes wrong in high dimensions.
   - Describe regression in high dimensions.
   - Define curse of dimensionality.


<table>
<thead>
<tr>
<th>Keywords</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arithmetic return (p. 3)</td>
<td></td>
</tr>
<tr>
<td>Geometric return (p. 3)</td>
<td></td>
</tr>
<tr>
<td>Time resolution (p. 5)</td>
<td></td>
</tr>
<tr>
<td>Time horizon (p. 5)</td>
<td></td>
</tr>
<tr>
<td>Random walk model (p. 7)</td>
<td></td>
</tr>
<tr>
<td>Autoregressive model (p. 8)</td>
<td></td>
</tr>
<tr>
<td>AR(1) model (p. 8)</td>
<td></td>
</tr>
<tr>
<td>Stationarity (p. 9)</td>
<td></td>
</tr>
<tr>
<td>Autocorrelation function (p. 10)</td>
<td></td>
</tr>
<tr>
<td>GARCH (1,1) (p. 13)</td>
<td></td>
</tr>
</tbody>
</table>

Learning Objectives

Demonstrate proficiency in the following areas:

1) Concepts of time series
   
   For example:
   - Define arithmetic and geometric returns.
   - Recognize and apply the relationship between arithmetic and geometric returns.
   - Describe the shape of the plotted line when geometric returns are plotted against arithmetic returns.
   - Define time resolution and time horizon.
   - Describe how time resolution and time horizon affect the distribution of financial data.

2) Statistical models
   
   For example:
   - Describe a random walk model and an autoregressive model.
   - Recognize and apply an AR(1) model.
   - Recognize and apply the variances of a random walk and an autoregressive model.
   - Define stationarity and autocorrelation function.
   - Recognize and apply the formula for the autocorrelation function.
Learning Objectives continued

3) Modeling volatility

   For example:
   - Describe a GARCH(1,1) model.
   - Recognize and apply the variance equation of a GARCH(1,1) model.
   - Describe the goodness-of-fit for a GARCH model.
Learning Objectives continued

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

Readings


Keywords

<table>
<thead>
<tr>
<th>Euclidean distance (p. 144)</th>
<th>Clustering (p. 164)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest neighbors (p. 144)</td>
<td>Hierarchical clustering (p. 165)</td>
</tr>
<tr>
<td>Combining function (p. 147)</td>
<td>Dendogram (p. 165)</td>
</tr>
<tr>
<td>Weighted voting (p. 150)</td>
<td>Linkage function (p. 167)</td>
</tr>
<tr>
<td>Manhattan distance (p. 159)</td>
<td>Cluster center or centroid (p. 170)</td>
</tr>
<tr>
<td>Cosine distance (p. 160)</td>
<td>k-means clustering (p. 170)</td>
</tr>
<tr>
<td>Edit distance or Levenshtein metric (p. 161)</td>
<td>Distortion (p. 173)</td>
</tr>
</tbody>
</table>

Learning Objectives

Demonstrate proficiency in the following areas:

1) Similarity and distance
   For example:
   - Recognize and apply the formula for calculating the general Euclidian distance.
   - Define nearest neighbors and combining function.
   - Explain how combining function can be used for classification.
   - Define weighted voting or similarity moderated voting.
   - Calculate contributions for weighted voting for classification.
   - Explain how k in k-NN can be used to address overfitting.
   - Discuss issues with nearest-neighbor methods with focus on
     - Intelligibility,
     - Dimensionality and domain knowledge, and
     - Computational efficiency.
Learning Objectives continued

2) Technical details related to similarities and neighbors

For example:

- Recognize and apply the formula for calculating the Manhattan distance and Cosine similarity.
- Define edit distance or Levenshtein metric.
- Define clustering, hierarchical clustering, and dendogram.
- Describe how dendogram can help decide the number of clusters.
- List the advantages of hierarchical clustering.
- Define linkage functions.
- Define “cluster center” or centroid and k-means clustering.
- Compare and contrast k-means clustering with hierarchical clustering.
- Describe the k-means algorithm.
- Describe the reason for running k-means algorithm many times.
- Define a cluster’s distortion.
- Describe the method for selecting k in k-means algorithm.


Keywords

<table>
<thead>
<tr>
<th>Logistic function (p. 132)</th>
<th>Loadings (p. 375)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odds (p. 132)</td>
<td>Bottom-up agglomerative clustering (p. 390)</td>
</tr>
<tr>
<td>Log odds (p. 132)</td>
<td>Linkage (p. 394)</td>
</tr>
<tr>
<td>Likelihood function (p. 133)</td>
<td>Inversion (p. 395)</td>
</tr>
<tr>
<td>Principal component analysis (p. 375)</td>
<td></td>
</tr>
</tbody>
</table>
Learning Objectives continued

Learning Objectives
Demonstrate proficiency in the following areas:

1) Logistic regression
For example:
- Describe the limitations of linear regression for categorical response variable.
- Recognize and apply the logistic function, odds, and log-odds.
- Recognize and apply the likelihood function.
- Interpret the logistic regression coefficients with single regressor.
- Recognize and apply predictions using the logistic regression.
- Interpret the multiple logistic regression coefficients.

2) Principal component analysis
For example:
- Define principal component analysis.
- Describe how principal components are found.
- Define loadings of a principal component.
- Describe another interpretation of the principal component.
- Describe the effect of scaling on the principal component.
- Describe uniqueness of the principal component.
- Define proportion of variance explained and scree plot.
- Describe how number of principal components is decided.

