

Ivo D. Dinov

Data Science and Predictive Analytics

Biomedical and Health Applications using R

 Springer

Contents

1	Motivation	1
1.1	DSPA Mission and Objectives	1
1.2	Examples of Driving Motivational Problems and Challenges	2
1.2.1	Alzheimer’s Disease	2
1.2.2	Parkinson’s Disease	2
1.2.3	Drug and Substance Use	3
1.2.4	Amyotrophic Lateral Sclerosis	4
1.2.5	Normal Brain Visualization	4
1.2.6	Neurodegeneration	4
1.2.7	Genetic Forensics: 2013–2016 Ebola Outbreak	5
1.2.8	Next Generation Sequence (NGS) Analysis	6
1.2.9	Neuroimaging-Genetics	7
1.3	Common Characteristics of Big (Biomedical and Health) Data	8
1.4	Data Science	9
1.5	Predictive Analytics	9
1.6	High-Throughput Big Data Analytics	10
1.7	Examples of Data Repositories, Archives, and Services	10
1.8	DSPA Expectations	11
2	Foundations of R	13
2.1	Why Use R?	13
2.2	Getting Started	15
2.2.1	Install Basic Shell-Based R	15
2.2.2	GUI Based R Invocation (RStudio)	15
2.2.3	RStudio GUI Layout	15
2.2.4	Some Notes	16
2.3	Help	16
2.4	Simple Wide-to-Long Data format Translation	17
2.5	Data Generation	18
2.6	Input/Output (I/O)	22

- 2.7 Slicing and Extracting Data 24
- 2.8 Variable Conversion 25
- 2.9 Variable Information 25
- 2.10 Data Selection and Manipulation 27
- 2.11 Math Functions 30
- 2.12 Matrix Operations 32
- 2.13 Advanced Data Processing 32
- 2.14 Strings 37
- 2.15 Plotting 39
- 2.16 QQ Normal Probability Plot 41
- 2.17 Low-Level Plotting Commands 45
- 2.18 Graphics Parameters 45
- 2.19 Optimization and model Fitting 47
- 2.20 Statistics 48
- 2.21 Distributions 49
 - 2.21.1 Programming 49
- 2.22 Data Simulation Primer 50
- 2.23 Appendix 56
 - 2.23.1 HTML SOCR Data Import 56
 - 2.23.2 R Debugging 57
- 2.24 Assignments: 2. R Foundations 60
 - 2.24.1 Confirm that You Have Installed R/RStudio 60
 - 2.24.2 Long-to-Wide Data Format Translation 61
 - 2.24.3 Data Frames 61
 - 2.24.4 Data Stratification 61
 - 2.24.5 Simulation 61
 - 2.24.6 Programming 62
- References 62
- 3 Managing Data in R 63**
 - 3.1 Saving and Loading R Data Structures 63
 - 3.2 Importing and Saving Data from CSV Files 64
 - 3.3 Exploring the Structure of Data 66
 - 3.4 Exploring Numeric Variables 66
 - 3.5 Measuring the Central Tendency: Mean, Median, Mode 67
 - 3.6 Measuring Spread: Quartiles and the Five-Number
Summary 68
 - 3.7 Visualizing Numeric Variables: Boxplots 70
 - 3.8 Visualizing Numeric Variables: Histograms 71
 - 3.9 Understanding Numeric Data: Uniform
and Normal Distributions 72
 - 3.10 Measuring Spread: Variance and Standard Deviation 73
 - 3.11 Exploring Categorical Variables 76

