1. Record Nr. UNINA9910746976103321

Autore	Emmert-Streib Frank
Titolo	Elements of Data Science, Machine Learning, and Artificial Intelligence Using R / / Frank Emmert-Streib, Salissou Moutari, and Matthias Dehmer
Pubbl/distr/stampa	Cham, Switzerland : , : Springer, , [2023] ©2023
ISBN	3-031-13339-0
Edizione	[First edition.]
Descrizione fisica	1 online resource (582 pages)
Disciplina	060
Soggetti	Artificial intelligence Machine learning R (Computer program language)
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Nota di bibliografia	Includes bibliographical references and index.
Nota di contenuto	Intro -- Preface -- Contents -- 1 Introduction to Learning from Data -- 1.1 What Is Data Science? -- 1.2 Converting Data into Knowledge -- 1.2.1 Big Aims: Big Questions -- 1.2.2 Generating Insights by Visualization -- 1.3 Structure of the Book -- 1.3.1 Part I -- 1.3.2 Part II -- 1.3.3 Part III -- 1.4 Our Motivation for Writing This Book -- 1.5 How to Use This Book -- 1.6 Summary -- Part I General Topics -- 2 General Prediction Models -- 2.1 Introduction -- 2.2 Categorization of Methods -- 2.2.1 Properties of the Data -- 2.2.2 Properties of the Optimization Algorithm -- 2.2.3 Properties of the Model -- 2.2.4 Summary -- 2.3 Overview of Prediction Models -- 2.4 Causal Model versus Predictive Model -- 2.5 Explainable AI -- 2.6 Fundamental Statistical Characteristics of Prediction Models -- 2.6.1 Example -- 2.7 Summary -- 2.8 Exercises -- 3 General Error Measures -- 3.1 Introduction -- 3.2 Motivation -- 3.3 Fundamental Error Measures -- 3.4 Error Measures -- 3.4.1 True-Positive Rate and True-Negative Rate -- 3.4.2 Positive Predictive Value and Negative Predictive Value -- 3.4.3 Accuracy -- 3.4.4 F-Score -- 3.4.5 False Discovery Rate and False Omission Rate -- 3.4.6 False-Negative Rate and False-Positive Rate -- 3.4.7 Matthews Correlation Coefficient -- 3.4.8 Cohen's Kappa -- 3.4.9 Normalized Mutual Information -- 3.4.10 Area Under the Receiver Operator

Characteristic Curve -- 3.5 Evaluation of Outcome -- 3.5.1 Evaluation of an Individual Method -- 3.5.2 Comparing Multiple Binary Decision-Making Methods -- 3.6 Summary -- 3.7 Exercises -- 4 Resampling Methods -- 4.1 Introduction -- 4.2 Resampling Methods for Error Estimation -- 4.2.1 Holdout Set -- 4.2.2 Leave-One-Out CV -- 4.2.3 K-Fold Cross-Validation -- 4.3 Extended Resampling Methods for Error Estimation -- 4.3.1 Repeated Holdout Set -- 4.3.2 Repeated K-Fold CV -- 4.3.3 Stratified K-Fold CV.
