FDP INSTITUTE® by CAIA

FDP CHARTER CANDIDATE STUDY GUIDE

April 10 - April 24, 2023

Learning objectives and keywords to facilitate your exam study

Brought to you by: CAIA ASSOCIATION®

INTRODUCTION TO THE FINANCIAL DATA PROFESSIONAL (FDP) PROGRAM

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

The digital revolution has disrupted the financial industry in recent years. It is critical for industry practitioners to have a working knowledge of the increasingly important roles played by big data, machine learning, and artificial intelligence in the financial industry. The FDP Institute has designed this self-study program to provide finance professionals with an efficient path to learn about financial data science's essential aspects. The FDP curriculum introduces Candidates to the central concepts of machine learning and big data, including ethical and privacy issues and their roles in various financial industry segments. Candidates will earn their FDP Charter once they pass the FDP exam and complete the online class requirements, which can be done before or after the FDP exam.

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 data science applications to the financial industry.

Passing the FDP examination is an important accomplishment that will require significant preparation. All Candidates will need to study and become familiar with the FDP curriculum material 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 (LOs) that define the content to be tested on the exam. The learning objectives are an important way for Candidates to organize their studies as they form the basis for examination questions. All learning objectives reflect the FDP curriculum content, and all exam questions are written to address the learning objectives directly. A Candidate who can 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 Charter must complete the FDP exam and the online requirements. Since the FDP program is designed for finance professionals, it is assumed that Candidates understand the central concepts of financial economics. Candidates are expected to have knowledge of various financial institutions and instruments' roles and characteristics and the financial models these institutions employ to value the instruments and measure their risk. These concepts are covered in CAIA®, CFA®, and FRM® exams and dedicated undergraduate or graduate courses covering financial markets, investments, and risk management.

FDP PROGRAM: ONLINE REQUIREMENTS

FDP Candidates must complete the following two components with a passing score before obtaining their FDP Charters.

- **FDP exam.**
- **Online classes covering Python or R programming.**

The FDP exam will not contain any coding questions. However, FDP Candidates must demonstrate some Python or R programming language knowledge before they obtain their FDP charter. FDP Candidates who do not have a verifiable academic background in Python or R programming can demonstrate their understanding of these languages by completing the online classes listed below. The online classes can be completed before or after a Candidate completes the FDP exam.

The FDP Institute recommends DataCamp's introductory online courses (https://www.datacamp.com) for completing the FDP Charter's requirement. The list of online classes offered by DataCamp appears on the FDP Institute's website and in this Study Guide.

The approved online classes offered by DataCamp are available as soon as a Candidate creates an account on DataCamp. Limited free access to their classes is available. DataCamp courses assume no prior knowledge of Python or R.

The Candidate Handbook, which can be found on the FDP website, describes the procedure for sending proof of successful completion of the online classes to the FDP Institute.

The following classes are recommended to complete the FDP Charter's programming knowledge requirement. These recommendations assume that a Candidate has no prior Python or R programming knowledge. If a Candidate has some knowledge of these languages, the Candidate is encouraged to take more advanced Python or R programming classes at DataCamp. If a Candidate has a verifiable academic background in Python or R, the Candidate can seek an exemption from the online classes. The approval of prior academic knowledge in Python or R programming is at the sole discretion of the FDP Institute. Please contact the FDP institute to learn more about this option.

FDP Candidates can satisfy the coding requirement of the FDP program by completing two (2) Python or two (2) R classes offered by DataCamp. DataCamp classes can be accessed through its website at https://www.datacamp.com/. Candidates are responsible for the cost of classes offered at DataCamp. Candidates are encouraged to take advantage of the limited free access offered by DataCamp to evaluate its teaching method. The classes listed below are short, and depending on the Candidate's background, each can be completed in four (4) to six (6) hours. Besides the classes mentioned below, Candidates have the option to complete any two (2) classes in either Python or R.

DataCamp: Python

1. **Introduction to Python**
   https://www.datacamp.com/courses/intro-to-python-for-data-science
2. **Intermediate Python**
   https://www.datacamp.com/courses/intermediate-python-for-data-science

DataCamp: R

1. **Introduction to R**
   https://www.datacamp.com/courses/free-introduction-to-r
2. **Intermediate R**
   https://www.datacamp.com/courses/intermediate-r

FDP EXAMINATION

The FDP examination, administered twice annually, is a four-hour computer-administered examination offered at test centers worldwide. The FDP examination consists of eighty (80) multiple choice questions weighted as 75% of the total points and two (2) to four (4) constructed response questions (multi-part essay type) weighted as 25% of the total points. The FDP exam will not contain any Python or R programming questions.

The FDP program is organized to facilitate quick learning and easy retention based on the study guide. 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 studies as they form the basis for examination questions. All learning objectives reflect the FDP curriculum content, and all examination questions are written to address the learning objectives directly.

For additional information about the FDP examination, please see the Candidate Handbook, which can be found on the FDP Institute website.

SAMPLE EXAM AND PRACTICE QUESTIONS

A sample exam is available for the Candidates to assist with their study efforts.
This sample exam contains eighty (80) multiple choice questions and several multi-part constructed response questions. There is also a set of practice questions available to Candidates. The set of practice questions contains more questions than the number of questions in the actual exam. In addition to helping the Candidates learn the topic material, the questions can also help the Candidates get familiar with the style and conventions used. An example is a simplifying convention of using the natural logarithm to solve any problem requiring the calculation of logarithm on the exam. This convention is announced at the beginning of the sample exam and on the actual exam. This convention is also described in the Candidate Handbook.

OTHER STUDY TOOLS AND RESOURCES

In addition to this Study Guide and the Candidate Handbook, the FDP Institute website directs Candidates to the readings covered in the curriculum. The readings are detailed below by topic area and include textbooks, often used across topics, and individual articles that are usually topic-specific. Both types of readings can be purchased from Amazon or the publisher, and whenever possible, they are posted on the FDP Institute website.

Page Number References for Keywords
For Candidates' convenience, six (6) articles published by PMR Journals are provided in one collection titled "Alternative Data and Machine Learning in the Financial Industry: *A Collection of Articles from PMR Journals*." It is available at a discounted price of $99 for registered Candidates. There are two sets of page numbers in this collection: one corresponds to the

collection's table of contents. In contrast, the other corresponds to each article's page number in the original journal. The page numbers appearing next to the keywords refer to the page numbers as they appeared in the original article.

Note: *Check if your employer has a subscription to Portfolio Management Research (PMR) as this might provide free access to the six (6) PMR readings.*

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 submit their certificates of completion for the required online classes. The rest of this document discusses the FDP curriculum. Below is the outline of the curriculum:

Topics		Approximate Weight %
1.	Introduction to Data Science	5-12
2.	Linear and Logistic Regression, Support Vector Machines, Regularization, and Time Series	10-15
3.	Decision Trees, Supervised Segmentation, and Ensemble Methods	10-15
4.	Classification, Clustering, and Naïve Bayes	5-12
5.	Neural Networks and Reinforcement Learning	5-12
6.	Performance Evaluation, Back-Testing, and False Discoveries	5-12
7.	Text Mining	5-12
8.	Ethical and Privacy Issues	5-12
9.	Fintech Applications	25-40

THE FDP CURRICULUM: COMPLETE READING LIST

The following is a complete list of all curriculum readings for the April 2023 FDP exam.

Two (2) out of the three (3) books and six (6) articles from the Alternative Data and Machine Learning in the Financial Industry: A Collection of Articles from PMR Journals must be purchased. One of the books is available free of charge from the authors' website.

The Topics in Financial Data Science articles are available free of charge. Candidates may access all materials from the authors' or publishers' websites or via the FDP website. Please use the web link below to access all curriculum materials.

https://fdpinstitute.org/Curriculum-Materials

A. Books

1. Provost, F., and T. Fawcett (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media Inc., 1st Edition. Chapters 1-10. Candidates should visit the book's errata page.

2. John C. Hull (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapters 1-11.

3. James, G., D. Witten, T. Hastie, and R. Tibshirani (2021). An Introduction to Statistical Learning: With Applications in R. Springer, 2nd Edition. Chapters 1, 2 (sections 1,2), Chapter 3 (sections 1-3), Chapter 6 (sections 1-3), Chapter 8 (sections 1,2). Candidates should visit the book's errata page.

B. Alternative Data and Machine Learning in the Financial Industry: *A Collection of Articles From PMR Journals*

1. Das S., M. Donini, J. Gelman, K. Haas, M. Hardt, J. Katzman, K. Kenthapadi, P. Larroy, P. Yilmaz, and M. B. Zafar (2021). Fairness Measures for Machine Learning in Finance. The Journal of Financial Data Science, 3(4): 33-64. Reading 8.3

2. Ekster, G. and Kolm, P. N. (2021). Alternative Data in Investment Management: Usage, Challenges, and Valuation. The Journal of Financial Data Science, 3(4): 10-32. Reading 9.1

3. Åstebro, T. (2021). An Inside Peek at AI Use in Private Equity. The Journal of Financial Data Science, 3(3): 97-107. Reading 9.5

4. Li, Y., Z. Simon, and D. Turkington. (2022). Investable and Interpretable Machine Learning for Equities. The Journal of Financial Data Science, 4(1): 54-74. Reading 9.6

5. López de Prado, M. (2018). The 10 Reasons Most Machine Learning Funds Fail. The Journal of Portfolio Management, 44 (6): 120-133. Reading 9.7

6. Harvey, C. R., and Y. Liu. (2014). Evaluating Trading Strategies. The Journal of Portfolio Management, 40(5): 108-118. Reading 9.8

C. Topics in Financial Data Science

1. Das, S., and H. Kazemi (2022). Time Series: A Financial Perspective. The FDP Institute. This reading is provided by the FDP Institute free of charge. Reading 2.4
2. Colquhoun, D. (2014). An Investigation of the False Discovery Rate and the Misinterpretation of p-values. Royal Society Open Science, 1 (3): 1-16. Reading 6.3
3. Zhao, F. (2017). Natural Language Processing – Part I: Primer. S&P Global: Market Intelligence. Reading 7.3
4. Institute of International Finance (May 2019). Machine Learning Thematic Series Part II: Bias and Ethical Implications. Reading 8.2
5. OECD (2021). Artificial Intelligence, Machine Learning and Big Data in Finance: Opportunities, Challenges, and Implications for Policy Makers. Reading 9.2
6. Financial Stability Board (2017). Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications. Reading 9.3
7. Zappa, D., M. Borrelli, G.P. Clemente, and N. Savelli. (2021). Text Mining in Insurance: From Unstructured Data to Meaning. Variance Journal, 14(1). Reading 9.4
8. Amler, H., L. Eckey, S. Faust, M. Kaiser, P. Sandner, and B. Schlosser. (2021). DeFi-ning DeFi: Challenges & Pathway. Reading 9.9
9. Nadini, M., L. Alessandretti, F. D. Giacinto, M. Martino, L. M. Aiello, and A. Baronchelli. (2021). Mapping the NFT Revolution: Market Trends, Trade Networks, and Visual Features. Reading 9.10

ACTION WORDS

In each learning objective that appears below, action words are used to direct Candidates' focus of study. The following table contains the list of all action words used in this study guide and their definitions.