- 3.12 Exploring Relationships Between Variables 77
- 3.13 Missing Data 79
 - 3.13.1 Simulate Some Real Multivariate Data 84
 - 3.13.2 TBI Data Example 98
 - 3.13.3 Imputation via Expectation-Maximization 122
- 3.14 Parsing Webpages and Visualizing Tabular HTML Data 130
- 3.15 Cohort-Rebalancing (for Imbalanced Groups) 135
- 3.16 Appendix 138
 - 3.16.1 Importing Data from SQL Databases 138
 - 3.16.2 R Code Fragments 139
- 3.17 Assignments: 3. Managing Data in R 140
 - 3.17.1 Import, Plot, Summarize and Save Data 140
 - 3.17.2 Explore some Bivariate Relations in the Data 140
 - 3.17.3 Missing Data 141
 - 3.17.4 Surface Plots 141
 - 3.17.5 Unbalanced Designs 141
 - 3.17.6 Aggregate Analysis 141
- References 141
- 4 Data Visualization 143**
 - 4.1 Common Questions 143
 - 4.2 Classification of Visualization Methods 144
 - 4.3 Composition 144
 - 4.3.1 Histograms and Density Plots 144
 - 4.3.2 Pie Chart 147
 - 4.3.3 Heat Map 149
 - 4.4 Comparison 152
 - 4.4.1 Paired Scatter Plots 152
 - 4.4.2 Jitter Plot 157
 - 4.4.3 Bar Plots 159
 - 4.4.4 Trees and Graphs 164
 - 4.4.5 Correlation Plots 167
 - 4.5 Relationships 171
 - 4.5.1 Line Plots Using `ggplot` 171
 - 4.5.2 Density Plots 173
 - 4.5.3 Distributions 173
 - 4.5.4 2D Kernel Density and 3D Surface Plots 174
 - 4.5.5 Multiple 2D Image Surface Plots 176
 - 4.5.6 3D and 4D Visualizations 178
 - 4.6 Appendix 183
 - 4.6.1 Hands-on Activity (Health Behavior Risks) 183
 - 4.6.2 Additional `ggplot` Examples 187

- 4.7 Assignments 4: Data Visualization 198
 - 4.7.1 Common Plots 198
 - 4.7.2 Trees and Graphs 198
 - 4.7.3 Exploratory Data Analytics (EDA) 199
- References 199
- 5 Linear Algebra & Matrix Computing 201**
 - 5.1 Matrices (Second Order Tensors) 202
 - 5.1.1 Create Matrices 202
 - 5.1.2 Adding Columns and Rows 203
 - 5.2 Matrix Subscripts 204
 - 5.3 Matrix Operations 204
 - 5.3.1 Addition 204
 - 5.3.2 Subtraction 205
 - 5.3.3 Multiplication 205
 - 5.3.4 Element-wise Division 207
 - 5.3.5 Transpose 207
 - 5.3.6 Multiplicative Inverse 207
 - 5.4 Matrix Algebra Notation 209
 - 5.4.1 Linear Models 209
 - 5.4.2 Solving Systems of Equations 210
 - 5.4.3 The Identity Matrix 212
 - 5.5 Scalars, Vectors and Matrices 213
 - 5.5.1 Sample Statistics (Mean, Variance) 215
 - 5.5.2 Least Square Estimation 218
 - 5.6 Eigenvalues and Eigenvectors 219
 - 5.7 Other Important Functions 220
 - 5.8 Matrix Notation (Another View) 220
 - 5.9 Multivariate Linear Regression 224
 - 5.10 Sample Covariance Matrix 227
 - 5.11 Assignments: 5. Linear Algebra & Matrix Computing 229
 - 5.11.1 How Is Matrix Multiplication Defined? 229
 - 5.11.2 Scalar Versus Matrix Multiplication 229
 - 5.11.3 Matrix Equations 229
 - 5.11.4 Least Square Estimation 230
 - 5.11.5 Matrix Manipulation 230
 - 5.11.6 Matrix Transpose 230
 - 5.11.7 Sample Statistics 230
 - 5.11.8 Least Square Estimation 230
 - 5.11.9 Eigenvalues and Eigenvectors 231
 - References 231
- 6 Dimensionality Reduction 233**
 - 6.1 Example: Reducing 2D to 1D 233
 - 6.2 Matrix Rotations 237
 - 6.3 Notation 242