3) Clustering methods
For example:
- Describe the algorithm for k-means clustering.
- Define bottom-up agglomerative clustering.
- Describe how to interpret a dendogram.
- Define linkage as well as the four types of linkage.
- Define inversion.
- Describe the choice of dissimilarity measure for clustering.
- Describe the practical issues with clustering.
Learning Objectives continued


Readings


Keywords

- Accuracy (p. 189)
- Confusion matrix (p. 189)
- False positive (p. 190)
- False negative (p. 190)
- Expected value (p. 194)
- Class prior (p. 201)
- Precision (p. 204)
- Recall (p. 204)
- F-measure (p. 204)
- Profit curve (p. 212)
- ROC graph (p. 215)
- AUC (p. 219)
- Cumulative response curve (p. 219)

Learning Objectives

Demonstrate proficiency in the following areas:

1) Evaluating classifiers

   For example:

   - Define accuracy.
   - Describe a confusion matrix.
   - Define false positives and false negatives.
   - Describe the problems with unbalanced data.
   - Describe the problems with unequal costs and benefits.
Learning Objectives continued

2) A key analytical framework: expected value
   For example:
   o Calculate expected value and expected benefit.
   o Describe how expected value can be used to frame classifier use.
   o Describe how expected value can be used to frame classifier evaluation.
   o Define class priors.
   o Recognize and apply expected profit using priors.
   o Describe the two pitfalls common to formulating cost-benefit analysis.
   o Define precision and recall.
   o Define F-measure.

3) Visualizing model performance
   For example:
   o Describe a ranking classifier.
   o Define a profit curve.
   o Describe the ROC graph.
   o Describe the four corners and the diagonal of the ROC graph.
   o Describe how the ROC space can be used to evaluate classifiers.
   o Define AUC.
   o Describe a cumulative response curve.


Keywords
Exaggerated positive (p. 68)
**Learning Objectives continued**

**Learning Objectives**
Demonstrate proficiency in the following areas:

1) **Backtesting Protocol in the Era of Machine Learning**  
   *For example:*  
   - List the five attractive features of the simulated strategy.  
   - Describe the lessons learned from the data-mined strategy.  
   - Describe the winner’s curse.  
   - Define exaggerated positive.  
   - Describe the six protocols suggested for avoiding false positives.


**Keywords**
- Specificity *(p. 2)*  
- Power *(p. 4)*  
- Sensitivity *(p. 2)*

**Learning Objectives**
Demonstrate proficiency in the following areas:

1) **An investigation of the false discovery rate and the misinterpretation of p-values**  
   *For example:*  
   - Define specificity and sensitivity.  
   - Describe the false discovery rate with the help of a tree diagram.  
   - Define power of a test.  
   - Describe the false discovery rate in simulated t-tests.  
   - Describe underpowered study.  
   - Describe the inflation effect in the context of false discovery.  
   - Describe what happens when we consider p=0.05 rather than p<=0.05.  
   - Describe Berger’s approach.
Learning Objectives continued


**Keywords**

*Selection bias under multiple testing (p. 99)*

Learning Objectives

Demonstrate proficiency in the following areas:

1) **A Data Science Solution to the Multiple-Testing Crisis**

   *For example:*

   - Define selection bias under multiple testing.
   - Describe the three properties satisfied by the heatmap of the correlation matrix of the returns of all trials.
   - Describe clustering of trials.
   - Describe cluster statistics and how it can reduce the probability of selecting a false positive.
   - Describe the implications for authors, journals, and financial firms.
Learning Objectives continued

Topic 6. Data Mining & Machine Learning: Representing & Mining Text


Keywords

<table>
<thead>
<tr>
<th>Linguistic structure (p. 252)</th>
<th>Term frequency (p. 254)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dirty data (p. 252)</td>
<td>Stemmed (p.255)</td>
</tr>
<tr>
<td>Document (p. 253)</td>
<td>Stopwords (p. 25)</td>
</tr>
<tr>
<td>Token (p. 253)</td>
<td>Inverse document frequency (p. 256)</td>
</tr>
<tr>
<td>Terms (p. 253)</td>
<td>n-grams (p. 265)</td>
</tr>
<tr>
<td>Corpus (p. 253)</td>
<td>Latent information model (p. 268)</td>
</tr>
<tr>
<td>Bag of words (p. 254)</td>
<td>Information triage (p. 276)</td>
</tr>
</tbody>
</table>

Learning Objectives

Demonstrate proficiency in the following areas:

1) Broad issues involved in mining text
   For example:
   - Explain why text is “dirty” which makes mining text is difficult.

2) Text representation
   For example:
   - Understand the meaning of terms when used in the field of information retrieval.
   - Describe the “bag of words” approach including the following steps:
     - Measuring term frequency (TF)
     - Measuring sparseness: inverse document frequency (IDF)
     - Combining them: TFIDF
   - Apply appropriate methods to search an example set of documents.
   - Express entropy in terms of the IDF.
Learning Objectives continued

3) Additional text representation approaches beyond “bag of words”
   
   For example:
   
   o Describe N-gram sequences.
   o Describe named entity extraction.
   o Describe topic models.

4) Mining news stories to predict stock price movement
   
   For example:
   
   o 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.
   o Describe required considerations for data preprocessing.
   o Choose and discuss appropriate methods for analyzing the results.


Keywords

<table>
<thead>
<tr>
<th>Sentiment analysis (p. 1)</th>
<th>Gold labels (p.11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probabilistic classifier (p. 2)</td>
<td>Precision (p.12)</td>
</tr>
<tr>
<td>Generative classifier (p. 2)</td>
<td>Recall (p.12)</td>
</tr>
<tr>
<td>Discriminative classifier (p. 2)</td>
<td>F-measure (p.13)</td>
</tr>
<tr>
<td>Linear classifier (p. 5)</td>
<td>Macroaveraging (p.13)</td>
</tr>
<tr>
<td>Sentiment lexicon (p. 9)</td>
<td>Microaverageing (p.13)</td>
</tr>
</tbody>
</table>

Learning Objectives

Demonstrate proficiency in the following areas:

1) Classification
   
   For example:
   
   o Describe common applications of classifying text.
   o Describe tasks often involved in classifying text.
   o Compare alternative methods of classification.
Learning Objectives continued

2) Math behind Naïve Bayes classifiers
   For example:
   o Explain why in the context of classifying a document the denominator can be dropped from Bayes Rule.
   o Explain the bag of words and naïve Bayes assumptions.
   o Explain why Naïve Bayes calculations are done in log space so that the predicted class is a linear function of input features.