Action Word	Meaning
Analyze	To examine methodically and in detail the constitution or structure of the information or concept covered by the LO. This is similar to offering an explanation and an interpretation. It is used chiefly to explain relationships.
Apply	To bring into action, use or employ a concept or a mathematical relationship (equation). If the LO is about an equation, the Candidate must memorize the equation (see Recognize below).
Calculate	It is similar to Apply but is related to a mathematical concept and equation. If the LO is about an equation, the Candidate must memorize the equation (see Recognize below).
Compare	Estimate, measure, or note the similarity or dissimilarity between two concepts or definitions.
Contrast	Similar to Compare. In this case, the emphasis is on the differences.
Define	A general action word. The Candidate is expected to state or describe precisely the nature, scope, or meaning of a concept. If the LO is about a mathematical equation, the Candidate is not expected to memorize the exact equation but is expected to describe its essential aspects.
Describe	Similar to Define. The Candidate should give an account in words of concepts covered by the LO. The Candidate is expected to cover all the relevant characteristics, qualities, or relationships the LO covers. If the LO is about a mathematical equation, the Candidate is not expected to memorize the exact equation but is expected to describe its essential aspects.
Discuss	It is similar to Analyze. To provide details about a key word or concept. If the LO is about an equation, the Candidate does not need to memorize the equation, but must know its uses and applications.

Action Word	Meaning
Explain	Similar to Describe. The Candidate is expected to clarify an idea, problem, or relationship by describing it in more detail or revealing relevant facts or ideas. If the LO is about a mathematical equation, the Candidate is not expected to memorize the exact equation but is expected to describe its essential aspects.
Identify	The Candidate is expected to recognize or establish as being a particular model, concept, or relationship. The LO may expect the Candidate to verify a given relationship or recognize a particular pattern. If the LO is about a mathematical equation, the Candidate is not expected to memorize the exact equation but is expected to describe its essential aspects.
Interpret	Similar to Explain. The Candidate is expected to give or provide an explanation for the observed pattern, relationship, or information. If the LO is about a mathematical equation, the Candidate is not expected to memorize the exact equation but is expected to describe its essential aspects.
List	The Candidate is expected to learn the list of related items or concepts covered by the LO. The Candidate is not expected to describe the members of the list. A separate LO may state that some or all of the members of the list must be explained.
Recognize	The Candidate is expected to identify an equation or model that has appeared in the readings. The Candidate is not expected to memorize the equation. The Candidate is expected to apply the equation or make some calculations using the equation provided on the exam.

LEARNING OBJECTIVES

Topic 1. Introduction to Data Science

Reading 1.1 Provost, F. and T. Fawcett (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media Inc., 1st Edition. Chapters 1 and 2.

Keywords

Data mining (p. 2)
Data Science (p. 2)
Data-driven decision making (p.5)
Big data (p. 8)
Classification (p. 20)
Regression (p. 21)
Similarity matching (p. 21)
Clustering (p. 21)
Co-occurrence grouping (p. 21)
Profiling (p. 22)
Link prediction (p. 22)
Data reduction (p. 22)
Causal modeling (p. 23)
Unsupervised learning (p. 24)
Supervised learning (p. 24)
Leak (p. 30)

Learning Objectives

Demonstrate proficiency in the following areas:

1.1.1 Data Analytic Thinking (Ch. 1)

For example:

A. List examples of data mining in finance, marketing, and customer relationship management.

B. Contrast data science with data mining.

C. Describe the two types of decisions that can benefit from data-driven decision making.

D. Describe the reason for the early adoption of automated decision making by finance and telecommunications industries.

E. Contrast data science with data processing.

F. Describe the usage of big data.

G. Explain why both appropriate data and data scientists are required to extract useful knowledge from data.

H. Explain why it is necessary to understand data science even if someone is not going to use data science directly.

I. List and describe the four fundamental concepts of data science.

1.1.2 Business Problems and Data Science Solutions (Ch. 2)

For example:

A. Describe when each type of data mining algorithm, such as classification, regression, similarity matching, clustering, co-occurrence grouping, profiling, link-prediction, data reduction, and causal modeling, should be used.

B. Explain the differences between regression and classification.

C. Contrast supervised learning with unsupervised learning.

D. List the algorithms that can be used for supervised and unsupervised learning.

E. Contrast data mining with the use of data mining results.

F. List and describe the steps used in Cross Industry Standard Process for Data Mining (CRISP-DM).

G. Explain the reason for having an iterative process involved in CRISP-DM.

H. Describe the characteristics of credit card and Medicare fraud.

I. List the reasons for deploying the data mining system itself rather than the models produced by a data mining system.

Reading 1.2 John C. Hull (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapter 1.

Keywords

Machine learning (p. 1)
Artificial intelligence (p. 1)
Features (p. 6)
Labels (p. 6)
Semi-supervised learning (p. 7)
Training set (p. 8)
Root-mean squared error (p. 9)
Bias-variance tradeoff (p. 15)
Numerical feature (p. 16)
Categorical feature (p. 16)
Outliers (p. 17)
Bayes' Theorem (p. 18)

Learning Objectives

Demonstrate proficiency in the areas of:

1.2.1 Introduction

For example:

A. List the advantages for the society of replacing human decision-making with machines.

B. Contrast machine learning to statistics.

C. Describe a training set, validation set, and test set.

D. Define instances.

E. Analyze the relationship between model error and model complexity.

F. Define bias and variance in the context of machine learning.

G. List the usage of the training set, validation set, and test set.

H. List and explain different data cleaning issues.

I. List the type of models that are least and most affected by outliers.

J. Calculate conditional probability using Bayes' Theorem.

Reading 1.3 James, G., D. Witten, T. Hastie, and R. Tibshirani. An Introduction to Statistical Learning: With Applications in R. Springer, 2nd Edition. Chapters 1, 2.1, and 2.2.

Keywords

Statistical learning (p. 1)
Flexible (p. 22)
Thin plate spline (p. 23)
Classification problems (p. 28)
Quantitative variables (p. 28)
Qualitative response (p. 28)
Binary response (p. 28)
Predictors (p. 29)
Mean squared error (MSE) (p. 29)
Test MSE (p. 30)
Test data (p. 30)
Training MSE (p. 30)
Degrees of freedom (p. 31)
Expected test MSE (p. 34)
Bias (p. 35)
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 areas of:

1.3.1 Organization and Resources of the Book a 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.

1.3.2 Statistical Learning (Ch. 2.1)

For example:

A. Explain why we estimate a function with data, including the role of input and output variables and their synonyms.

B. Explain various error terms (reducible and irreducible), the expected value of error squared, and the variance of error terms.

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

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

E. Explain when a supervised learning model is preferable to unsupervised or semi-supervised learning models.

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

1.3.3 Assessing Model Accuracy (Ch. 2.2)

For example:

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

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

C. Explain the purpose of cross-validation.

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

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

F. Calculate the Bayes error rate.

G. Explain and apply the Bayesian classifier.

H. Explain how the K-nearest neighbors (KNN) classifier is related to the Bayes classifier and how the choice of K impacts results.

I. Calculate the conditional probability of a point belonging to a particular class.

J. Analyze the relationship between the value of K and the bias-variance tradeoff for a KNN classifier.

K. Explain what happens to the decision boundary as K increases in a KNN classifier.

Topic 2. Linear and Logistic Regression, Support Vector Machines, Regularization, and Time Series

Reading 2.1 Provost, F. and T. Fawcett (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media Inc., 1st Edition. Chapter 4.

Keywords

Parameter learning or parametric modeling (p. 81)
Linear classifier (p. 85)
Linear discriminant (p. 86)
Hyperplane (p. 86)
Parameterized model (p. 86)
Objective function (p. 88)
Margin (p. 92)
Support vector machine (SVM) (p. 92)
Hinge-loss (p. 94)
Zero-one loss (p. 95)
Squared error (p. 95)
Odds (p. 97)
Log-odds (p. 99)
Logistic function (p. 101)
Nonlinear SVM (p. 107)
Neural networks (p. 108)

Learning Objectives

Demonstrate proficiency in the areas of:

2.1.1 Classification via Mathematical Functions

For example:

A. Apply the equation of a straight line using slope and intercept.
B. Describe, apply, and interpret a linear discriminant.
C. Recognize the classification function for a linear discriminant.
D. Calculate the best value for the parameters of a linear discriminant for a set of instances.
E. Describe decision boundaries in 2-dimensions, 3-dimensions, and higher dimensions.
F. Interpret the magnitude of a feature's weight in a general linear model.
G. Describe the general idea behind optimizing the objective function for a linear discriminant for a particular data set.
H. Describe how linear discriminant functions can be used for scoring and ranking instances.
I. Analyze the relationship between the distance from decision boundary of a linear discriminant and the likelihood of response.
J. Describe the important idea behind the Support Vector Machine (SVM).
K. Describe the objective function of the SVM.
L. Explain how the objective function used in SVM utilizes the concept of hinge-loss function.
M. Describe the reason for not using a squared loss function in classification problems.

2.1.2 Regression via Mathematical Functions

For example:

A. Describe the major drawback of least-squares regression.
B. Calculate odds and log odds.
C. List the important features of logistic regression.
D. Calculate the log-odds linear function.
E. Calculate class probability using the logistic function.
F. Describe the shape of the logistic function.
G. Describe the decision boundary for the logistic regression.
H. Describe how an objective function is formed in the logistic regression.
I. Compare and contrast classification trees with linear classifiers.
J. Explain the basic idea behind nonlinear SVMs and neural networks.

Reading 2.2 John C. Hull (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapters 3 and 5.

Keywords

Polynomial regression (p. 53)
One-hot encoding (p. 54)
Dummy variable trap (p. 55)
Regularization (p. 56)
Logistic regression (p. 69)
Sigmoid function (p. 70)
Balanced data set (p. 108)
Support vectors (p. 110)
Hard margin classification (p. 114)
Soft margin classification (p. 114)
Gaussian radial basis function (p. 118)
SVM regression (p. 119)

Learning Objectives

Demonstrate proficiency in the following areas:

2.2.1 Supervised Learning (Ch. 3)

For example:

A. List the conditions that must be satisfied for linear regression to be valid.
B. List the steps used in the gradient descent method.
C. Calculate the probability of a positive outcome using the sigmoid function.
D. Recognize the cost function for the logistic regression.
E. Analyze the effect of using different types of regularization on logistic regression.

Note that this chapter contains many other topics with no learning objectives specified in this section. Candidates are still encouraged to read these sections to understand subsequent material better. Questions from these missing topics will primarily be asked from the book An Introduction to Statistical Learning: With Applications in R by James, G., D. Witten, T. Hastie, and R. Tibshirani (Reading 2.3).