- 6.4 Summary (PCA vs. ICA vs. FA) 242
- 6.5 Principal Component Analysis (PCA) 243
 - 6.5.1 Principal Components 243
- 6.6 Independent Component Analysis (ICA) 250
- 6.7 Factor Analysis (FA) 254
- 6.8 Singular Value Decomposition (SVD) 256
- 6.9 SVD Summary 258
- 6.10 Case Study for Dimension Reduction (Parkinson’s Disease) 258
- 6.11 Assignments: 6. Dimensionality Reduction 265
 - 6.11.1 Parkinson’s Disease Example 265
 - 6.11.2 Allometric Relations in Plants Example 266
- References 266
- 7 Lazy Learning: Classification Using Nearest Neighbors 267**
 - 7.1 Motivation 268
 - 7.2 The kNN Algorithm Overview 269
 - 7.2.1 Distance Function and Dummy Coding 269
 - 7.2.2 Ways to Determine k 270
 - 7.2.3 Rescaling of the Features 270
 - 7.2.4 Rescaling Formulas 271
 - 7.3 Case Study 271
 - 7.3.1 Step 1: Collecting Data 271
 - 7.3.2 Step 2: Exploring and Preparing the Data 272
 - 7.3.3 Normalizing Data 273
 - 7.3.4 Data Preparation: Creating Training and Testing Datasets 274
 - 7.3.5 Step 3: Training a Model On the Data 274
 - 7.3.6 Step 4: Evaluating Model Performance 274
 - 7.3.7 Step 5: Improving Model Performance 275
 - 7.3.8 Testing Alternative Values of k 276
 - 7.3.9 Quantitative Assessment (Tables 7.2 and 7.3) 282
 - 7.4 Assignments: 7. Lazy Learning: Classification Using Nearest Neighbors 286
 - 7.4.1 Traumatic Brain Injury (TBI) 286
 - 7.4.2 Parkinson’s Disease 286
 - 7.4.3 KNN Classification in a High Dimensional Space 287
 - 7.4.4 KNN Classification in a Lower Dimensional Space 287
 - References 287
- 8 Probabilistic Learning: Classification Using Naive Bayes 289**
 - 8.1 Overview of the Naive Bayes Algorithm 289
 - 8.2 Assumptions 290
 - 8.3 Bayes Formula 290
 - 8.4 The Laplace Estimator 292

- 8.5 Case Study: Head and Neck Cancer Medication 293
 - 8.5.1 Step 1: Collecting Data 293
 - 8.5.2 Step 2: Exploring and Preparing the Data 293
 - 8.5.3 Step 3: Training a Model on the Data 299
 - 8.5.4 Step 4: Evaluating Model Performance 300
 - 8.5.5 Step 5: Improving Model Performance 301
 - 8.5.6 Step 6: Compare Naive Bayesian against LDA 302
- 8.6 Practice Problem 303
- 8.7 Assignments 8: Probabilistic Learning: Classification
 - Using Naive Bayes 304
 - 8.7.1 Explain These Two Concepts 304
 - 8.7.2 Analyzing Textual Data 305
- References 305
- 9 Decision Tree Divide and Conquer Classification 307**
 - 9.1 Motivation 307
 - 9.2 Hands-on Example: Iris Data 308
 - 9.3 Decision Tree Overview 310
 - 9.3.1 Divide and Conquer 311
 - 9.3.2 Entropy 312
 - 9.3.3 Misclassification Error and Gini Index 313
 - 9.3.4 C5.0 Decision Tree Algorithm 313
 - 9.3.5 Pruning the Decision Tree 315
 - 9.4 Case Study 1: Quality of Life and Chronic Disease 316
 - 9.4.1 Step 1: Collecting Data 316
 - 9.4.2 Step 2: Exploring and Preparing the Data 316
 - 9.4.3 Step 3: Training a Model On the Data 319
 - 9.4.4 Step 4: Evaluating Model Performance 322
 - 9.4.5 Step 5: **Trial** Option 323
 - 9.4.6 Loading the Misclassification Error Matrix 324
 - 9.4.7 Parameter Tuning 325
 - 9.5 Compare Different Impurity Indices 331
 - 9.6 Classification Rules 331
 - 9.6.1 Separate and Conquer 331
 - 9.6.2 The One Rule Algorithm 332
 - 9.6.3 The RIPPER Algorithm 332
 - 9.7 Case Study 2: QoL in Chronic Disease (Take 2) 332
 - 9.7.1 Step 3: Training a Model on the Data 332
 - 9.7.2 Step 4: Evaluating Model Performance 333
 - 9.7.3 Step 5: Alternative Model1 334
 - 9.7.4 Step 5: Alternative Model2 334
 - 9.8 Practice Problem 337