3) Training the Naïve Bayes classifiers
   For example:
   o Explain why Laplace smoothing is commonly used in Bayes text categorization.
   o Explain how stop words and unknown words are treated during training.
   o Calculate the prior probabilities of two classes given a training set categorized into two classes.
   o Determine the class that a test sentence belongs to using the Naïve Bayes classifier.

4) Optimizing for sentiment analysis
   For example:
   o Explain how binary multinomial Naïve Bayes differs from Naïve Bayes.
   o Explain why binary multinomial Naïve Bayes (also called binary NB) might improve results.
   o Describe two other methods (besides binary NB) that can improve the results of sentiment analysis.

5) Evaluation of sentiment analysis results
   For example:
   o Calculate precision and recall statistics given system output and gold standard label results.
   o Describe the F-measure and various methods of weighting precision and recall.
   o Compare macroaveraging and microaveraging approaches to evaluating the categorization performance of multiple classes.
   o Compare 10-fold cross-validation with bootstrap tests.
Learning Objectives continued


<table>
<thead>
<tr>
<th>Key Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data trust deficit (p. 2)</td>
</tr>
<tr>
<td>Veracity (p. 6)</td>
</tr>
</tbody>
</table>

**Learning Objectives**

Demonstrate proficiency in the following areas:

1) **Big data for business**
   
   *For example:*
   
   - Discuss the potential and concerns of big data for business.
   - Explain how the new term “data trust deficit” developed.

2) **Ethical issues**
   
   *For example:*
   
   - List five methods of protecting human rights in the ‘Era of Big Data.’
   - Provide an example of a concern for each of three main areas of privacy issues: customer profiling, group privacy and data security.
   - Discuss what constitutes informed consent.
   - Provide an example of how to improve the veracity of data.

3) **The Ethics Test**
   
   *For example:*
   
   - List six questions which Ethics Professionals within an organization using big data can ask themselves.
Learning Objectives continued


<table>
<thead>
<tr>
<th>Keywords</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial intelligence (p. 1)</td>
<td>Code of ethics (p. 6)</td>
</tr>
</tbody>
</table>

Learning Objectives

Demonstrate proficiency in the following areas:

1) **The nature of and business risks of artificial intelligence (AI)**  
   *For example:*
   - List three main features characterizing artificial intelligence.
   - List three immediate risks of artificial intelligence.

2) **Values that form the cornerstone of an ethical framework of artificial intelligence in business**  
   *For example:*
   - Discuss each of the following as they impact the ethical nature of applications of artificial intelligence in business:
     - Accurate results
     - Respect of privacy
     - Transparency and openness
     - Interpretability of algorithms
     - Fairness to stakeholders
     - Integrity and due diligence
     - Control of humans relative to machines
     - Impact of a new technology
     - Accountability assignment
     - Learning about how the AI technologies work

3) **The role of business decision makers**  
   *For example:*
   - List five measures organizations can take to minimize the risk of ethical lapses due to improper use of AI technologies.
   - List some questions addressing the use of AI that could be included in a code of ethics.
Learning Objectives continued


Keywords
General Data Protection Regulation (p. 1) People risk (p. 3)

Learning Objectives

Demonstrate proficiency in the following areas:

1) General Data Protection Regulation (GDPR)
   For example:
   - Describe the primary purpose of the GDPR.
   - Describe the key changes in data protection regulation including the meaning of
     - rights of the individual
     - informed consent
     - notification
     - data portability
     - supervision and enforcement, and
     - liability

2) Separating ethics and compliance
   For example:
   - Distinguish between two types of threats of personal data breaches.
   - Discuss ‘people risk.’
   - List key questions around the role an ethical culture plays in preventing data breaches.

3) Maintaining privacy of personal data
   For example:
   - Describe how an organization must build awareness regarding employees’ roles in
     protecting data.
   - Discuss liability if the 72-hour notification deadline is missed.
Learning Objectives continued

4) The GDPR Embedding Wheel

*For example:*

- Describe how the tone from the top can help foster an ethical culture and compliance with the GDPR.
- Describe how establishing the boundaries and standards can help foster an ethical culture and compliance with the GDPR.
- Describe how communication and training can help foster an ethical culture and compliance with the GDPR.
- Describe how choice of the individual can help foster or hinder an ethical culture and compliance with the GDPR.
- Describe how monitoring outcomes can help foster an ethical culture and compliance with the GDPR.
Learning Objectives continued


Readings


   
   The paper posted on the FDP Institute’s website will be the version used for exam questions.


Learning Objectives continued


<table>
<thead>
<tr>
<th>Keywords</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Big data (p. 4)</td>
<td>Fraud detection (p. 11)</td>
</tr>
<tr>
<td>Artificial intelligence (p. 4)</td>
<td>RegTech (p. 11)</td>
</tr>
<tr>
<td>Natural language processing (p. 5)</td>
<td>InsurTech (p. 13)</td>
</tr>
<tr>
<td>Machine learning (ML) (p. 4)</td>
<td>Chatbots (p. 14)</td>
</tr>
<tr>
<td>Supervised learning (p. 5)</td>
<td>Know your customer (KYC) (p. 20)</td>
</tr>
<tr>
<td>Unsupervised learning (p. 5) SupTech (p. 21)</td>
<td>Reinforcement learning (p. 5) Audibility (p.33)</td>
</tr>
<tr>
<td>Deep learning (p. 5)</td>
<td>Fintech (p. 35)</td>
</tr>
<tr>
<td>Sentiment indicators (p. 10)</td>
<td>Rob-advisors (p.35)</td>
</tr>
<tr>
<td>Trading signals (p. 11)</td>
<td>Tonality analysis (p.36)</td>
</tr>
</tbody>
</table>

Learning Objectives

Demonstrate proficiency in the following areas:

1) Categories of machine learning algorithms

*For example:*

- Define four categories of machine learning algorithms based on the degree of human intervention.
- Describe the role of machine learning algorithms in determining causality vs correlation.
- Define 'augmented intelligence'.