2.2.2 Support Vector Machines (Ch. 5)

For example:

A. List the advantages and disadvantages of using support vector machines (SVM).

B. Describe the reason for normalizing data before using it in SVM.

C. Calculate the dimension of a separating hyperplane.

D. Recognize the equation of a separating hyperplane with m features.

E. Describe the reasons for using regularization in SVM.

F. Recognize the objective function used in creating SVM with m features.

G. Recognize the objective function for a soft margin classification.

H. Describe the type of regularization used in soft margin classification.

I. Describe how violations and misclassifications are measured in soft margin classification.

J. Analyze the relationship between the hyperparameter, C, and the width of the pathway for soft margin classification.

K. Describe the general approach to finding a non-linear boundary when using a linear model.

L. Calculate Gaussian radial basis function (RBF) for an observation.

M. Explain the effect of the parameter gamma on RBF.

2.2.3 SVM Regression (Ch. 5)

For example:

A. Describe how an error is calculated in SVM regression.

B. Recognize the equations of hyperplanes in SVM regression.

C. Recognize the objective function used in SVM regression.

D. Describe the interaction of the two terms in the objective function of an SVM regression.

E. Contrast simple linear regression with SVM.

Reading 2.3 James, G., D. Witten, T. Hastie, and R. Tibshirani. An Introduction to Statistical Learning: With Applications in R. Springer, 2nd Edition. Chapters 3.1, 3.2, 3.3, 6.1, 6.2, and 6.3.

Keywords

Residual (p.61)
Residual sum of squares (p. 63)
Population regression line (p. 63)
Least squares line (p. 63)
Bias (p. 65)
Unbiased (p. 65)
Standard error (pg. 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. 68)
Total sum of squares (p. 70)
F-statistic (p. 75)
Forward selection (p. 79)
Backward selection (p. 79)
Mixed selection (p. 79)
Prediction interval (p. 82)
Dummy variable (p. 83)
Additive (p. 87)
Linear (p. 87)
Hierarchical principle (p. 89)
Residual plot (p. 93)
Heteroscedasticity (p. 96)
Outlier (p. 97)
Collinearity (p. 99)
Power (p. 101)
Multicollinearity (p. 102)
Variance inflation factor (p. 102)
Feature selection (p. 226)
Variable selection (p. 226)
Best subset selection (p. 227)
Deviance (p. 228)
Forward stepwise selection (p. 229)
Backward stepwise selection (p. 231)
C_p (p. 233)
Akaike information criterion (AIC) (p. 233)
Bayesian information criterion (BIC) (p. 233)
Adjusted R^2 (p. 233)
Ridge regression (p. 237)
Tuning parameter (p. 237)
Shrinkage penalty (p. 237)
ℓ_2 norm (p. 238)
Scale equivariant (p. 239)
Lasso (p. 241)
ℓ_1 norm (p. 241)
Sparse (p. 242)
Soft-thresholding (p. 248)
Signal and noise variables (p. 250)
Dimension reduction methods (p. 251)
Linear combination (p. 251)
Principal component analysis (p. 252)
Principal component scores (p. 254)
Orthogonal (p. 256)
Principal component regression (p. 252)
Partial least squares (p. 260)

Learning Objectives

Demonstrate proficiency in the following areas:

2.3.1 Simple Linear Regression (Ch. 3.1)

For example:

A. Calculate the value of RSS.

B. Calculate the least-squares coefficient estimates.

C. Interpret least-squares coefficients.

D. Calculate the standard error of a statistic.

E. Apply standard errors of linear regression.

F. Calculate the 95% confidence interval.

G. Calculate the t-statistic.

H. Explain the rules for rejecting the null hypothesis using p-values.

I. Explain the accuracy of linear regression.

J. Calculate and interpret the R^2 statistic.

K. Describe the advantages of the R^2 statistic over the RSE.

L. Calculate correlation from R^2 for the simple linear regression.

2.3.2 Multiple Linear Regression (Ch. 3.2)

For example:

A. Interpret the coefficients of multiple linear regression.

B. Describe how a multiple linear regression tests the relationship between responses and predictors.

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

D. Explain how the F-statistic can be used for hypothesis testing.

E. Explain why the value of the t-statistic can be a misleading indicator of variable importance in a multiple regression.

F. Describe how to determine the importance of variables in a multiple regression.

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

H. Calculate RSE given the values of RSS, n, and p.

2.3.3 Considerations in the Regression Model (Ch. 3.3)

For example:

A. Apply dummy variables.

B. Describe how to use qualitative variables with more than two levels in a multiple regression.

C. Interpret the coefficients of a dummy variable.

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

E. Describe the interaction effect.

F. Interpret the coefficients of an interaction term.

G. Explain when an interaction term should be added to a multiple regression model.

H. Describe the potential problems related to non-linearity, correlation of error terms, non-constant variance of error terms, outliers, high-leverage points, and collinearity for a linear regression model.

I. Explain what happens to standard errors and confidence intervals in the presence of correlated errors.

J. Explain how heteroscedasticity can be mitigated using data transformation.

K. Describe high leverage points and leverage statistic.

L. Explain how high leverage points can be detected using the leverage statistic.

M. Describe the range of values for the variance inflation factor.

N. Calculate the variance inflation factor.

2.3.4 Subset Selection (Ch. 6.1)

For example:

A. Define the best subset selection.

B. List the steps used in the best subset selection.

C. Analyze the relationship between the number of variables and RSS (or R^2) for a multiple linear regression.

D. Explain the effect of low RSS (or high R^2) on training and test error.

E. Explain the role of deviance in a logistic regression model.

F. Analyze the relationship between the value of deviance and fit of a model.

G. Describe the key drawback of using the best subset selection.

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

I. Explain the advantage of forward stepwise regression over the best subset selection method.

J. Describe a disadvantage of forward stepwise regression and backward stepwise regression relative to the best subset selection model.

K. Describe a key requirement for the number of samples and predictors when using the backward stepwise regression.

L. Describe the hybrid approach of using forward and backward stepwise regression together.

M. List the two common approaches used to select the best model concerning test error.

N. Explain the reason for not using training set RSS and training set R^2 for selecting the best model from a set of models with different predictors.

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

P. Describe the decision rule for selecting a model based on C_p, AIC, and BIC.

Q. Analyze the interaction between the RSS and the penalty term in C_p, AIC, and BIC.

R. Calculate the adjusted R^2.

2.3.5 Ridge Regression (Ch. 6.2)

For example:

A. Recognize the objective function of ridge regression.

B. Explain the effect of the tuning parameter on the coefficients in ridge regression.

C. Explain when the ridge regression is equivalent to the least-squares regression model.

D. Explain when the ridge regression is equivalent to the null model.

E. Calculate the ℓ_2 norm.

F. Explain the effect of multiplying a predictor by a factor before using it in the ridge regression.

G. Describe standardizing the predictors.

H. Explain what happens to the bias-variance trade-off as the tuning parameter changes in the ridge regression.

I. Describe what happens to the least-squares coefficients when the number of variables is as large as the number of observations.

J. Describe when ridge regression can be used, but least-squares regression cannot be used.

K. Describe the advantage of ridge regression over best subset selection.

2.3.6 The Lasso (Ch. 6.2)

For example:

A. Describe the key disadvantage of ridge regression.

B. Describe the advantage of Lasso over ridge regression.

C. Recognize the objective function of Lasso.

D. Calculate the ℓ_1 norm.

E. Describe the variable selection property of Lasso.

F. Explain the effect of the tuning parameter on the coefficients in Lasso.

G. Recognize the alternative formulation of the objective function for Lasso and ridge regression.

H. Analyze the impact of the size of the budget in estimating Lasso and ridge regression.

I. Explain the graphical interpretation of Lasso and ridge regression when there are two features.

J. Describe the geometric shape of the constraint for Lasso and ridge regression in two or more dimensions.

K. List the key advantage of Lasso over ridge regression.

L. Describe when Lasso is expected to perform better than ridge regression and when ridge regression is expected to perform better than Lasso.

M. Explain the relationship between best subset selection and Lasso or ridge regression.

N. Explain the type of shrinkage done by Lasso and ridge regression.

O. Describe the rule governing the selection of the tuning parameter for Lasso and ridge regression.

P. Analyze the expected values of coefficients for the signal and noise variables for a robust regression model.

2.3.7 Principal Component Analysis (Ch. 6.3)

For example:

A. List the two ways of controlling variance.

B. Describe the relationship between the number of features and the number of parameters estimated in a dimension reduction method.

C. List the two major steps used in a dimension reduction method.

D. Explain the characteristics of the first principal component.

E. Explain the meaning of projecting a point on a line.

F. Describe the constraint that must be used to find the loadings for the principal components.

G. Describe an alternative interpretation for principal component analysis (PCA).

H. Explain the information content of the first principal component.

I. Explain the effect of zero correlation between the first and the second principal component.

J. Explain orthogonal properties of principal components.

K. Explain the expected information content of the second principal component when there are two predictors.

L. Analyze the relationship between the number of principal components and the number of features.

M. Describe the key assumption behind the use of principal component regression (PCR).

N. Explain the problem mitigated by using PCR provided the assumptions underlying PCR holds.

O. Explain when PCR is expected to perform better than linear regression with all features.

P. Explain why PCR is not a feature selection method.

Q. Describe the equivalence between the PCR and ridge regression.

R. Explain the process of selecting the number of principal components.

S. Describe when to standardize features before using PCR and when not to standardize the features for using them in PCR.

2.3.8 Partial Least Squares (Ch. 6.3)

For example:

A. List the key drawback of principal component regression (PCR).

B. Describe the key difference between PCR and partial least squares (PLS).

C. Describe the way the first PLS is found.

D. Analyze the impact of least-squares coefficients from the simple linear regression of each feature on the weight of the first PLS.

E. Describe the process of finding the second PLS.

Reading 2.4 Das, S. and H. Kazemi (2022). Time Series: A Financial Perspective. The FDP Institute.

Keywords

Strictly stationary (p. 3)
Weakly stationary (p. 4)
Gaussian White noise (p. 4)
Random walk (p. 4)
Simple moving average (p. 6)
Weighted moving average (p. 7)
Exponentially weighted moving average (p. 7)
Autoregressive model (p. 10)
Moving average process (p. 18)
Autoregressive moving average models (p. 21)
Homoskedasticity (p. 21)
Volatility clustering (p. 25)
Engle's ARCH test (p. 27)
Persistence Parameter (p. 28)

Learning Objectives

Demonstrate proficiency in the following areas:

2.4.1 Stationary Time Series and Moving Average Methods

For example:

A. Explain the reason for making a time series stationary before analyzing it.

B. Describe how stock prices can be converted to a stationary series.

C. List the conditions that are satisfied by a strictly stationary time series.

D. Describe the process that can be used to detect the presence of stationarity.

E. Describe the characteristics of the autocorrelation function of a stationary series and a non-stationary series.

F. List the conditions satisfied by Gaussian white noise.

G. Analyze the effect of the size of the window on a simple moving average (SMA).

H. List a key advantage of using an SMA.

I. Describe the most common range of values for the weighting parameter in the exponentially weighted moving average (EWMA).