4.4 Bootstrap -- 4.4.1 Resampling With versus Resampling Without Replacement -- 4.5 Subsampling -- 4.6 Different Types of Prediction Data Sets -- 4.7 Sampling from a Distribution -- 4.8 Standard Error -- 4.9 Summary -- 4.10 Exercises -- 5 Data -- 5.1 Introduction -- 5.2 Data Types -- 5.2.1 Genomic Data -- 5.2.2 Network Data -- 5.2.3 Text Data -- 5.2.4 Time-to-Event Data -- 5.2.5 Business Data -- 5.3 Summary -- Part II Core Methods -- 6 Statistical Inference -- 6.1 Exploratory Data Analysis and Descriptive Statistics -- 6.1.1 Data Structure -- 6.1.2 Data Preprocessing -- 6.1.3 Summary Statistics and Presentation of Information -- 6.1.4 Measures of Location -- 6.1.4.1 Sample Mean -- 6.1.4.2 Trimmed Sample Mean -- 6.1.4.3 Sample Median -- 6.1.4.4 Quartile -- 6.1.4.5 Percentile -- 6.1.4.6 Mode -- 6.1.4.7 Proportion -- 6.1.5 Measures of Scale -- 6.1.5.1 Sample Variance -- 6.1.5.2 Range -- 6.1.5.3 Interquartile Range -- 6.1.6 Measures of Shape -- 6.1.6.1 Skewness -- 6.1.6.2 Kurtosis -- 6.1.7 Data Transformation -- 6.1.8 Example: Summary of Data and EDA -- 6.2 Sample Estimators -- 6.2.1 Point Estimation -- 6.2.2 Unbiased Estimators -- 6.2.3 Biased Estimators -- 6.2.4 Sufficiency -- 6.3 Bayesian Inference -- 6.3.1 Conjugate Priors -- 6.3.2 Continuous Parameter Estimation -- 6.3.2.1 Example: Continuous Bayesian Inference Using R -- 6.3.3 Discrete Parameter Estimation -- 6.3.4 Bayesian Credible Intervals -- 6.3.5 Prediction -- 6.3.6 Model Selection -- 6.4 Maximum Likelihood Estimation -- 6.4.1 Asymptotic Confidence Intervals for MLE -- 6.4.2 Bootstrap Confidence Intervals for MLE -- 6.4.3 Meaning of Confidence Intervals -- 6.5 Expectation-Maximization Algorithm -- 6.5.1 Example: EM Algorithm -- 6.6 Summary -- 6.7 Exercises -- 7 Clustering -- 7.1 Introduction -- 7.2 What Is Clustering? -- 7.3 Comparison of Data Points -- 7.3.1 Distance Measures.
7.3.2 Similarity Measures -- 7.4 Basic Principle of Clustering Algorithms -- 7.5 Non-hierarchical Clustering Methods -- 7.5.1 K-Means Clustering -- 7.5.2 K-Medoids Clustering -- 7.5.3 Partitioning Around Medoids (PAM) -- 7.6 Hierarchical Clustering -- 7.6.1 Dendrograms -- 7.6.2 Two Types of Dissimilarity Measures -- 7.6.3 Linkage Functions for Agglomerative Clustering -- 7.6.4 Example -- 7.7 Defining Feature Vectors for General Objects -- 7.8 Cluster Validation -- 7.8.1 External Criteria -- 7.8.2 Assessing the Numerical Values of Indices -- 7.8.3 Internal Criteria -- 7.9 Summary -- 7.10 Exercises -- 8 Dimension Reduction -- 8.1 Introduction -- 8.2 Feature Extraction -- 8.2.1 An Overview of PCA -- 8.2.2 Geometrical Interpretation of PCA -- 8.2.3 PCA Procedure -- 8.2.4 Underlying Mathematical Problems in PCA -- 8.2.5 PCA Using Singular Value Decomposition -- 8.2.6 Assessing PCA Results -- 8.2.7 Illustration of PCA Using R -- 8.2.8 Kernel PCA -- 8.2.9 Discussion -- 8.2.10 Non-negative Matrix Factorization -- 8.2.10.1 NNMF Using the Frobenius Norm as Objective Function -- 8.2.10.2 NNMF Using the Generalized Kullback-Leibler Divergence as Objective Function -- 8.2.10.3 Example of NNMF Using R -- 8.3 Feature Selection -- 8.3.1 Filter Methods Using Mutual Information -- 8.4 Summary -- 8.5 Exercises -- 9 Classification -- 9.1 Introduction -- 9.2 What Is Classification? -- 9.3 Common Aspects of Classification Methods -- 9.3.1 Basic Idea of a Classifier --

9.3.2 Training and Test Data -- 9.3.3 Error Measures -- 9.3.3.1 Error Measures for Multi-class Classification -- 9.4 Naive Bayes Classifier -- 9.4.1 Educational Example -- 9.4.2 Example -- 9.5 Linear Discriminant Analysis -- 9.5.1 Extensions -- 9.6 Logistic Regression -- 9.7 k-Nearest Neighbor Classifier -- 9.8 Support Vector Machine -- 9.8.1 Linearly Separable Data -- 9.8.2 Nonlinearly Separable Data.