2) Drivers of the growth in use of fintech and adaptation of artificial intelligence

*For example:*

- Discuss the supply factors related to advances in computing technologies and changes in the financial sector.
- Discuss the supply factors related to search for higher profits, increased competition and changes in the regulatory environment.

3) Use cases of artificial intelligence and machine learning in financial sector

*For example:*

- Describe customer-focused uses, such as credit scoring, insurance and client-facing chatbots.
- Describe operations-focused uses, such as optimal allocation of capital, risk management modeling, market impact analysis.
Learning Objectives

o Describe portfolio management and trading uses.

o Describe regulatory compliance and supervision uses by financial institutions, central banks macroprudential authorities, and market regulators.

4) The micro effects of uses of artificial intelligence and machine learning in the financial sector.

For example:

o Describe the uses of artificial intelligence and machine learning in information gathering and processing their potential impacts on financial markets.

o Describe the uses of artificial intelligence and machine learning in improving efficiency of financial institutions.

o Describe the uses of artificial intelligence and machine learning by financial institutions and their potential impacts on customers and investors.

5) The macro effects of uses of artificial intelligence and machine learning in the financial sector

For example:

o Describe economic growth and enhanced economic efficiency that could result from the applications of artificial intelligence and machine learning to financial services.

o Describe the implications of uses of artificial intelligence and machine learning by financial institutions for market concentration and systemic importance of those institutions.

o Describe how the uses of artificial intelligence and machine learning by financial institutions could be sources of greater stability and vulnerability in financial markets.

o Describe how the uses of artificial intelligence and machine learning by insurance industry could affect both moral hazard and adverse selection problems.

o Describe challenges posed by the lack of interpretability or auditability in applications of artificial intelligence and machine learning in the financial industry.

6) The terms listed in the glossary

For example:

o Describe the following terms: Algorithm, Artificial intelligence, Augmented intelligence, Big data, Chatbots, Cluster analysis, Deep learning, FinTech, InsurTech, Internet of things, Machine learning, Natural Language Processing, RegTech, Reinforcement learning, Robo-advisors, Social trading SupTech, Supervised learning, Tonality analysis, Topic modelling, and Unsupervised learning.
Learning Objectives continued


**Keywords**
- Fintech (p. 79)
- Robo-advisor (p. 80)
- Work-flow (p. 83)
- D2C platforms (p. 86)
- Hybrid (p. 86)
- B2B platforms (p. 86)

Learning Objectives

Demonstrate proficiency in the following areas:

1) **Robo-advisors, their characteristics and services offered by them**
   
   *For example:*
   - List the three areas of passive strategies.
   - Compare and contrast North American and Asian robo-advisors.
   - Describe the generic workflow employed by robo-advisors to provide financial advice.
   - List the key characteristics of robo-advisors.
   - List the three main areas of private wealth management.
   - Describe areas of private wealth management that robo-advisors could be more effective.
     - Investment advisory and asset management
     - Retirement planning
     - Estate planning
     - Tax planning


**Keywords**
- Alternative data (p. 14)
- Social media (p. 14)
- Microdata (p. 14)
- Data exhaust (p. 14)
- Rivalry (p. 16)
- Excludability (p. 16)
- Defensive strategies (p. 17)
- Defensible strategies (p. 18)
- Operational alpha (p. 19)
- Aggregation (p. 19)
- Disaggregation (p. 19)
- Volume (p. 21)
- Velocity (p. 21)
- Variety (p. 21)
- Veracity (p. 21)
- Granularity (p. 21)
- Relationality (p. 21)
- Flexibility (p. 21)
- Actionability (p. 22)
- Excludable (p. 28)
- Data hoarding (p. 29)
Learning Objectives continued

Learning Objectives

Demonstrate proficiency in the following areas:

1) Alternative data and institutional investors

For example:

- List the most commonly used types of alternative data.
- Discuss advantages and disadvantages that institutional investors may have in using alternative data.
- Discuss why the deepest value proposition alternative data has for institutional investors entails defensive and defensible strategies.
- Discuss the applications of alternative data to risk measurement and management for institutional investors.
- Describe the operational alpha gains by institutional investors through the use of alternative datasets.
- Describe types of alternative datasets in terms of the origins of dataset.
- Discuss why the volume, veracity and velocity of big data may not determine value of alternative data for institutional investors.
- Describe the six-dimensional characterization of alternative data.
- Discuss external asset managers and alternative data providers as methods of accessing alternative data.
- Describe rivalry and excludability as determinants of alternative dataset’s value.
- Discuss the consequences of the increased use of alternative data on risk for institutional investors.


Keywords

Factors (p. 32)  CART (p. 34)
Linear (p. 32)  Binary recursive partitioning (p. 34)
Nonlinear (p. 32)  Bagging (p. 34)
Random forest (p. 33)  Out-of-bag data (p. 34)
Supervised (p. 34)  Feature importance (p. 35)
Unsupervised (p. 34)  Mean decrease accuracy (p. 34)
Root node (p. 34)  Fama-French-Carhart (p. 37)
Decision node (p. 34)  Probabilistic Sharpe ratio (p. 42)
Terminal node (p. 34)
Learning Objectives *continued*

Learning Objectives

Demonstrate proficiency in the following areas:

1) Applications of random forest regression algorithm to factor models

*For example:*

- Discuss two shortcomings of nonlinear factor models that are developed to address shortcomings of linear models.
- Discuss the ability of random forest algorithm to overcome one shortcoming of linear models.
- Discuss the ability of random forest algorithm to overcome one shortcoming of nonlinear models.
- List 4 components of the decision tree when applied to the regression problem of factor models.
- 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.
- Describe the role of out-of-bag observations in a random forest algorithm.
- Discuss mean decrease accuracy approach to estimating feature importance in a random forest algorithm.
- Recognize and apply the probabilistic Sharpe ratio.