J. Calculate the value of EWMA for a series.

K. Explain the impact of the weighting parameter on EWMA.

L. Describe the choice of the weighting parameter that makes EWMA equivalent to SMA.

2.4.2 Autoregressive Models

For example:

A. Describe one of the key differences between autoregressive models and moving average methods.

B. List the conditions required for an AR(1) process to be stationary.

C. Analyze the process that pulls a stationary AR(1) process close to its mean.

D. Calculate the mean, variance, autocovariance, and autocorrelation of a stationary AR(1) process.

E. Analyze the effect of the coefficients of a stationary AR(1) process on autocorrelation.

F. Calculate the values of an AR(1) process.

G. Explain how a shock affects a stationary AR(1) process.

H. Calculate the mean and the variance of a random walk.

I. Explain how a shock affects a random walk.

J. Explain why some financial time series cannot be modeled as a random walk process.

K. Calculate conditional forecast of mean and variance for a stationary AR(1) model.

2.4.3 Moving Average Models

For example:

A. Define the order of a moving average model.

B. Calculate the unconditional values of mean, variance, and autocovariance of a moving average model.

C. Calculate the conditional values of the mean and variance of an MA(1) model.

D. Calculate the unconditional mean of an ARMA(p, q) model.

E. Explain the characteristics of the autocorrelation function for an ARMA(p, q) model.

2.4.4 Volatility Models

For example:

A. Analyze the effect of heteroskedasticity on the standard errors and confidence intervals for least-squares regression.

B. Describe the advantages of ARCH and GARCH models.

C. List the stylized facts that indicate financial time series error terms are not homoscedastic.

D. List the challenges in creating a time-varying volatility model.

E. Explain how an ARCH(1) model satisfies the challenges of creating a time-varying volatility model.

F. Calculate the conditional and unconditional variance for the error term when an ARCH(1) model is used.

G. List the conditions that must be satisfied by the parameters of an ARCH(1) model for the model to be stationary.

H. Describe the weaknesses of the ARCH model.

I. Describe the restrictions imposed on GARCH(1,1) model parameters.

J. Calculate the long-term mean of volatility for a GARCH(1,1) model.

K. Explain the effect of the persistent parameter on a GARCH(1,1) model.

L. Analyze the equivalence between an ARCH(1) and a GARCH(1,1) model.

M. Calculate the forecasted value of volatility using a GARCH(1,1) model.

Topic 3. Decision Trees, Supervised Segmentation, and Ensemble Methods

Reading 3.1 Provost, F. and T. Fawcett (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media Inc., 1st Edition. Chapters 3 and 5.

Keywords

Information (p. 43)
Tree induction (p. 44)
Predictive model (p. 45)
Instance (p. 46)
Descriptive modeling (p. 46)
Feature vector (p.46)
Target variable (p. 46)
Attributes or features (p. 46)
Model induction (p. 47)
Deduction (p. 47)
Training data (p. 47)
Labeled data (p. 47)
Supervised segmentation (p. 48)
Entropy (p. 51)
Information gain (p. 51)
Parent set (p. 52)
Children set (p. 52)
Variance (p. 56)
Entropy graph/chart (p. 58)
Leaf (p. 63)
Decision nodes (p. 63)
Classification tree (p. 63)
Regression tree (p. 64)
Probability estimation tree (p. 64)
Decision line (p. 69)
Decision surface or boundary (p. 69)
Frequency-based estimation of class partitions (p.70)
Membership probability (p. 72)
Laplace correction (p. 73)
Generalization (p. 112)
Generalization performance (p. 113)
Overfitting (p. 113)
Fitting graph (p. 113)
Holdout data (p. 113)
Test set (p. 114)
Base rate (p. 115)
Sweet spot (p. 117)
Cross-validation (p. 126)
Folds (p. 127)
Learning curve (p. 131)
Complexity (p. 131)
Sub-training set (p. 134)
Pruning (p. 134)
Validation set (p. 134)
Nested holdout testing (p. 134)
Nested cross-validation (p. 135)
Sequential forward selection (p. 135)
Sequential backward elimination (p. 135)
Penalty function (p. 138)

Learning Objectives

Demonstrate proficiency in the following areas:

3.1.1 Models, Induction and Prediction (Ch. 3)

For example:

A. Define prediction in the context of data science.

B. Compare and contrast predictive modeling with descriptive modeling.

C. Compare and contrast induction with deduction.

3.1.2 Supervised Segmentation (Ch. 3)

For example:

A. List the complications arising from selecting informative attributes.

B. Calculate the value of entropy.

C. Recognize and apply entropy with the maximum and minimum disorder.

D. Contrast the parent set with the children set.

E. Calculate information gain for children sets from a parent set.

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

G. Discuss the application of variance to numeric variables for supervised segmentation.

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

3.1.3 Visualizing Segmentations and Probability Estimation (Ch. 3)

For example:

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

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

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

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

E. Calculate the value of the Laplace correction.

F. Explain how one can determine the predictive power of each attribute.

3.1.4 Generalization, Overfitting, and Its Avoidance (Ch. 5)

For example:

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

B. Describe the relationship between complexity and error rates.

C. Describe the relationship between tree size and accuracy.

D. Apply the concept of overfitting in mathematical functions.

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

F. Explain why overfitting should be of concern.

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

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

I. List and describe strategies that can be used to avoid overfitting in tree induction.

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

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

L. Explain nested cross-validation.

M. Describe the main idea behind regularization.

N. Analyze the relationship between overfitting and multiple comparisons.

Reading 3.2 John C. Hull (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapter 4.

Keywords

Gini measure (p. 88)
Naïve Bayesian classifier (p. 94)
Bagging (p. 94)
Random forest (p. 95)
Boosting (p. 95)
Ensemble learning (p. 102)

Learning Objectives

Demonstrate proficiency in the following areas:

3.2.1 Decision Trees

For example:

A. Describe the advantages of decision trees over linear or logistic regression.

B. Describe and calculate entropy, information gain, and Gini measures.

C. Describe and calculate the confusion matrix for a decision tree.

D. Describe and calculate various points of a ROC curve given various confusion matrices.

3.2.2 The Naïve Bayes Classifier

For example:

A. Describe and apply Bayes' theorem.

B. Calculate conditional probabilities using Bayes' formula.

C. Explain the conditions under which the Naïve Bayes classifier can be applied.

D. Apply Naïve Bayes classifier to a decision tree problem.

E. Describe the criterion for determining the optimal feature choice and its threshold when the target is a continuous variable.

3.2.3 Ensemble Learning

For example:

A. Describe the primary idea behind ensemble learning.

B. Describe bagging with or without replacement.

C. Describe random forest.

D. Describe boosting.

Reading 3.3 James, G., D. Witten, T. Hastie, and R. Tibshirani. An Introduction to Statistical Learning: With Applications in R. Springer, 2nd Edition. Chapters 8.1 and 8.2.

Keywords

Tree based method (p. 327)
Terminal nodes or leaves (p. 329)
Internal nodes (p. 329)
Stratification (p. 330)
Top-down approach (p. 330)
Bottom-up approach (p. 330)
Recursive binary splitting (p. 330)
Subtree (p. 331)
Cost complexity (p. 332)
Weakest link (p. 332)
Classification error rate (p. 335)
Gini index (p. 336)
Weak learner (p. 340)
Majority vote (p. 341)
Out-of-bag observations (p. 342)
Variable importance (p. 343)
Stump (p. 347)
Interaction depth (p. 347)
Bayesian additive regression trees (p. 348)

Learning Objectives

Demonstrate proficiency in the following areas:

3.3.1 The Basics of Decision Trees (Ch. 8.1)

For example:

A. Apply and interpret a decision tree's predictions.
B. Explain and apply a regression tree and a partition.
C. Calculate and interpret RSS for a given partition (box).
D. Calculate RSS to perform recursive binary splitting.
E. Describe tree pruning, specifically cost complexity (weakest link) pruning.
F. Compare regression and classification trees.
G. Describe the construction of classification trees using the classification error rate, Gini index, and entropy.
H. Calculate the Gini Index.
I. Contrast tree-based methods and linear models.
J. Describe the advantages and disadvantages of trees.

3.3.2 Bagging, Random Forests, Boosting, and Bayesian Additive Regression Tree (Ch. 8.2)

For example:

A. Describe bagging and out-of-bag error estimation.
B. Explain how low variance procedures can be created from high variance ones.
C. Describe how qualitative targets are predicted using bagging.
D. Describe the out-of-bag error and its importance.
E. Describe how variable importance measures can be created using the Gini index.

F. Describe how random forest attempts to decorrelate trees.

G. Compare and contrast random forests to bagging.

H. Describe boosting as an approach for improving the prediction results from decision trees.

I. Explain why boosting is described as a slow learner.

J. Describe the key difference between BART and other ensemble methods, such as random forest and boosting.

Topic 4. Classification, Clustering, and Naïve Bayes

Reading 4.1 Provost, F. and T. Fawcett (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media Inc., 1st Edition. Chapters 6 and 9.

Keywords

Euclidean distance (p. 143)
Nearest neighbors (p. 144)
Combining function (p. 147)
Weighted voting (p. 150)
Similarity moderate voting (p. 150)
Complexity parameter (p. 152)
Classification boundaries (153)
Intelligibility (p. 154)
Feature selection (p. 156)
Domain knowledge (p. 156)
Manhattan distance (p. 158)
Jaccard distance (p.159)
Levenshtein metric (p. 160)
Hierarchical clustering (p. 164)
Dendrogram (p. 164)
Linkage function (p. 166)
Centroids (p. 169)
Clusters' distortion (p. 172)
CRISP process (p. 183)
Joint probability (p. 236)
Independent events (p. 236)
Unconditional probability (p. 237)
Bayes' Rule (p. 237)
Prior (p. 238)
Posterior probability (p. 238)
Likelihood (p. 240)
Conditional independence (p. 241)
Naïve Bayes equation (p. 241)
Lift (p. 244)

Learning Objectives

Demonstrate proficiency in the following areas:

4.1.1 Calculating and Interpreting Similarity and Distance (Ch. 6)

For example:

A. Calculate Euclidean distance.
B. Explain how combining functions can be used for classification.
C. Calculate the probability of belonging to a class based on the nearest neighbor classification.
D. Explain weighted voting (scoring) or similarity moderated voting (scoring).
E. Calculate contributions and class probabilities using weighted voting.
F. Explain how k in k-NN (Nearest-Neighbor) can be used to address overfitting.
G. Describe issues with nearest-neighbor methods focusing on intelligibility, dimensionality, domain knowledge, and computational efficiency.
H. Describe two aspects of intelligibility.
I. Explain how the curse of dimensionality could be fixed using domain knowledge.
J. Interpret and calculate Manhattan and Cosine distance.