- 9.9 Assignments 9: Decision Tree Divide and Conquer
 - Classification 342
 - 9.9.1 Explain These Concepts 342
 - 9.9.2 Decision Tree Partitioning 342
- References 343
- 10 Forecasting Numeric Data Using Regression Models 345**
 - 10.1 Understanding Regression 345
 - 10.1.1 Simple Linear Regression 345
 - 10.2 Ordinary Least Squares Estimation 347
 - 10.2.1 Model Assumptions 349
 - 10.2.2 Correlations 349
 - 10.2.3 Multiple Linear Regression 350
 - 10.3 Case Study 1: Baseball Players 352
 - 10.3.1 Step 1: Collecting Data 352
 - 10.3.2 Step 2: Exploring and Preparing the Data 352
 - 10.3.3 Exploring Relationships Among Features:
The Correlation Matrix 356
 - 10.3.4 Visualizing Relationships Among Features:
The Scatterplot Matrix 356
 - 10.3.5 Step 3: Training a Model on the Data 358
 - 10.3.6 Step 4: Evaluating Model Performance 359
 - 10.4 Step 5: Improving Model Performance 361
 - 10.4.1 Model Specification: Adding Non-linear
Relationships 369
 - 10.4.2 Transformation: Converting a Numeric Variable
to a Binary Indicator 370
 - 10.4.3 Model Specification: Adding Interaction Effects 371
 - 10.5 Understanding Regression Trees and Model Trees 373
 - 10.5.1 Adding Regression to Trees 373
 - 10.6 Case Study 2: Baseball Players (Take 2) 374
 - 10.6.1 Step 2: Exploring and Preparing the Data 374
 - 10.6.2 Step 3: Training a Model On the Data 375
 - 10.6.3 Visualizing Decision Trees 375
 - 10.6.4 Step 4: Evaluating Model Performance 377
 - 10.6.5 Measuring Performance with Mean Absolute Error 378
 - 10.6.6 Step 5: Improving Model Performance 378
 - 10.7 Practice Problem: Heart Attack Data 380
 - 10.8 Assignments: 10. Forecasting Numeric Data Using
Regression Models 381
 - References 381