<table>
<thead>
<tr>
<th>Keywords</th>
<th>Forecast horizon (p. 10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine learning (p. 3)</td>
<td>Bagging (p. 11)</td>
</tr>
<tr>
<td>Overfitting (p. 5)</td>
<td>Boosting (p. 11)</td>
</tr>
<tr>
<td>Signal-to-noise ratio (p. 4)</td>
<td>Fundamental factors (p. 13)</td>
</tr>
<tr>
<td>Ensemble algorithms (p. 6)</td>
<td>Technical factors (p. 11)</td>
</tr>
<tr>
<td>Feature engineering (p. 8)</td>
<td></td>
</tr>
</tbody>
</table>

Learning Objectives

Demonstrate proficiency in the following areas:

1) The applications of machine learning algorithms to stock selection

*For example:*

- Describe the role of signal to noise ratio in creating overfitted models.
- 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.
Learning Objectives continued

- Describe the four different approaches to bagging and boosting employed by the paper to avoid overfitting.
- Explain the importance of feature engineering in mitigating the overfitting problem.
- Describe the three decisions that must be made with regard to the forecasting goals of the machine learning algorithms.
- Describe the bias versus variance tradeoff.
- Explain the role of bagging and boosting in affecting the bias versus variance tradeoff.


<table>
<thead>
<tr>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Learning (p. 2)</td>
</tr>
<tr>
<td>Regularization (p. 2)</td>
</tr>
<tr>
<td>Mean squared error (p. 9)</td>
</tr>
<tr>
<td>Ordinary least squares (p. 9)</td>
</tr>
<tr>
<td>Heavy tails (p. 11)</td>
</tr>
<tr>
<td>Huber loss function (p. 12)</td>
</tr>
<tr>
<td>Penalized linear models (p. 11)</td>
</tr>
<tr>
<td>Loss function (p. 11)</td>
</tr>
<tr>
<td>Penalty function (p. 11)</td>
</tr>
<tr>
<td>Net elastic (p. 11)</td>
</tr>
<tr>
<td>Hyperparameters (p. 12)</td>
</tr>
<tr>
<td>Tuning parameters (p. 12)</td>
</tr>
<tr>
<td>LASSO (p. 13)</td>
</tr>
<tr>
<td>Principal components (p. 13)</td>
</tr>
<tr>
<td>Partial least squares (p. 13)</td>
</tr>
<tr>
<td>Approximation error (p. 15)</td>
</tr>
<tr>
<td>Estimation error (p. 15)</td>
</tr>
<tr>
<td>Intrinsic error (p. 15)</td>
</tr>
<tr>
<td>Terminal nodes (p. 16)</td>
</tr>
<tr>
<td>Impurity (p. 16)</td>
</tr>
<tr>
<td>Weak learners (p. 17)</td>
</tr>
<tr>
<td>Boosting (p. 18)</td>
</tr>
<tr>
<td>Random forest (p. 18)</td>
</tr>
<tr>
<td>Gradient boosted regression trees (p. 18)</td>
</tr>
<tr>
<td>Neural network (p. 19)</td>
</tr>
<tr>
<td>Hidden layers (p. 19)</td>
</tr>
<tr>
<td>Feed-forward networks (p. 19)</td>
</tr>
<tr>
<td>Input layer (p. 20)</td>
</tr>
<tr>
<td>Output layer (p. 20)</td>
</tr>
<tr>
<td>ReLU function (p. 21)</td>
</tr>
<tr>
<td>Stochastic gradient descent (p. 21)</td>
</tr>
<tr>
<td>Early stopping (p. 22)</td>
</tr>
<tr>
<td>Sharpe ratio (p. 36)</td>
</tr>
</tbody>
</table>

Learning Objectives

Demonstrate proficiency in the following areas:

1) Applications of machine learning algorithms to empirical asset pricing

*For example:*

- Describe the three components of the definition of machine learning.
- Describe the three aspects of empirical asset pricing model that makes it attractive for the applications of machine learning algorithms.
Learning Objectives continued

- Compare and contrast the overall performance of linear versus nonlinear models in predicting individual stock returns and portfolio returns.
- Explain one potential short coming of machine learning algorithms when used to predict asset returns.
- Describe the roles of “training” set, “validation” set and “testing” set in using machine learning algorithms in to predict stock returns.
- Recognize the Huber loss function.
- Describe the benefit of using the Huber loss function as opposed to standard least squares method to the estimation of linear models.
- Recognize the “elastic net” approach for modeling penalized linear models.
- Compare and contrast “elastic net” penalty versus LASSO and Ridge Regression.
- Compare and contrast the principle components regression versus partial regression.
- Describe the boosting regularization method in the context of regression trees.
- Describe the random forest regularization method in the context of regression trees.
- Describe the dropout method in the context of random forest regression trees.
- Recognize rectified linear unit (ReLU) activation function in the context of neural networks.
- Describe a method for discovering variable importance when machine learning algorithms are employed to predict stock returns.