K. Describe and interpret combining functions.

L. Describe the primary idea behind clustering.

M. Describe the primary idea behind hierarchical clustering.

N. Describe the general approach to k-means clustering using centroids.

O. Explain the role of supervised learning in interpreting cluster analysis results.

4.1.2 Combining Evidence Probabilistically (Ch. 9)

For example:

A. Calculate joint probability for independent and dependent events.

B. Explain and apply Bayes' Rule with the help of an example.

C. Calculate posterior probability, prior, and likelihood.

D. Explain and apply the naïve Bayes classifier.

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

F. List the advantages and disadvantages of the naïve Bayes classifier.

G. Explain and calculate lift in the context of the naïve Bayes method.

H. Define generative model and Naïve-Naïve Bayes.

Reading 4.2 John C. Hull (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapter 2.

Keywords

Scaled feature (p. 24)
Z-score (p. 24)
Min-max scaling (p. 24)
k-means (p. 25)
Inertia (p. 30)
Elbow method (p. 30)
Silhouette method (p. 31)
Gap statistic (p. 32)
Curse of dimensionality (p. 33)
Cosine function (p. 33)
Principal component (p. 41)
Factor loading (p. 42)

Learning Objectives

Demonstrate proficiency in the following areas:

4.2.1 Unsupervised Learning

For example:

A. Calculate and interpret feature scaling using Z-score and mini-max.

B. Calculate and interpret Euclidean distance.

C. Calculate and interpret the centroid of a cluster.

D. Describe the primary features and the process of implementing the k-means algorithm.

E. Calculate and interpret inertia as a measure of the clustering algorithm.
F. Describe the elbow method for selecting the number of clusters.
G. Describe and apply the silhouette method for selecting the number of clusters.
H. Describe and apply the gap statistic for selecting the number of clusters.
I. Describe the primary features of the hierarchical clustering method.
J. Describe the primary features of principal component analysis and how it relates to cluster analysis.

Topic 5. Neural Networks and Reinforcement Learning

Reading 5.1 John C. Hull (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapters 6, 7, and 8.

Keywords

Artificial neural network (ANN) (p. 125)
Multi-layer perceptrons (p. 125)
Hidden layer (p. 125)
Input layer (p. 126)
Output layer (p. 126)
Bias (p. 126)
Activation function (p. 126)
Cost function (p. 128)
Universal approximation theorem (p. 130)
ReLU activation function (p. 132)
Leaky ReLU activation function (p. 132)
Hyperbolic tangent activation function (p. 132)
Learning rate (p. 134)
Gradient descent algorithm (p. 134)
Backpropagation (p. 139)
L1 regularization (p. 140)
L2 regularization (p. 140)
Epoch (p. 140)
Mini-batch stochastic gradient descent (p. 140)
Gradient descent with momentum (p. 140)
Gradient descent with adaptive learning rate (p. 140)
Learning rate decay (p. 141)
Gradient descent with dropouts (p. 141)
Adam (p. 141)
Stopping rule (p. 141)
Implied volatility (p. 148)
Moneyness (p. 148)
Delta (p. 148)
Volatility surface (p. 148)
Autoencoders (p. 155)
Latent variables (p. 156)
Encoder (p. 157)
Decoder (p. 157)
Variational autoencoders (p. 160)
Kullback-Leibler divergence (p. 161)
Generative adversarial networks (p. 161)
Recurrent neural network (RNN) (p. 163)
Long short-term memory (LSTM) (p. 165)
Convolutional neural networks (CNN) (p. 165)
Feature map (p. 165)
Receptive field (p. 166)
Filter (p. 167)
Pooling (p. 167)
Flattening (p. 167)
Stride (p. 167)
Padding (p. 167)
Temporal convolutional network (TCN) (p. 168)
Reinforcement learning (p. 171)
Rewards (p. 171)
Exploitation choice (p. 172)
Exploration choice (p. 172)
Greedy action (p. 172)
Non-greedy action (p. 172)
Temporal difference learning (p. 182)
n-step bootstrapping (p. 185)
Deep reinforcement learning or deep Q-learning (p. 186)

Learning Objectives

Demonstrate proficiency in the following areas:

5.1.1 ANNs and Activation Functions (Ch. 6)

For example:

A. Describe an artificial neural network (ANN) with a single hidden layer.

B. Explain the downside of having a linear (or identity) activation function.
C. Calculate the value of a sigmoid function from weights and bias.
D. Explain the reason for using the linear activation function for numerical output values.
E. Explain when using the sigmoid function for the output layer is appropriate.
F. Calculate the number of parameters to be estimated for an ANN with a single hidden layer.
G. Recognize the cost function for an ANN.
H. Calculate the output of ReLU, leaky ReLU, and hyperbolic tangent activation functions.
I. Identify the shapes of sigmoid, ReLU, leaky ReLU, and hyperbolic tangent activation functions.

5.1.2 Gradient Descent Algorithm (Ch 6)

For example:

A. Calculate the change in the value of a function using the learning rate.
B. Explain the reason for requiring a good value for the learning rate.
C. Explain the reason for scaling all variables before using them in the gradient descent algorithm.
D. Calculate the gradient of a function with multiple features.
E. Calculate the relationship between a function and its scaled version.
F. Explain the reason for using backpropagation.
G. Describe how the partial derivative of an objective function can be calculated using backpropagation.
H. Describe the usage of L1 and L2 regularization in the objective function of neural networks.
I. Analyze the effect of L1 and L2 regularization in the objective function of neural networks.
J. Describe mini-batch stochastic gradient descent.
K. Explain the way Adam selects the learning rate.
L. Analyze the relationship between the learning rate and the different iteration stages for a neural network.
M. Explain the adjustment required when gradient descent with dropouts is used.
N. Explain the reason for not minimizing the cost function with many parameters for the training set.
O. Describe the most commonly used stopping rule.

5.1.3 Volatility Modeling (Ch. 6)

For example:

A. Describe the advantage of neural networks over Monte Carlo simulation.

B. List the advantages and disadvantages of using neural networks for derivative pricing.

C. Explain the reason for observing many variations in the pattern of implied volatility.

D. List the reasons for the need to understand movement in the volatility surface.

5.1.4 Applications of Neural Networks (Ch 7)

For example:

A. Describe the objective of an autoencoder.

B. Explain how the number of neurons in the hidden layer is determined in an autoencoder.

C. Recognize the objective function of an autoencoder.

D. List the advantages of PCA and autoencoders.

E. Describe variational autoencoder (VAE) and its key objective.

F. Contrast an autoencoder with a VAE.

G. Describe the two components of the objective function of a VAE.

H. Analyze the effect of the single hyperparameter on the twin objective of a VAE.

I. Describe a generative adversarial network (GAN) and the two types of networks.

J. Describe the objective of a GAN.

K. Explain what happens to the maximum likelihood function when the GAN gets everything correct and when it does not get everything correct.

L. Describe the key difference between a recurrent neural network (RNN) and an ANN.

M. List applications of RNN.

N. Describe the relationship between RNN and an exponentially weighted moving average.

O. Explain the key problem of RNN that is overcome by the Long Short-term Memory (LSTM) network.

P. Describe the key difference between an ANN and a convolutional neural network (CNN).

Q. Analyze the effect of applying a filter to a feature map.

R. List the key advantages of the architecture used in a CNN.

S. Describe how stride reduces the size of feature maps.

T. Explain the reason for using padding.

U. Describe the temporal convolutional network (TCN).

5.1.5 Reinforcement Learning (Ch. 8)

For example:

A. Describe the objective of a reinforcement learning algorithm.

B. Analyze the relationship between the probability of exploration and the number of trials.

C. Calculate the probability of exploration using a decay factor.

D. List the two quantities that are needed for updating expected rewards.

E. Explain what happens when a low number is chosen for the decay factor in the multi-arms bandit problem.

F. Recognize the objective function having discount factor for reinforcement learning with changing environment.

G. Describe the characteristics of a state in reinforcement learning with changing environment.

H. Explain the reason for assigning more weights to later trials for reinforcement learning with changing environments.

I. Describe the key concept of dynamic programming.

J. Calculate the updated values of reward using temporal difference updating.

K. Identify the reason for using an artificial neural network with reinforcement learning.

L. Explain the process of converting the Q-values as the probability of winning.

M. List applications of reinforcement learning.

N. List the problems faced in using reinforcement learning to treat a patient.

O. Explain the application of reinforcement learning to portfolio management and hedging a derivatives portfolio.

P. Describe how reinforcement learning can be used when limited data is available.

Topic 6. Performance Evaluation, Back-Testing, and False Discoveries

Reading 6.1 Provost, F. and T. Fawcett (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media Inc., 1st Edition. Chapters 7 and 8.

Keywords

Accuracy (p. 189)
Confusion matrix (p. 189)
False positive (p. 190)
False negative (p. 190)
True positive (p. 200)
True negative (p. 200)
Class prior (p. 201)
Precision (p.203)
Recall (p.203)
F-measure (p. 204)
Sensitivity (p. 204)
Specificity (p. 204)
Majority Classifier (p. 205)
Ranking classifier (p.210)
Profit curve (p. 212)
ROC graph (p. 215)
Hit rate (p. 216)
False alarm rate (p. 216)
Conservative classifier (p. 216)
Permissive classifier (p. 217)
AUC (p. 219)
Lift curve (p. 219)
Cumulative response curve (p. 219)

Learning Objectives

Demonstrate proficiency in the following areas:

6.1.1 Describing and Evaluating Classifiers (Ch. 7)

For example:

A. Calculate accuracy and error rate.
B. Identify false positives and false negatives within a confusion matrix.
C. Describe unbalanced data and the problems associated with unbalanced data.
D. Calculate accuracy of a model developed using a balanced dataset but applied to an unbalanced dataset.
E. Discuss the problems with unequal costs and benefits of errors.

6.1.2 Describing a Key Analytical Framework and Calculating Expected Values (Ch. 7)

For example:

A. Calculate expected value and expected benefit.
B. Describe how expected value can be used to frame classifier use.
C. Calculate the minimum probability of response for which a customer should be targeted.
D. Describe how expected value can be used to frame classifier evaluation.
E. Calculate expected profit for a classifier with and without using priors.
F. Describe the two common pitfalls to formulating cost-benefit analysis.

G. Calculate true positive, false positive, true negative, and false negative rates for a confusion matrix.
H. Calculate and interpret precision and recall.
I. Calculate the value of the F-measure.
J. Calculate specificity and sensitivity.
K. Describe the reasons for the need to have a baseline model.