11 Black Box Machine-Learning Methods: Neural Networks and Support Vector Machines 383

11.1 Understanding Neural Networks 383

 11.1.1 From Biological to Artificial Neurons 383

 11.1.2 Activation Functions 384

 11.1.3 Network Topology 386

 11.1.4 The Direction of Information Travel 386

 11.1.5 The Number of Nodes in Each Layer 386

 11.1.6 Training Neural Networks with Backpropagation 387

11.2 Case Study 1: Google Trends and the Stock Market: Regression 388

 11.2.1 Step 1: Collecting Data 388

 11.2.2 Step 2: Exploring and Preparing the Data 389

 11.2.3 Step 3: Training a Model on the Data 391

 11.2.4 Step 4: Evaluating Model Performance 392

 11.2.5 Step 5: Improving Model Performance 393

 11.2.6 Step 6: Adding Additional Layers 394

11.3 Simple NN Demo: Learning to Compute $\sqrt{\cdot}$ 394

11.4 Case Study 2: Google Trends and the Stock Market – Classification 396

11.5 Support Vector Machines (SVM) 398

 11.5.1 Classification with Hyperplanes 399

11.6 Case Study 3: Optical Character Recognition (OCR) 403

 11.6.1 Step 1: Prepare and Explore the Data 404

 11.6.2 Step 2: Training an SVM Model 405

 11.6.3 Step 3: Evaluating Model Performance 406

 11.6.4 Step 4: Improving Model Performance 408

11.7 Case Study 4: Iris Flowers 409

 11.7.1 Step 1: Collecting Data 409

 11.7.2 Step 2: Exploring and Preparing the Data 409

 11.7.3 Step 3: Training a Model on the Data 411

 11.7.4 Step 4: Evaluating Model Performance 412

 11.7.5 Step 5: RBF Kernel Function 413

 11.7.6 Parameter Tuning 413

 11.7.7 Improving the Performance of Gaussian Kernels 415

11.8 Practice 416

 11.8.1 Problem 1 Google Trends and the Stock Market 416

 11.8.2 Problem 2: Quality of Life and Chronic Disease 416

11.9 Appendix 420

11.10 Assignments: 11. Black Box Machine-Learning Methods: Neural Networks and Support Vector Machines 421

 11.10.1 Learn and Predict a Power-Function 421

 11.10.2 Pediatric Schizophrenia Study 421

References 422

- 12 Apriori Association Rules Learning** 423
 - 12.1 Association Rules 423
 - 12.2 The Apriori Algorithm for Association Rule Learning 424
 - 12.3 Measuring Rule Importance by Using Support and Confidence 424
 - 12.4 Building a Set of Rules with the Apriori Principle 425
 - 12.5 A Toy Example 426
 - 12.6 Case Study 1: Head and Neck Cancer Medications 427
 - 12.6.1 Step 1: Collecting Data 427
 - 12.6.2 Step 2: Exploring and Preparing the Data 427
 - 12.6.3 Step 3: Training a Model on the Data 432
 - 12.6.4 Step 4: Evaluating Model Performance 433
 - 12.6.5 Step 5: Improving Model Performance 435
 - 12.7 Practice Problems: Groceries 438
 - 12.8 Summary 441
 - 12.9 Assignments: 12. Apriori Association Rules Learning 442
 - References 442
- 13 k-Means Clustering** 443
 - 13.1 Clustering as a Machine Learning Task 443
 - 13.2 Silhouette Plots 446
 - 13.3 The k-Means Clustering Algorithm 447
 - 13.3.1 Using Distance to Assign and Update Clusters 447
 - 13.3.2 Choosing the Appropriate Number of Clusters 448
 - 13.4 Case Study 1: Divorce and Consequences on Young Adults 448
 - 13.4.1 Step 1: Collecting Data 448
 - 13.4.2 Step 2: Exploring and Preparing the Data 449
 - 13.4.3 Step 3: Training a Model on the Data 450
 - 13.4.4 Step 4: Evaluating Model Performance 451
 - 13.4.5 Step 5: Usage of Cluster Information 454
 - 13.5 Model Improvement 455
 - 13.5.1 Tuning the Parameter k 457
 - 13.6 Case Study 2: Pediatric Trauma 459
 - 13.6.1 Step 1: Collecting Data 459
 - 13.6.2 Step 2: Exploring and Preparing the Data 460
 - 13.6.3 Step 3: Training a Model on the Data 461
 - 13.6.4 Step 4: Evaluating Model Performance 462
 - 13.6.5 Practice Problem: Youth Development 465
 - 13.7 Hierarchical Clustering 467
 - 13.8 Gaussian Mixture Models 470
 - 13.9 Summary 472
 - 13.10 Assignments: 13. k-Means Clustering 472
 - References 473

- 14 Model Performance Assessment 475**
 - 14.1 Measuring the Performance of Classification Methods 475
 - Evaluation Strategies 477
 - 14.1.1 Binary Outcomes 477
 - 14.1.2 Confusion Matrices 478
 - 14.1.3 Other Measures of Performance Beyond Accuracy . . . 480
 - 14.1.4 The Kappa (κ) Statistic 481
 - 14.1.5 Computation of Observed Accuracy and Expected Accuracy 484
 - 14.1.6 Sensitivity and Specificity 485
 - 14.1.7 Precision and Recall 486
 - 14.1.8 The F-Measure 487
 - 14.2 Visualizing Performance Tradeoffs (ROC Curve) 488
 - 14.3 Estimating Future Performance (Internal Statistical Validation) 491
 - 14.3.1 The Holdout Method 491
 - 14.3.2 Cross-Validation 492
 - 14.3.3 Bootstrap Sampling 494
 - 14.4 Assignment: 14. Evaluation of Model Performance 495
 - References 496
- 15 Improving Model Performance 497**
 - 15.1 Improving Model Performance by Parameter Tuning 497
 - 15.2 Using `caret` for Automated Parameter Tuning 497
 - 15.2.1 Customizing the Tuning Process 501
 - 15.2.2 Improving Model Performance with Meta-learning . . . 502
 - 15.2.3 Bagging 503
 - 15.2.4 Boosting 505
 - 15.2.5 Random Forests 506
 - 15.2.6 Adaptive Boosting 508
 - 15.3 Assignment: 15. Improving Model Performance 510
 - 15.3.1 Model Improvement Case Study 511
 - References 511
- 16 Specialized Machine Learning Topics 513**
 - 16.1 Working with Specialized Data and Databases 513
 - 16.1.1 Data Format Conversion 514
 - 16.1.2 Querying Data in SQL Databases 515
 - 16.1.3 Real Random Number Generation 521
 - 16.1.4 Downloading the Complete Text of Web Pages 522
 - 16.1.5 Reading and Writing XML with the XML Package 523
 - 16.1.6 Web-Page Data Scraping 524
 - 16.1.7 Parsing JSON from Web APIs 525
 - 16.1.8 Reading and Writing Microsoft Excel Spreadsheets Using XLSX 526