Keywords

<table>
<thead>
<tr>
<th>Backtesting (p. 122)</th>
<th>Precision (p. 128)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume clock (p. 123)</td>
<td>Recall (p. 128)</td>
</tr>
<tr>
<td>Dollar bars (p. 123)</td>
<td>F1-score (p. 128)</td>
</tr>
<tr>
<td>Stationary (p. 123)</td>
<td>Walk-forward approach (p. 129)</td>
</tr>
<tr>
<td>Integer differentiation (p. 123)</td>
<td>Leakage (p. 129)</td>
</tr>
<tr>
<td>Fractional differentiation (p. 124)</td>
<td>Probabilistic Sharpe ratio (p. 132)</td>
</tr>
<tr>
<td>Triple barrier method (p. 127)</td>
<td>Deflated Sharpe ratio (p. 132)</td>
</tr>
</tbody>
</table>
Learning Objectives continued

Learning Objectives
Demonstrate proficiency in the following areas:

1) The most common errors made when machine learning techniques are applied to financial data sets
   
   For example:
   
   - Compare and contrast the silo approach in discretionary strategies versus meta-strategy in machine learning strategies.
   - Compare and contrast repeated backtesting using machine learning versus examining feature importance of the results from a machine learning application.
   - Describe the two problems with data samples generated using time bars.
   - Describe the advantages of dollar bars over time bars in creating data for machine learning algorithms.
   - Describe the benefit of using fractional differentiation in generating stationary series and preserving memory.
   - Explain the triple-barrier method for labeling observed returns.
   - Describe the definitions of precision, recall and F1-score as features of machine learning algorithms.
   - Explain the role of non-independent identically distributed in returns in the failure of k-fold cross-validation in finance.
   - Describe walk forward (WF) approach to backtesting of trading strategies.
   - Describe advantages and disadvantages of walk forward approach.
   - Explain the relationship between the maximum Sharpe ratio obtained from several backtested strategies and the return volatility of those strategies.
   - Describe the concept of probabilistic Sharpe ratio.
   - List the impacts of nonnormalized Sharpe ratio, tail record, skewness and kurtosis on probabilistic Sharpe ratio.


Keywords

- T-statistics (p. 110)
- Family-wise error rate (p. 111)
- False discovery rate (p. 111)
- Bonferroni test (p. 112)
- Holm test (p. 112)
- BHY hurdle (p. 112)
- Type I error (p. 113)
- Type II error (p. 113)
- p-value (p. 114)
Learning Objectives continued

Learning Objectives

Demonstrate proficiency in the following areas:

1) Using statistical techniques to evaluate trading strategies in the presence of multiple tests
   For example:
   - Describe why standard statistical tools, such as p-values and t-statistics can lead to false discoveries in the presence of multiple tests.
   - Calculate the t-statistic based on reported Sharpe ratio for testing a single trading strategy.
   - Describe the Bonferroni tests in the context of the family-wise error rate (FWER) approach to adjusting p-values for multiple tests.
   - Describe the Holm method in the context of the family-wise error rate (FWER) approach to adjusting p-values for multiple tests.
   - Recognize and apply the Holm function to calculate adjusted p-values.
   - Describe the false discovery approach to adjusting p-values in the presence of multiple tests.
   - Recognize and apply the BHY formula to calculate adjusted p-values.
   - Explain the relationship between avoiding false discoveries and missing on profitable opportunities.


Keywords

<table>
<thead>
<tr>
<th>Mainstream (p. 336)</th>
<th>Part of Speech Tagging (p. 349)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary source (p. 336)</td>
<td>Stemming (p. 350)</td>
</tr>
<tr>
<td>Social media (p. 337)</td>
<td>Lemmatization (p. 350)</td>
</tr>
<tr>
<td>Sentiment analysis (p. 339)</td>
<td>Naïve Bayes (p. 355)</td>
</tr>
<tr>
<td>Natural language processing (p. 347)</td>
<td>FNN (p. 363)</td>
</tr>
<tr>
<td>Tokenization (p. 348)</td>
<td>RNN (p. 363)</td>
</tr>
<tr>
<td>Word filter (p. 348)</td>
<td>CNN (p. 363)</td>
</tr>
</tbody>
</table>
Learning Objectives continued

Learning Objectives

Demonstrate proficiency in the following areas:

1) Natural language processing of financial news

   For example:
   - Describe the three categories of sources of news data.
   - Explain the advantages and disadvantages of using the new category of social media.
   - Describe sentiment analysis.
   - Describe the word list approach to sentiment analysis.
   - Describe the three challenges associated with sentiment analysis.
   - Describe the four steps — pre-processing, feature representation, inference and evaluation — in applying NLP to texts.


Keywords

<table>
<thead>
<tr>
<th>Natural language processing (p.8)</th>
<th>Episode (p. 11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RegTech (p. 8)</td>
<td>Topic Net sentiment (p. 16)</td>
</tr>
<tr>
<td>Fintech (p. 11)</td>
<td>Polarity (p. 16)</td>
</tr>
<tr>
<td>Fabula (p. 11)</td>
<td>Disagreement (p. 16)</td>
</tr>
<tr>
<td>Syuzhet (p. 11)</td>
<td>Term document matrix (TDM) (p. 29)</td>
</tr>
<tr>
<td>Topic modeling (p. 11)</td>
<td>Document term matrix (DTM) (p. 29)</td>
</tr>
</tbody>
</table>

Learning Objectives

Demonstrate proficiency in the following areas:

1) Using linguistic analysis to perform risk analysis of investments.

   For example:
   - Explain the difficulties associated with manual parsing of unstructured text.
   - Describe how content and structure of emails could be used for risk analysis.
   - Explain the effectiveness of textual versus hard numbers in corporate risk analysis.
   - Define RegTech.
   - Apply the net sentiment metric to calculate the polarity of a text using its Pos and Neg figures.
   - Calculate the disagreement measure of a text using Pos and Neg figures of a text.
   - Describe the findings of the paper regarding the effectiveness of email length as predictor of risk analysis of Enron.
Action Words

In each of the above learning objectives, action words are used to direct your study focus. Below is a list of all action words used in this study guide, along with definitions and two examples of usage, in a question example and in a description. Should you not understand what is required for any learning objective, we suggest you refer to the table below for clarification.