6.1.3 Visualizing Model Performance (Ch. 8)

For example:

A. Describe how thresholding can create different confusion matrices.
B. Calculate a confusion matrix using a threshold.
C. List the variables used on both axes of a profit curve.
D. Describe the properties of a profit curve.
E. Calculate points on a profit curve.
F. Calculate the proportion of sample data that can be targeted when a fixed budget is available.
G. List the two critical conditions that must be met for using the profit curve.
H. Describe the ROC graph, including the variables used on the x-axis and the y-axis.
I. Calculate points on a ROC graph using data from a confusion matrix.
J. Describe the four corners and the diagonal of the ROC graph.
K. Analyze the behavior of a random classifier on the ROC graph.
L. Describe how to use the ROC space to evaluate classifiers.
M. Describe a key advantage of using the ROC graph.
N. Explain the equivalence between the AUC and the Gini Index.
O. List the variables used on the x-axis and the y-axis for the cumulative response curve.
P. Explain the equivalence between the lift curve and the cumulative response curve.
Q. Describe the key assumption in creating the lift curve or the cumulative response curve.
R. Calculate points on a cumulative response curve.

Reading 6.2 John C. Hull (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapter 10.

Keywords

Model interpretability (p. 214)
White boxes (p. 215)
Black boxes (p. 215)
Partial dependence plot (p. 223)
Shapley values (p. 223)
Local interpretable model-agnostic explanations (LIME) (p. 226)

Learning Objectives

Demonstrate proficiency in the following areas:

6.2.1 Model Interpretability

For example:

A. Explain the reason for the need to understand how predictions are made.

B. List examples of black boxes and white boxes.

C. Interpret the value of weights in linear regression.

D. Interpret the value of bias in a linear regression when the features are measured as the difference from their means.

E. Calculate confidence limits for sensitivities using the t-statistic.

F. Explain the impact of a particular feature when the difference from the mean of the feature is used in a linear regression.

G. List an important reason for using regularization.

H. Calculate the combined impact of all features in a linear regression when the difference from the mean is used as features.

I. Calculate the probability of a positive and negative outcome for logistic regression.

J. Calculate the probability of an increase in positive outcomes in a logistic regression for small changes in the value of a continuous or categorical feature.

K. Calculate the odds against a given probability.

L. Calculate probabilities from odds on or odds against.

M. List the steps used in creating an expected conditional prediction to understand the role of a particular feature in the prediction.

N. Describe the shape of a partial dependence plot for the linear regression.

O. Explain the difficulty in measuring the combined effect of all features for a non-linear model.

P. Calculate the contribution of features using Shapley values.

Q. List the properties illustrated by the use of Shapley values.

R. List the limitations of Shapley values.

S. List the steps used in LIME.

Reading 6.3 Colquhoun, D. (2014). An Investigation of the False Discovery Rate and the Misinterpretation of p-values. Royal Society Open Science, London, U.K.: Royal Society Open Science.

Keywords

Positive predictive power (p.2) *Inflation effect (p. 9)*

Learning Objectives

Demonstrate proficiency in the following areas:

6.3.1 An Investigation of the False Discovery Rate and the Misinterpretation of p-values

For example:

A. Describe the false discovery rate with the help of a tree diagram.

B. Calculate the probability of real effect given a result is significant.

C. Calculate the false discovery rate.

D. Describe an underpowered study.

E. Describe the inflation effect in the context of false discovery.

F. Describe what happens when we consider p=0.05 rather than p<=0.05.

G. Describe Berger's approach.

H. Calculate the false discovery rate using conditional probabilities.

I. Calculate the conditional probability of the real effect.

J. Calculate the odds ratio using the Bayes approach.

Topic 7. Text Mining

Reading 7.1 Provost, F. and T. Fawcett (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media Inc., 1st Edition. Chapter 10.

Keywords

Linguistic structure (p. 250)
Dirty (p. 250)
Document (p. 251)
Corpus (p. 251)
Tokens (p. 251)
Terms (p. 251)
Bag of words (p. 252)
Term frequency (p. 252)
Inverse document frequency (p. 254)
TFIDF (p. 256)
N-grams (p. 263)
Bi-grams (p. 263)
Named entity extraction (p. 264)
Topic models (p. 264)
Latent information model (p. 266)
Information triage (p. 274)

Learning Objectives

Demonstrate proficiency in the following areas:

7.1.1 Broad Issues Involved in Mining Text

For example:

A. Explain why text is "dirty," which makes mining text difficult.

7.1.2 Text Representation

For example:

A. Describe the meaning of "terms" (or "tokens") when used in information retrieval.
B. List the steps used in converting a document to a term frequency representation.
C. Calculate term frequency (TF), inverse document frequency (IDF), and term frequency inverse document frequency (TFIDF).
D. Describe the treatment for rare and common words when deciding the weight of a term.
E. Identify the general shape of IDF when plotted against the number of documents containing the term.
F. Describe the relationship between a corpus and IDF.
G. Describe the relationship between a document and TFIDF.
H. List the drawbacks of the "bag of words" approach.
I. Calculate IDF using the probability of a term in a set of documents.
J. Calculate the entropy of a term using IDF.

7.1.3 Additional Text Representation Approaches Beyond "Bag of Words"

For example:

A. Explain the term "bag of n-grams up to three."

B. Describe when n-gram sequences would be more useful than their component words.

C. List the main disadvantage of n-gram sequences.

D. Describe key requirements for using the named entity extraction.

E. Contrast topic models with the "bag of words" approach.

F. Describe the process used to learn about topics in topic models.

G. Compare the topic model to the latent information model.

7.1.4 Mining News Stories to Predict Stock Price Movement

For example:

A. Describe how a given task, such as recommending a news story that is likely to result in a significant change in a stock's price, must be formulated into a problem with simplifying assumptions.

B. Describe the required considerations for data preprocessing.

C. Identify and discuss appropriate methods for analyzing the results.

Reading 7.2 John C. Hull (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapter 9.

Keywords

Sentiment analysis (p. 196)
Web scraping (p. 197)
Tokenization (p. 199)
Stop words (p. 199)
Stemming (p. 199)
Lemmatization (p. 200)
Laplace smoothing (p. 205)
Word vectors (p. 209)
Word embedding (p. 209)

Learning Objectives

Demonstrate proficiency in the following areas:

7.2.1 Natural Language Processing (NLP)

For example:

A. List the reasons that make it difficult to develop NLP applications.

B. List applications of NLP.

C. Explain why one should not rush into developing a trading strategy based on NLP.

D. Describe the best approach to creating labeled data for NLP.

E. Describe the steps used in tokenization.

F. Describe a common approach to creating a list of stop words.
G. Contrast stemming from lemmatization.
H. Describe the treatment for rarely occurring words and abbreviations during pre-processing.
I. Describe how a bag-of-words approach can be used to convert a sentence to a numerical array.
J. Identify the drawbacks of the bag-of-words approach.
K. Calculate the number of n-grams that can be created from a sentence.
L. Discuss the key assumption made in using the Naïve Bayes classifier.
M. Calculate the conditional probability of a document having a particular sentiment.
N. Explain the key drawback of the Naïve Bayes classifier.
O. Calculate the conditional probability of a document having a particular label using Laplace smoothing.
P. List the advantages of the logistic regression and SVM over the Naïve Bayes classifier
Q. List applications of word sequences.
R. List some of the algorithms used in translating from one language to another.

Reading 7.3 Zhao, F. (2017). Natural Language Processing – Part I: Primer. S&P Global: Market Intelligence.

Keywords

Natural language processing (NLP) (p. 2)
Structured data (p. 3)
Unstructured data (p. 3)
Deep learning (p. 3)
Noise removal (p. 4)
Lexicon normalization (p. 4)
Object standardization (p. 4)
Dependency grammar (p. 4)
Part of speech tagging (p. 4)
Statistical feature (p. 5)
Gunning Fog Index (p. 10)

Learning Objectives

Demonstrate proficiency in the following areas:

7.3.1 Definitions and Key Concepts

For example:

A. Explain the main difference between machine learning and deep learning.
B. List the major steps used in NLP.
C. Describe the noise removal process for text data.
D. List the types of analysis used in syntactical parsing.
E. Describe triplet relation.
F. Explain the key idea of syntactical parsing.

G. List examples of statistical features used in NLP.

H. Explain what is indicated by the numerical values in word embedding.

I. Explain why NLP is important.

7.3.2 Usage of NLP

For example:

A. List and describe the attributes that have made the dictionary of Loughran and McDonald (2011) such a useful tool for financial research.

B. Explain the difficulty in gaming the dictionary of Loughran and McDonald (2011).

C. List examples of positive and negative words.

D. Define the sentiment of an earnings call.

E. Describe the process of creating industry-level sentiment.

F. Describe the way industry sentiment can be used in the investment process.

G. List the components of the Gunning Fog Index.

H. Explain using the Gunning Fog Index when earnings news is bad.

I. Describe how answers to questions from analysts vary during earnings calls when earnings are good and when earnings are bad.

J. Describe the empirical relationship between language complexity and analyst selectivity for an earnings call.

K. Describe the empirical relationship between analyst selectivity ratio and future return.

Topic 8. Ethical and Privacy Issues

Reading 8.1 John C. Hull (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapter 11.

Keywords

Global Data Protection Regulation (GDPR) (p. 230)
Trolley problem (p. 232)
Adversarial machine learning (p. 233)
Spoofing (p. 233)
Four industrial revolutions (p. 235)

Learning Objectives

Demonstrate proficiency in the following areas:

8.1.1 Data Privacy

For example:

A. Discuss the Global Data Protection Regulation (GDPR) and list its requirements.

B. List the consequences of violating the GDPR.

8.1.2 Biases

For example:

A. Discuss biases, including representativeness and data availability.

B. Discuss how biases can arise from cleaning data, which models are used, and how models are interpreted.

C. Discuss what constitutes informed consent.

8.1.3 Ethics

For example:

A. Discuss whether machine learning models and their applications, such as warfare, can be ethical or unethical.

B. Explain the trolley problem and how it applies to algorithms used for driverless cars.

C. Explain Microsoft's "Thinking About You" and how decisions in the model building can lead to unexpected results.

8.1.4 Adversarial Machine Learning

For example:

A. List an example of adversarial machine learning.
B. List approaches to limiting adversarial machine learning.

8.1.5 Legal Issues

For example:

A. List the potential legal liabilities of algorithms, including ownership and use of data, biased algorithms, and assignment of liability for actions of autonomous systems.

8.1.6 Man vs. Machine

For example:

A. Discuss the four industrial revolutions, including concerns and benefits, and the implications for job markets.

B. Discuss the skill of monitoring machine learning algorithms.

Reading 8.2 Institute of International Finance (May 2019). Machine Learning Thematic Series Part II: Bias and Ethical Implications.