- 16.2 Working with Domain-Specific Data 527
 - 16.2.1 Working with Bioinformatics Data 527
 - 16.2.2 Visualizing Network Data 528
- 16.3 Data Streaming 533
 - 16.3.1 Definition 533
 - 16.3.2 The `stream` Package 534
 - 16.3.3 Synthetic Example: Random Gaussian Stream 534
 - 16.3.4 Sources of Data Streams 536
 - 16.3.5 Printing, Plotting and Saving Streams 537
 - 16.3.6 Stream Animation 538
 - 16.3.7 Case-Study: SOCR Knee Pain Data 540
 - 16.3.8 Data Stream Clustering and Classification (DSC) 542
 - 16.3.9 Evaluation of Data Stream Clustering 545
- 16.4 Optimization and Improving the Computational Performance 546
 - 16.4.1 Generalizing Tabular Data Structures with `dplyr` 547
 - 16.4.2 Making Data Frames Faster with `Data.Table` 548
 - 16.4.3 Creating Disk-Based Data Frames with `ff` 548
 - 16.4.4 Using Massive Matrices with `bigmemory` 549
- 16.5 Parallel Computing 549
 - 16.5.1 Measuring Execution Time 550
 - 16.5.2 Parallel Processing with Multiple Cores 550
 - 16.5.3 Parallelization Using `foreach` and `doParallel` 552
 - 16.5.4 GPU Computing 553
- 16.6 Deploying Optimized Learning Algorithms 553
 - 16.6.1 Building Bigger Regression Models with `biglm` 553
 - 16.6.2 Growing Bigger and Faster Random Forests with `bigrf` 553
 - 16.6.3 Training and Evaluation Models in Parallel with `caret` 554
- 16.7 Practice Problem 554
- 16.8 Assignment: 16. Specialized Machine Learning Topics 555
 - 16.8.1 Working with Website Data 555
 - 16.8.2 Network Data and Visualization 555
 - 16.8.3 Data Conversion and Parallel Computing 555
- References 556
- 17 Variable/Feature Selection 557**
 - 17.1 Feature Selection Methods 557
 - 17.1.1 Filtering Techniques 557
 - 17.1.2 Wrapper Methods 558
 - 17.1.3 Embedded Techniques 558
 - 17.2 Case Study: ALS 559
 - 17.2.1 Step 1: Collecting Data 559
 - 17.2.2 Step 2: Exploring and Preparing the Data 559

17.2.3	Step 3: Training a Model on the Data	560
17.2.4	Step 4: Evaluating Model Performance	564
17.3	Practice Problem	569
17.4	Assignment: 17. Variable/Feature Selection	571
17.4.1	Wrapper Feature Selection	571
17.4.2	Use the PPMI Dataset	571
	References	572
18	Regularized Linear Modeling and Controlled Variable Selection	573
18.1	Questions	574
18.2	Matrix Notation	574
18.3	Regularized Linear Modeling	574
18.3.1	Ridge Regression	576
18.3.2	Least Absolute Shrinkage and Selection Operator (LASSO) Regression	579
18.3.3	Predictor Standardization	582
18.3.4	Estimation Goals	582
18.4	Linear Regression	582
18.4.1	Drawbacks of Linear Regression	583
18.4.2	Assessing Prediction Accuracy	583
18.4.3	Estimating the Prediction Error	583
18.4.4	Improving the Prediction Accuracy	584
18.4.5	Variable Selection	585
18.5	Regularization Framework	586
18.5.1	Role of the Penalty Term	586
18.5.2	Role of the Regularization Parameter	586
18.5.3	LASSO	587
18.5.4	General Regularization Framework	587
18.6	Implementation of Regularization	588
18.6.1	Example: Neuroimaging-Genetics Study of Parkinson's Disease Dataset	588
18.6.2	Computational Complexity	590
18.6.3	LASSO and Ridge Solution Paths	590
18.6.4	Choice of the Regularization Parameter	598
18.6.5	Cross Validation Motivation	599
18.6.6	n -Fold Cross Validation	599
18.6.7	LASSO 10-Fold Cross Validation	600
18.6.8	Stepwise OLS (Ordinary Least Squares)	601
18.6.9	Final Models	602
18.6.10	Model Performance	604
18.6.11	Comparing Selected Features	604
18.6.12	Summary	605
18.7	Knock-off Filtering: Simulated Example	605
18.7.1	Notes	607