9.8.3 Nonlinear Support Vector Machines -- 9.8.4 Examples -- 9.9 Decision Tree -- 9.9.1 What Is a Decision Tree? -- 9.9.1.1 Three Principal Steps to Get a Decision Tree -- 9.9.2 Step 1: Growing a Decision Tree -- 9.9.3 Step 2: Assessing the Size of a Decision Tree -- 9.9.3.1 Intuitive Approach -- 9.9.3.2 Formal Approach -- 9.9.4 Step 3: Pruning a Decision Tree -- 9.9.4.1 Alternative Way to Construct Optimal Decision Trees: Stopping Rules -- 9.9.5 Predictions -- 9.10 Summary -- 9.11 Exercises -- 10 Hypothesis Testing -- 10.1 Introduction -- 10.2 What Is Hypothesis Testing? -- 10.3 Key Components of Hypothesis Testing -- 10.3.1 Step 1: Select Test Statistic -- 10.3.2 Step 2: Null Hypothesis H0 and AlternativeHypothesis H1 -- 10.3.3 Step 3: Sampling Distribution -- 10.3.3.1 Examples -- 10.3.4 Step 4: Significance Level -- 10.3.5 Step 5: Evaluate the Test Statistic from Data -- 10.3.6 Step 6: Determine the p-Value -- 10.3.7 Step 7: Make a Decision about the Null Hypothesis -- 10.4 Type 2 Error and Power -- 10.4.1 Connections between Power and Errors -- 10.5 Confidence Intervals -- 10.5.1 Confidence Intervals for a Population Mean with Known Variance -- 10.5.2 Confidence Intervals for a Population Mean with Unknown Variance -- 10.5.3 Bootstrap Confidence Intervals -- 10.6 Important Hypothesis Tests -- 10.6.1 Student's t-Test -- 10.6.1.1 One-Sample t-Test -- 10.6.1.2 Two-Sample t-Test -- 10.6.1.3 Extensions -- 10.6.2 Correlation Tests -- 10.6.3 Hypergeometric Test -- 10.6.3.1 Null Hypothesis and Sampling Distribution -- 10.6.3.2 Examples -- 10.6.4 Finding the Correct Hypothesis Test -- 10.7 Permutation Tests -- 10.8 Understanding versus Applying Hypothesis Tests -- 10.9 Historical Notes and Misinterpretations -- 10.10 Summary -- 10.11 Exercises -- 11 Linear Regression Models -- 11.1 Introduction -- 11.1.1 What Is Linear Regression?.
11.1.2 Motivating Example -- 11.2 Simple Linear Regression -- 11.2.1 Ordinary Least Squares Estimation of Coefficients -- 11.2.2 Variability of the Coefficients -- 11.2.3 Testing the Necessity of Coefficients -- 11.2.4 Assessing the Quality of a Fit -- 11.3 Preprocessing -- 11.4 Multiple Linear Regression -- 11.4.1 Testing the Necessity of Coefficients -- 11.4.2 Assessing the Quality of a Fit -- 11.5 Diagnosing Linear Models -- 11.5.1 Error Assumptions -- 11.5.2 Linearity Assumption of the Model -- 11.5.3 Leverage Points -- 11.5.4 Outliers -- 11.5.5 Collinearity -- 11.5.6 Discussion -- 11.6 Advanced Topics -- 11.6.1 Interactions -- 11.6.2 Nonlinearities -- 11.6.3 Categorical Predictors -- 11.6.4 Generalized Linear Models -- 11.6.4.1 How to Determine Which Family to Use When Fitting a GLM -- 11.6.4.2 Advantages of GLMs over Traditional OLS Regression -- 11.6.4.3 Example: Poisson Regression -- 11.6.4.4 Example: Logistic Regression -- 11.7 Summary -- 11.8 Exercises -- 12 Model Selection -- 12.1 Introduction -- 12.2 Difference Between Model Selection and Model Assessment -- 12.3 General Approach to Model Selection -- 12.4 Model Selection for Multiple Linear Regression Models -- 12.4.1 R2 and Adjusted R2 -- 12.4.2 Mallow's Cp Statistic -- 12.4.3 Akaike's Information Criterion (AIC) and Schwarz's BIC -- 12.4.4 Best Subset Selection -- 12.4.5 Stepwise Selection -- 12.4.5.1 Forward Stepwise Selection -- 12.4.5.2 Backward Stepwise Selection -- 12.5 Model Selection for Generalized Linear Models -- 12.5.1 Negative Binomial

Regression Model -- 12.5.2 Zero-Inflated Poisson Model -- 12.5.3 Quasi-Poisson Model -- 12.5.4 Comparison of GLMs -- 12.6 Model Selection for Bayesian Models -- 12.7 Nonparametric Model Selection for General Models with Resampling -- 12.8 Summary -- 12.9 Exercises -- Part III Advanced Topics -- 13 Regularization -- 13.1 Introduction.
13.2 Preliminaries.