Keywords

Statistical definition of bias (p. 7)
Social definition of bias (p. 7)
Disparate treatment (p. 7)
Disparate impact (p. 7)
Fairness (p. 8)
Dataset bias (p. 10)
Bias by omission (p. 11)
Association bias (p. 11)
Direct or encoded bias (p. 11)
Cleaning and transformation bias (p. 12)
Interaction bias (p. 13)
Automation bias (p. 13)
European Global Data Protection Regulation (GDPR) (p. 13)
Data minimization (p. 14)
Oversampling (p. 20)
Accuracy equity (p. 20)
Conditional accuracy equity (p. 20)
Disparate mistreatment (p. 20)
A posteriori bias correction (p. 21)
Challenger model (p. 21)
Blacklist (p. 21)
FATE and FEAT (Fairness, Accountability, Transparency and Ethics) (p. 22)

Learning Objectives

Demonstrate proficiency in the following areas:

8.2.1 Bias and Ethics in Machine Learning

For example:

A. Discuss and contrast the statistical and social definitions of bias.

B. Explain the trade-off between variance and bias in models.

C. Define ethics.

8.2.2 Types of Bias in Machine Learning

For example:

A. Explain how machine learning methods can increase or decrease model discrimination and prejudices.

B. List and explain types of biases, including dataset bias, association bias, cleaning and transformation bias, interaction bias, and automation bias.

8.2.3 Data Protection Laws and Biases

For example:

A. Explain GDPR.

B. Explain when protected data should and should not be used in modeling processes.

8.2.4 Approaches for Ensuring Fairness

For example:

A. List and explain the approaches for ensuring fairness, including conceptual soundness, data use, data governance, modeling, outcome analysis and controls, and tuning and monitoring.

B. Explain the approaches of financial institutions, such as tracking the lifecycle of a training dataset, suitability assessment of customer profiles, oversampling, variable inclusion, improving accuracy, demographics, averaging results, challenger models, blacklisting, post-processing calibration, and using sensitive attributes.

8.2.5 Development of High-Level Principles

For example:

A. List and explain how fairness, accountability, transparency, and ethics are used as high-level principles in Singapore and the EU, including the four ethical principles and the seven voluntary requirements.

Reading 8.3 Das S., M. Donini, J. Gelman, K. Haas, M. Hardt, J. Katzman, K. Kenthapadi, P. Larroy, P. Yilmaz, and M. B. Zafar (2021). Fairness Measures for Machine Learning in Finance. The Journal of Financial Data Science, 3(4): 33-64. Only pages 33-50 from this reading will be used for the FDP exam.

Keywords

Protected characteristic (p. 34)
Attribute of interest (p. 34)
Biased labels (p. 36)
Biased features (p. 37)
Objective function bias (p. 37)
Homogenization bias (p. 37)
Active bias (p. 37)
Unanticipated machine decisions (p. 37)
Class imbalance (p. 39)
Class imbalance (p. 39)
Conditional demographic disparity in labels (CDDL) (p. 40)
Difference in conditional acceptance (DCA) (p. 42)
Difference in conditional rejection (DCR) (p. 42)
Difference in acceptance rates (DAR) (p. 43)
Difference in rejection rates (DRR) (p. 43)
Matched sample (p. 44)
Unintentional discrimination (p. 47)
Unintentional discrimination (p. 47)

Learning Objectives

Demonstrate proficiency in the following areas:

8.3.1 Algorithmic Biases and Finance

For example:

A. List the three broad approaches to fairness-aware machine learning (FAML).

B. Explain the practical challenges in FAML, including where bias appears in models and how metrics of fairness may conflict with other metrics.

8.3.2 Bias Metrics

For example:

A. List and explain the six categories of bias.

B. Recognize and explain class imbalance and conditional demographic disparity in labels (CDDL).

C. Recognize and explain the difference in conditional acceptance (DCA) and difference in conditional rejection (DCR).

D. Recognize and explain the difference in acceptance rates (DAR) and the difference in rejection rates (DRR).

8.3.3 Bias Mitigation

For example:

A. List and explain four methods of bias correction and mitigation.

Topic 9. Fintech Applications

Reading 9.1 Ekster, G. and Kolm, P. N. (2021). Alternative Data in Investment Management: Usage, Challenges, and Valuation. The Journal of Financial Data Science, 3(4): 10-32.

Keywords

Alternative data (Alt-data) (p. 11)
Originators (p. 12)
Intermediaries (p. 12)
Alpha decay (p. 13)
Entity mapping (p. 14)
Ticker tagging (p. 14)
Panel (p. 15)
Unbalanced panel (p. 15)
Balanced panel (p. 15)
Panel stabilization (p. 15)
Debiasing (p. 16)
Golden triangle event study methodology (p. 17)
Public information test (p. 17)
Market reaction test (p. 17)
Report card (p. 18)
Leave-one-out (LOO) cross-validation (p. 23)

Learning Objectives

Demonstrate proficiency in the following areas:

9.1.1 Background

For example:

A. List examples of alt-datasets.

9.1.2 The Alternative Data Ecosystem

For example:

A. List and discuss the constituents in the alt-data ecosystem, including originators, intermediaries, and investment professionals.

B. List the mistakes and critical steps in preparing and cleaning data.

C. Explain alpha decay and the types of data that have less alpha potential.

D. Compare the use of alt-data in fundamental funds vs. quantitative funds.

9.1.3 Challenges With Alternative Data

For example:

A. Explain entity mapping, ticker tagging, panel stabilization, and debiasing.

9.1.4 The Value of Alternative Data

For example:

A. List the two fundamental methods of evaluating alt-datasets.

B. List and discuss the three steps of the golden triangle event study methodology.

C. Explain using a report card to determine the value of an alternative dataset.

D. Explain the relationship between a dataset's structure and investment performance.

9.1.5 Issues in Preprocessing Data

For example:

A. Explain entity tagging, outlier detection and resolution, panel stabilization and imputation, and imputation error estimation.

9.1.6 Trends in the Alternative Data Space

For example:

A. Discuss cost-benefit analysis of intermediaries vs. originators.

Reading 9.2 OECD (2021). Artificial Intelligence, Machine Learning and Big Data in Finance: Opportunities, Challenges, and Implications for Policy Makers.

Keywords

Governance/accountability (p. 8)
Non-financial risks (p. 8)
Explainability (p. 8)
Robustness/resilience (p. 8)
AI systems (p. 16)
AI subsets (p. 17)
The four V's (p. 18)
Volume (p. 18)
Velocity (p. 18)
Variety (p. 18)
Veracity (p. 18)
Regtech/Suptech (p. 20)
Algo wheel (p. 26)
Thin files (p. 30)
Smart contracts (p. 34)

Learning Objectives

Demonstrate proficiency in the following areas:

9.2.1 Artificial Intelligence (AI) in Finance

For example:

A. Describe the two avenues through which the deployment of AI in finance is expected to drive competitive advantages for financial firms.

B. Describe the primary issues and risks stemming from the deployment of AI in finance.

C. Describe the primary impacts of AI on business models and activities in the financial sector.

D. Describe the main features of the AI system.

E. List and describe the four V's of big data.

F. Describe AI in regulatory and supervisory technology.

G. List back-office applications of AI in financial markets.

H. List middle-office applications of AI in financial markets.

I. List front-office applications of AI in financial markets.

J. Describe the potential risks or benefits stemming from many asset managers using the same AI models.

K. Describe the primary difference between AI-managed trading and systematic trading.

L. Describe the unintended consequences and risks of deploying AI systems in the financial sector.

M. Describe the potential benefits and risks associated with credit intermediation and assessment of creditworthiness using AI systems.

N. Describe the potential benefits of integrating AI systems with blockchain technology, including augmenting the capabilities of smart contracts.

O. Describe the potential risks related to the representativeness and relevance of big data.

P. Describe the potential risks related to privacy and confidentiality related to big data.

Q. Describe risks of bias and discrimination as they relate to using big data.

R. Explain the problems and risks that arise from the lack of the explainability of AI/ML models deployed in the financial sector.

S. Explain the importance of governance of AI systems and accountability when AI systems are deployed in the financial sector.

T. Explain risks that could arise from outsourcing AI techniques to third parties.

Reading 9.3 Financial Stability Board. (2017). Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications.

Keywords

Sentiment indicators (p. 10)
Fraud detection (p. 11)
RegTech (p. 11)
Trading signals (p. 11)
InsurTech (p. 13)
Chatbots (p. 14)
Know your customer (KYC) (p. 20)
SupTech (p. 21)
Auditability (p. 33)
Fintech (p. 35)
Robo-advisors (p.35)
Tonality analysis (p.36)

Learning Objectives

Demonstrate proficiency in the following areas:

9.3.1 Regulatory and Supervisory Issues Around FinTech

For example:

A. Identify factors that may contribute to increases in third party dependencies among financial institutions.

B. Explain why unexpected forms of interconnectedness among institutions could be created.

C. Explain why new forms of macro-level risks could emerge.

D. Explain why new risk management tools and techniques may be required.

9.3.2 Relationships Among Artificial Intelligence, Machine Learning, Big Data, and Algorithms

For example:

A. Describe the two recent developments that have contributed to increased interest in AI.

B. List factors contributing to making the markets more efficient.

C. Describe the relationship between AI, machine learning, and the three algorithms in Figure 1.

9.3.3 Categories of Machine Learning Algorithms

For example:

A. Define four categories of machine learning algorithms based on the degree of human intervention.

B. Describe the role of machine learning algorithms in determining causality vs. correlation.

C. Define augmented intelligence.

D. Explain the limitations of machine learning algorithms in determining causality and correlations.

9.3.4 Drivers of the Growth in the Use of Fintech and Adoption of Artificial Intelligence

For example:

A. Discuss the supply factors related to advances in computing technologies and changes in the financial sector.

B. Discuss the demand factors related to the search for higher profits, increased competition, and changes in the regulatory environment.

9.3.5 Use Cases of Artificial Intelligence and Machine Learning in the Financial Sector

For example:

A. Describe customer-focused uses, such as credit scoring, insurance, and client-facing chatbots.

B. Describe operations-focused uses, such as optimal capital allocation, risk management modeling, and market impact analysis.

C. Describe portfolio management and trading uses.

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

9.3.6 The Micro-Financial Analysis of Artificial Intelligence and Machine Learning Uses

For example:

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

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

C. Describe financial institutions' uses of artificial intelligence and machine learning and their potential impacts on customers and investors.

9.3.7 The Macro-Financial Analysis of Uses of Artificial Intelligence and Machine Learning

For example:

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

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

C. Describe how financial institutions' uses of artificial intelligence and machine learning could be sources of greater instability and vulnerability in financial markets.

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

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

9.3.8 Define the Terms Listed in the Glossary

For example:

A. 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 modeling, and Unsupervised learning.

Reading 9.4 Zappa, D., M. Borrelli, G.P. Clemente, and N. Savelli. (2019) Text Mining in Insurance: From Unstructured Data to Meaning.