**NOTE: The question examples in this table are NOT sample questions for the current exam.**

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
<th>Question Example</th>
<th>Example of Term Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analyze</td>
<td>Study the interrelations</td>
<td>George has identified an opportunity for a convertible arbitrage reverse hedge. What risks are associated with this hedge?</td>
<td>You have to <strong>analyze</strong> the positions and factors impacting them. <strong>Correct Answer: B</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apply</td>
<td>Make use of</td>
<td>Alicia Weeks, CFA, Real Estate Investment Advisor, works in an Asian country where there are no securities laws or regulations. According to CFA Institute Standard I, Fundamental Responsibilities, Alicia:</td>
<td>You have to <strong>apply</strong> CFA Institute Standard I to find the correct answer. <strong>Correct Answer: C</strong></td>
</tr>
</tbody>
</table>
|             | **Note:** If you are asked to apply a model to data, you will be expected to have the appropriate equation memorized, unless the question also contains the action word “recognize”. | A. Must adhere to the standards as defined in a neighboring country that has the strictest laws and regulations.  
B. Need not concern herself with ethics codes and standards.  
C. Must adhere to the CFA Institute’s codes and standards.  
D. Must adhere to the standards as defined in a neighboring country that has the least strict laws and regulations. |                                                                                  |
| Argue       | Prove by reason or by presenting the associated pros and cons; debate | Why did the shape of the supply curve for venture capital funds change after 1979? | You have to describe how the curve has changed AND **argue** why it changed by providing reasons and supporting the reasons with statements of facts (e.g., change in regulations). |
### Action Words continued

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
<th>Question Example</th>
<th>Example of Term Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assess</td>
<td>Determine importance, size, or value</td>
<td>How are lower capital gains taxes expected to impact firm commitments?</td>
<td>You must <strong>assess</strong> the significance of the change in the tax rate for firm commitments.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A. Through increased supply of capital, firm commitments are expected to rise.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>B. Through decreased supply of capital, firm commitments are expected to rise.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C. Through decreased after-tax return on venture investments, firm commitments are expected to rise.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D. Through increased after-tax return on venture investments, firm commitments are expected to decline.</td>
<td></td>
</tr>
<tr>
<td>Calculate</td>
<td>Determine a value mathematically</td>
<td>Consider a set of 100 people. Eighty percent have feature A and twenty percent do not have feature A. What is the entropy for this set?</td>
<td>You have to <strong>calculate</strong> entropy based on the given probabilities.</td>
</tr>
<tr>
<td></td>
<td><strong>Note:</strong> You will be expected to have the appropriate equation memorized, unless the question also contains the action word “recognize”.</td>
<td>A. 0.72</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>B. 0.88</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C. 0.93</td>
<td></td>
</tr>
<tr>
<td>Compare</td>
<td>Describe similarities and differences</td>
<td>Which of the following least accurately compares the Sharpe and Treynor ratios?</td>
<td>You must <strong>compare</strong> the ratios based on their most important similarities and their most important differences.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D. Both ratios contain excess return in the numerator.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>E. Both ratios express a measure of return per unit of some measure of risk.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F. The Sharpe ratio is based on total risk, while the Treynor ratio is based on systematic risk.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>G. The Sharpe ratio is the inverse of the Treynor ratio.</td>
<td></td>
</tr>
</tbody>
</table>
### Action Words continued

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
<th>Question Example</th>
<th>Example of Term Use</th>
</tr>
</thead>
</table>
| **Compare and Contrast** | Examine in order to note similarities or differences | A comparison of monthly payments and loan balances of a constant payment mortgage with a constant amortization mortgage with the same loan terms will show that:  
A. The initial payment will be the same.  
B. The payments of the constant payment mortgage are initially greater than those of the constant amortization mortgage, but at some point the payments of the constant payment mortgage become less.  
C. The present value of the payment streams of the two loan types are the same.  
D. The constant payment mortgage loan balance exceeds that of the constant amortization mortgage during the first six months of the loan. | You must compare indices to arrive at the answer.  
**Correct Answer: C**                                                                                     |
| **Construct**       | Make or form by combining or arranging parts or elements | A reverse convertible arbitrage hedge consists of a:  
A. Short convertible position plus a put option on the stock.  
B. Long convertible position plus a put option on the stock.  
C. Short convertible position plus a call option on the stock.  
D. Short convertible position plus a long position in the stock. | You must combine positions to construct the hedge.  
**Correct Answer: D**                                                                                     |
| **Contrast**        | Expound on the differences                       | Which of the following best characterizes a difference between value at risk (VaR) and modified VaR?  
A. Modified VaR is expressed as a percent while VaR is a dollar value.  
B. Modified VaR uses a user defined confidence interval while VaR uses a 99% interval.  
C. Modified VaR incorporates non-normality while traditional VaR assumes normality.  
D. Modified VaR is for a single trading period while traditional VaR is multiple period. | You have to contrast the assumptions of the first model to those of the second model so that the differences are clear.  
**Correct Answer: C**                                                                                     |
| **Define**          | State the precise meaning                        | The interest rate charged by banks with excess reserves at a Federal Reserve Bank to banks needing overnight loans to meet reserve requirements is called the:  
A. Prime rate.  
B. Discount rate.  
C. Federal funds rate.  
D. Call money rate. | You must define, in this case, the federal funds rate.  
**Correct Answer: C**                                                                                     |
### Action Words continued