- 18.8 PD Neuroimaging-Genetics Case-Study 608
 - 18.8.1 Fetching, Cleaning and Preparing the Data 608
 - 18.8.2 Preparing the Response Vector 609
 - 18.8.3 False Discovery Rate (FDR) 617
 - 18.8.4 Running the Knockoff Filter 620
- 18.9 Assignment: 18. Regularized Linear Modeling and Knockoff Filtering 621
- References 622
- 19 Big Longitudinal Data Analysis 623**
 - 19.1 Time Series Analysis 623
 - 19.1.1 Step 1: Plot Time Series 626
 - 19.1.2 Step 2: Find Proper Parameter Values for ARIMA Model 628
 - 19.1.3 Check the Differencing Parameter 629
 - 19.1.4 Identifying the AR and MA Parameters 630
 - 19.1.5 Step 3: Build an ARIMA Model 632
 - 19.1.6 Step 4: Forecasting with ARIMA Model 637
 - 19.2 Structural Equation Modeling (SEM)-Latent Variables 638
 - 19.2.1 Foundations of SEM 638
 - 19.2.2 SEM Components 641
 - 19.2.3 Case Study – Parkinson’s Disease (PD) 642
 - 19.2.4 Outputs of Lavaan SEM 647
 - 19.3 Longitudinal Data Analysis-Linear Mixed Models 648
 - 19.3.1 Mean Trend 648
 - 19.3.2 Modeling the Correlation 652
 - 19.4 GLMM/GEE Longitudinal Data Analysis 653
 - 19.4.1 GEE Versus GLMM 655
 - 19.5 Assignment: 19. Big Longitudinal Data Analysis 657
 - 19.5.1 Imaging Data 657
 - 19.5.2 Time Series Analysis 658
 - 19.5.3 Latent Variables Model 658
 - References 658
- 20 Natural Language Processing/Text Mining 659**
 - 20.1 A Simple NLP/TM Example 660
 - 20.1.1 Define and Load the Unstructured-Text Documents 661
 - 20.1.2 Create a New VCorpus Object 663
 - 20.1.3 To-Lower Case Transformation 664
 - 20.1.4 Text Pre-processing 664
 - 20.1.5 Bags of Words 666
 - 20.1.6 Document Term Matrix 667