Keywords

Text mining (p. 1)
Document term matrix (p. 8)
Term document matrix (p. 8)
Continuous bag of words (p. 13)
Part-of-speech tagging (p. 17)

Learning Objectives

Demonstrate proficiency in the following areas:

9.4.1 Text Mining and Its Applications in the Insurance Industry

For example:

A. Describe an example of a competitive advantage that an insurance company can gain through text mining.

B. Describe and apply N-grams to analyze a document.

C. Describe and apply the chain rule to calculate the joint probability of words in a text data document.

D. Describe and apply the chain rule of probability calculation when applied to an N-gram language model.

E. Describe the tokenization process when applied to a document.

F. Describe the use of stop-words in reducing the number of features of a document.

G. Describe the stemming pre-processing approach when applied to a document.

H. Describe the lemmatization pre-processing approach when applied to a document.

I. Describe, interpret, and apply term frequency and inverse term frequency algorithms. (***Note:*** *equation (2.4) of this reading is different from the one used in LO 7.1.2. Any FDP exam question on this topic will use the formula referenced in LO 7.1.2.)*

J. Explain the objective of the simplest version of a continuous bag of words algorithm.

K. Describe the part-of-speech tagging pre-processing approach when applied to a document.

Reading 9.5 Åstebro, T. (2021). An Inside Peek at AI Use in Private Equity. The Journal of Financial Data Science, 3(3): 97-107.

Keywords

Operational efficiency (p. 97)
Decision Support System (DSS) (p. 97)
Spray and pray (p. 105)

Learning Objectives

Demonstrate proficiency in the following areas:

9.5.1 The Motivation for PE Firms to Deploy AI Techniques

For example:

A. List and describe the benefits and operational efficiencies PE firms can gain from implementing AI techniques.

9.5.2 The Approaches PE Firms Take to Deploy AI Techniques

For example:

A. Describe and contrast AI techniques and data sources, such as web scraping, structured databases, and unstructured data.

B. Describe a machine learning model to predict which startups would successfully raise a Series A round of funding.

C. Explain the three reasons why AI-enhanced prediction models have only been adopted recently.

9.5.3 AI's Impact on Deal Making

For example:

A. Discuss why AI has mostly impacted seed and series A funding rounds.

B. List the key sources that are valuable inputs to natural language processing models.

C. Describe the potential for a technological arms race and an eventual industry shakeout.

9.5.4 Jolt Capital's Ninja Decision Support System (DSS)

For example:

A. List and explain the three insights underlying the construction of the Ninja system.

B. List and explain the two machine learning algorithms deployed by Ninja, including their inputs and outputs.

C. Discuss how the interactions of experienced VC partners with Ninja can be implemented as supervised learning and used to help train other investors.

9.5.5 The Future of AI in PE/VC Decision Making

For example:

A. Describe operational efficiency, beating others to the punch, add-on services, changing investment models, and syndication of data, platforms, and algorithms.

Reading 9.6 Li, Y., Z. Simon, and D. Turkington. (2022). Investable and Interpretable Machine Learning for Equities. The Journal of Financial Data Science, 4(1): 54-74.

The material in the appendix is not tested.

Keywords

Model fingerprint (p. 55)

Learning Objectives

Demonstrate proficiency in the following areas:

9.6.1 Training and Testing

For example:

A. Explain the two steps in training the model.

9.6.2 Interpretation with Model Fingerprinting

For example:

A. Explain interpretation with model fingerprints.

B. List and explain the desirable properties of the fingerprint attribution method, including symmetry, dummy, additivity, and completeness.

C. Explain the linear and nonlinear effects of predictors, and interaction effects of pairs of predictors in machine learning models.

9.6.3 Goal Setting

For example:

A. Explain how a model's behavior can be adjusted through changing prediction goals.

B. Explain how investability can be improved by changing the model's prediction horizon.

Reading 9.7 López de Prado, M. (2018). The 10 Reasons Most Machine Learning Funds Fail. The Journal of Portfolio Management, 44 (6): 120-133.

Keywords

Backtesting (p. 122)
Volume clock (p. 123)
Dollar bars (p. 123)
Stationary (p. 123)
Integer differentiation (p. 123)
Fractional differentiation (p. 124)
Triple barrier method (p. 127)
F1-score (p. 128)
Walk-forward approach (p. 129)
Leakage (p. 129)
Deflated Sharpe ratio (p. 132)
Probabilistic Sharpe ratio (p. 132)

Learning Objectives

Demonstrate proficiency in the following areas:

9.7.1 The Most Common Errors Made When Machine Learning Techniques are Applied to Financial Data Sets

For example:

A. Compare and contrast the silo approach in discretionary strategies versus the meta-strategy in machine learning strategies.

B. Compare and contrast repeated backtesting using machine learning versus examining feature importance of a machine learning application results.

C. Describe the two problems with data samples generated using time bars.

D. Describe the advantages of dollar bars over time bars in creating data for machine learning algorithms.

E. Describe the benefit of fractional differentiation in generating stationary series while preserving memory.

F. Explain the triple-barrier method for labeling observed returns.

G. Describe the definitions of precision, recall, and F1-score as features of machine learning algorithms.

H. Explain the role of non-independent identically distributed returns in the failure of k-fold cross-validation in finance.

I. Describe the walk forward (WF) approach to backtesting of trading strategies.

J. Describe the advantages and disadvantages of the walk forward approach.

K. Explain the relationship between the maximum Sharpe ratio obtained from several backtested strategies and the return volatility of those strategies.

L. Describe the concept of the probabilistic Sharpe ratio.

M. List the impacts of nonnormalized Sharpe ratio, length of track record, skewness, and kurtosis on the probabilistic Sharpe ratio.

Reading 9.8 Harvey, C. R. and Y. Liu. (2014). Evaluating Trading Strategies. The Journal of Portfolio Management, 40(5): 108-118.

Keywords

Family-wise error rate (p. 111)
False discovery rate (p. 111)
Holm test (p. 112)
BHY hurdle (p. 112)
Bonferroni test (p. 112)
Type I error (p. 113)
Type II error (p. 113)

Learning Objectives

Demonstrate proficiency in the following areas:

9.8.1 Using Statistical Techniques to Evaluate Trading Strategies in the Presence of Multiple Tests

For example:

A. Describe why standard statistical tools, such as p-values and t-statistics, can lead to false discoveries in the presence of multiple tests.

B. Calculate the t-statistic based on the reported Sharpe ratio for testing a single trading strategy.

C. Describe and apply Bonferroni tests in the context of the family-wise error rate (FWER) approach to adjusting p-values for multiple tests.

D. Describe the Holm method in the context of the false discovery rate (FDR) approach to adjusting p-values for multiple tests.

E. Recognize and apply the Holm function to calculate adjusted p-values.

F. Describe the process of accepting and rejecting tests using the Holm method.

G. Describe the false discovery approach to adjusting p-values in the presence of multiple tests.

H. Recognize and apply the BHY formula to calculate adjusted p-values.

I. Describe the process of accepting and rejecting tests using the BHY method.

J. Explain the relationship between avoiding false discoveries and missing profitable opportunities.

Reading 9.9 Amler, H., L. Eckey, S. Faust, M. Kaiser, P. Sandner, and B. Schlosser. (2021). DeFi-ning DeFi: Challenges & Pathway.

Keywords

Decentralized finance (p. 1)
Financial Lego (p. 1)
Permissionless (p. 2)
Trustless (p. 2)
Self-sovereignty (p. 2)
Flash loans (p. 3)
Decentralized exchanges (DEXes) (p. 3)
Oracles (p. 4)
Prediction markets (p. 4)
Airdrop (p. 4)
Locked up value (p. 4)
Gas price (gwei) (p. 5)
Smart contract vulnerabilities (p. 5)
Infrastructural risk (p. 5)
Interdependence weaknesses (p. 5)
Frontrunning (p. 6)
Atomic or atomicity (p. 6)
Timelocks (p. 6)
Sharding (p. 6)
Layer-2 scaling solutions (p. 7)
Markets in crypto assets (MiCA) (p. 8)
On-ramping and off-ramping (p. 8)

Learning Objectives

Demonstrate proficiency in the following areas:

9.9.1 The Structure of Decentralized Financial System

For example:

A. Describe the characteristics of decentralized financial systems, such as smart contracts, trustless transactions, and composability.

9.9.2 The Advantages of the DeFi Ecosystem

For example:

A. Explain the characteristics of permissionless, trustless, transparent, interconnected, decentrally governed, and enabling self-sovereignty.

9.9.3 The Overview of Decentralized Financial Services

For example:

A. Explain the various applications available in DeFi, including lending platforms, assets, decentralized exchanges, derivative services, payment networks, oracles, and prediction markets.

9.9.4 Decentralized Governance and Economics

For example:

A. List and explain various governance models, including inflationary and deflationary dynamics.

B. Discuss the weakness and counterparty risk in DeFi, especially regarding the Tether (USDT) stablecoin.

C. Explain the growth and factors driving the economic growth of DeFi applications.

9.9.5 Challenges in the DeFi Market

For example:

A. List and discuss the challenges and potential solutions in the DeFi market, including security, limited scalability, oracles, regulation, on- and off-ramping, and privacy.

Reading 9.10 Nadini, M., L. Alessandretti, F. D. Giacinto, M. Martino, L. M. Aiello, and A. Baronchelli. (2021). Mapping the NFT Revolution: Market Trends, Trade Networks and Visual Features. Only pages 1-19 from this reading will be used for the FDP exam.

Keywords

Non-fungible tokens (NFTs) (p. 1)
Network of trades (p. 7)
Network links (p. 8)
Network structure (p. 9)
Network of NFTs (p. 10)

Learning Objectives

Demonstrate proficiency in the following areas:

9.10.1 The NFT Market

For example:

A. Explain the NFT market and list the six categories of NFTs.

B. Discuss the growth of the NFT market and the categories of NFTs with the highest prices and trading volumes.

9.10.2 The Networks of NFT Trades

For example:

A. Explain the properties of network links and network structure.

B. Explain how NFTs are connected to one another.

C. Explain the implications of behaviors on NFT network structure.

9.10.3 Visual Features

For example:

A. Explain the process of clustering visual features of NFTs, including the cosine distance methodology.

9.10.4 Predicting Sales

For example:

A. List the four features the regression model found to be important in predicting the market values of NFTs.

B. Explain how AdaBoost is used to predict the probability of a secondary sale of an NFT.

9.10.5 Data and Methods

For example:

A. List and define the six categories of NFTs, including art, collectibles, games, metaverse, utility, and other.

FDP EDITORIAL STAFF

Hossein Kazemi, Ph.D., CFA, Senior Advisor, CAIA Association
Satya Das, CAIA, CFA, Senior Advisor, Curriculum and Exams, FDP Institute
Kathryn Wilkens, Ph.D., CAIA, Consultant, FDP Institute
Kim Durand, Project & General Operations Manager, FDP Institute

Design: Gabriele Villamena, Sichtwerk, Inc.