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
<th>Question Example</th>
<th>Example of Term Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Describe</td>
<td>Convey or characterize an idea</td>
<td>Which of the following words best describes expected return?</td>
<td>You need to choose the word that best describes the concept from a list.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A. Spread</td>
<td>Correct Answer: B</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B. Average</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C. Spread squared</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D. Average squared</td>
<td></td>
</tr>
<tr>
<td>Differentiate</td>
<td>Constitute the distinction between; distinguish</td>
<td>What type of convertible hedge entails shorting a convertible and going long in the underlying stock?</td>
<td>You must differentiate one type of hedge from another.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A. Call-option hedge</td>
<td>Correct Answer: D</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B. Traditional convergence hedge</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C. Implied volatility convergence hedge</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D. Reverse hedge</td>
<td></td>
</tr>
<tr>
<td>Discuss</td>
<td>Examine or consider a subject</td>
<td>Discuss the limitations of private equity data.</td>
<td>You must present a discussion of a set of ideas in a list or paragraph.</td>
</tr>
<tr>
<td>Explain</td>
<td>Illustrate the meaning</td>
<td>1. Explain why return on assets (ROA) rather than return on equity (ROE) might be the preferred measure of performance in the case of hedge funds. or 2. Which of the following best explains risk from the standpoint of investment?</td>
<td>1. You must place a series of thoughts together as an explanation of a term or issue.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A. Investors will lose money.</td>
<td>2. You need to identify the term that best explains a term or issue.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B. Terminal wealth will be less than initial wealth.</td>
<td>Correct Answer: D</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C. Final wealth will be greater than initial wealth.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D. More than one outcome is possible.</td>
<td></td>
</tr>
<tr>
<td>Identify</td>
<td>Establish the identity</td>
<td>The investments that have historically performed best during periods of recession are:</td>
<td>You must identify the term that best meets the criterion of the question</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A. Commodities.</td>
<td>Correct Answer: B</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B. Treasury bills.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C. Stocks and bonds.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D. Gold.</td>
<td></td>
</tr>
<tr>
<td>Illustrate</td>
<td>Clarify through examples or comparisons</td>
<td>For two types of convergence hedges, what situations present profitable opportunities, how are the hedges set up, and what are the associated risks?</td>
<td>You must provide an example for each hedge or compare the two to illustrate how they work.</td>
</tr>
</tbody>
</table>
### Action Words continued

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
<th>Question Example</th>
<th>Example of Term Use</th>
</tr>
</thead>
</table>
| **Interpret** | Explain the meaning | Your certificate of deposit will mature in one week, and you are considering how to invest the proceeds. If you invest in a 30-day CD, the bank will pay you 4% interest. If you invest in a 2-year CD, the bank will pay you 6% interest.

You should choose the:

A. 30-day CD, no matter what you expect interest rates to do in the future.

B. 2-year CD, no matter what you expect interest rates to do in the future.

C. 30-day CD if you expect that interest rates will fall in the future.

D. 2-year CD if you expect that interest rates will fall in the future. | You must **interpret** the features of an investment scenario. **Correct Answer: D** |

| List | Create a series of items | List the determinants of real interest rates. | You must **differentiate** from a list those items that are consistent with the question. |

| Outline | Summarize tersely | Which of the following best characterizes the steps in computing a geometric mean return based on a series of periodic returns from T time periods?

A. Add one to each return, add them together, divide by T and subtract one.

B. Add one to each return, multiply them together, divide by T and subtract one.

C. Add one to each return, add them together, take the T<sup>th</sup> root and subtract one.

D. Add one to each return, multiply them together, take the T<sup>th</sup> root and subtract one. | You must **outline** the study’s most important findings rather than explain them in detail. **Correct Answer: D** |

| Recognize | Recall the purpose of a given equation or term, and its name when appropriate. **Note:** When the action word “recognize” is used and applied to an equation, the equation will be provided within the question stem or the correct answer choice. | What is the following equation called and used for in the context of artificial neural networks?

\[ \sigma(z) \equiv \frac{1}{1 + e^{-z}} \]

A. It is called a neuron and used to make a NAND gate.

B. It is called a sigmoid function and is used to model sigmoid neurons that better enable learning than perceptrons do.

C. It is called a perceptron which is used to create a smoother function than a logistic function. | You must **recognize** that this is a sigmoid function, also referred to as a logistic function, which is used as the basis for a sigmoid neuron. **Correct Answer: B.** |
## Action Words continued

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
<th>Question Example</th>
<th>Example of Term Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relate</td>
<td>Show or establish logical or causal connection</td>
<td>Which of the following effects does NOT help to explain growth in the venture capital industry?</td>
<td>You must relate effects or factors (e.g., the prudent man rule) to another result or concept (e.g.) growth in an industry.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A. Amendments to the prudent man rule</td>
<td>Correct Answer: C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B. The rise of limited partnerships as an organizational form</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C. Decline in the valuations of small capitalization stocks</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>The activities of investment advisors in the venture capital market</td>
<td></td>
</tr>
</tbody>
</table>
CAIA Editorial Staff

Hossein Kazemi, PhD, CFA, Senior Advisor, The CAIA Association
Mirjam Dekker, Project Manager, The FDP Institute
Kathryn Wilkens, PhD, CAIA, Curriculum Advisor, The FDP Institute
Satyabrota Das, CAIA, CFA, Exam Advisor, The FDP Institute

No part of this publication may be reproduced or used in any form (graphic, electronic or mechanical, including photocopying, recording, taping or information storage and retrieval systems) without permission by Financial Data Professional Institute, Inc. (“FDP”). The views and opinions expressed in the book are solely those of the authors. This book is intended to serve as a study guide only; it is not a substitute for seeking professional advice.

FDP disclaims all warranties with respect to any information presented herein, including all implied warranties of merchantability and fitness. All content contained herein is provided “AS IS” for general informational purposes only. In no event shall FDP be liable for any special, indirect or consequential changes or any damages whatsoever, whether in an action of contract, negligence or other action, arising out of or in connection with the content contained herein. The information presented herein is not financial advice and should not be taken as financial advice. The opinions and statements made in all articles and introductions herein do not necessarily represent the views or opinions of FDP.