- 20.2 Case-Study: Job Ranking 669
 - 20.2.1 Step 1: Make a VCorpus Object 670
 - 20.2.2 Step 2: Clean the VCorpus Object 670
 - 20.2.3 Step 3: Build the Document Term Matrix 670
 - 20.2.4 Area Under the ROC Curve 674
- 20.3 TF-IDF 676
 - 20.3.1 Term Frequency (TF) 676
 - 20.3.2 Inverse Document Frequency (IDF) 676
 - 20.3.3 TF-IDF 677
- 20.4 Cosine Similarity 685
- 20.5 Sentiment Analysis 686
 - 20.5.1 Data Preprocessing 686
 - 20.5.2 NLP/TM Analytics 689
 - 20.5.3 Prediction Optimization 692
- 20.6 Assignment: 20. Natural Language Processing/Text Mining 694
 - 20.6.1 Mining Twitter Data 694
 - 20.6.2 Mining Cancer Clinical Notes 695
- References 695
- 21 Prediction and Internal Statistical Cross Validation 697**
 - 21.1 Forecasting Types and Assessment Approaches 697
 - 21.2 Overfitting 698
 - 21.2.1 Example (US Presidential Elections) 698
 - 21.2.2 Example (Google Flu Trends) 698
 - 21.2.3 Example (Autism) 700
 - 21.3 Internal Statistical Cross-Validation is an Iterative Process 701
 - 21.4 Example (Linear Regression) 702
 - 21.4.1 Cross-Validation Methods 703
 - 21.4.2 Exhaustive Cross-Validation 703
 - 21.4.3 Non-Exhaustive Cross-Validation 704
 - 21.5 Case-Studies 704
 - 21.5.1 Example 1: Prediction of Parkinson’s Disease
Using Adaptive Boosting (AdaBoost) 705
 - 21.5.2 Example 2: Sleep Dataset 708
 - 21.5.3 Example 3: Model-Based (Linear Regression)
Prediction Using the Attitude Dataset 710
 - 21.5.4 Example 4: Parkinson’s Data (ppmi_data) 711
 - 21.6 Summary of CV output 712
 - 21.7 Alternative Predictor Functions 712
 - 21.7.1 Logistic Regression 713
 - 21.7.2 Quadratic Discriminant Analysis (QDA) 714
 - 21.7.3 Foundation of LDA and QDA for Prediction,
Dimensionality Reduction, and Forecasting 715
 - 21.7.4 Neural Networks 717
 - 21.7.5 SVM 718

21.7.6	k-Nearest Neighbors Algorithm (k-NN)	719
21.7.7	k-Means Clustering (k-MC)	720
21.7.8	Spectral Clustering	727
21.8	Compare the Results	730
21.9	Assignment: 21. Prediction and Internal Statistical Cross-Validation	733
	References	734
22	Function Optimization	735
22.1	Free (Unconstrained) Optimization	735
22.1.1	Example 1: Minimizing a Univariate Function (Inverse-CDF)	736
22.1.2	Example 2: Minimizing a Bivariate Function	738
22.1.3	Example 3: Using Simulated Annealing to Find the Maximum of an Oscillatory Function	739
22.2	Constrained Optimization	740
22.2.1	Equality Constraints	740
22.2.2	Lagrange Multipliers	740
22.2.3	Inequality Constrained Optimization	741
	Quadratic Programming (QP)	747
22.3	General Non-linear Optimization	748
22.3.1	Dual Problem Optimization	749
22.4	Manual Versus Automated Lagrange Multiplier Optimization	753
22.5	Data Denoising	756
22.6	Assignment: 22. Function Optimization	761
22.6.1	Unconstrained Optimization	761
22.6.2	Linear Programming (LP)	761
22.6.3	Mixed Integer Linear Programming (MILP)	762
22.6.4	Quadratic Programming (QP)	762
22.6.5	Complex Non-linear Optimization	762
22.6.6	Data Denoising	763
	References	763
23	Deep Learning, Neural Networks	765
23.1	Deep Learning Training	766
23.1.1	Perceptrons	766
23.2	Biological Relevance	768
23.3	Simple Neural Net Examples	770
23.3.1	Exclusive OR (XOR) Operator	770
23.3.2	NAND Operator	771
23.3.3	Complex Networks Designed Using Simple Building Blocks	772
23.4	Classification	773
23.4.1	Sonar Data Example	774
23.4.2	MXNet Notes	781

- 23.5 Case-Studies 782
 - 23.5.1 ALS Regression Example 783
 - 23.5.2 Spirals 2D Data 785
 - 23.5.3 IBS Study 789
 - 23.5.4 Country QoL Ranking Data 792
 - 23.5.5 Handwritten Digits Classification 795
- 23.6 Classifying Real-World Images 806
 - 23.6.1 Load the Pre-trained Model 806
 - 23.6.2 Load, Preprocess and Classify New Images – US Weather Pattern 806
 - 23.6.3 Lake Mapourika, New Zealand 810
 - 23.6.4 Beach Image 811
 - 23.6.5 Volcano 812
 - 23.6.6 Brain Surface 814
 - 23.6.7 Face Mask 815
- 23.7 Assignment: 23. Deep Learning, Neural Networks 816
 - 23.7.1 Deep Learning Classification 816
 - 23.7.2 Deep Learning Regression 817
 - 23.7.3 Image Classification 817
- References 817

- Summary 819**

- Glossary 823**

- Index 825**