Contents

Preface vii

1 Introduction 1

2 Overview of Supervised Learning 9

2.1 Introduction 9

2.2 Variable Types and Terminology 9

2.3 Two Simple Approaches to Prediction: Least Squares and Nearest Neighbors 11

2.3.1 Linear Models and Least Squares 11

2.3.2 Nearest-Neighbor Methods 14

2.3.3 From Least Squares to Nearest Neighbors 16

2.4 Statistical Decision Theory 18

2.5 Local Methods in High Dimensions 22

2.6 Statistical Models, Supervised Learning and Function Approximation 28

2.6.1 A Statistical Model for the Joint Distribution $\Pr(X, Y)$ 28

2.6.2 Supervised Learning 29

2.6.3 Function Approximation 29

2.7 Structured Regression Models 32

2.7.1 Difficulty of the Problem 32

2.8 Classes of Restricted Estimators 33

2.8.1 Roughness Penalty and Bayesian Methods 34

2.8.2 Kernel Methods and Local Regression 34
2.8.3 Basis Functions and Dictionary Methods 35
2.9 Model Selection and the Bias–Variance Tradeoff 37
Bibliographic Notes 39
Exercises 39

3 Linear Methods for Regression 41
3.1 Introduction 41
3.2 Linear Regression Models and Least Squares 42
3.2.1 Example: Prostate Cancer 47
3.2.2 The Gauss–Markov Theorem 49
3.3 Multiple Regression from Simple Univariate Regression . . 50
3.3.1 Multiple Outputs 54
3.4 Subset Selection and Coefficient Shrinkage 55
3.4.1 Subset Selection 55
3.4.2 Prostate Cancer Data Example (Continued) 57
3.4.3 Shrinkage Methods 59
3.4.4 Methods Using Derived Input Directions 66
3.4.5 Discussion: A Comparison of the Selection and Shrinkage Methods 68
3.4.6 Multiple Outcome Shrinkage and Selection 73
3.5 Computational Considerations 75
Bibliographic Notes 75
Exercises 75

4 Linear Methods for Classification 79
4.1 Introduction 79
4.2 Linear Regression of an Indicator Matrix 81
4.3 Linear Discriminant Analysis 84
4.3.1 Regularized Discriminant Analysis 90
4.3.2 Computations for LDA 91
4.3.3 Reduced-Rank Linear Discriminant Analysis 91
4.4 Logistic Regression 95
4.4.1 Fitting Logistic Regression Models 98
4.4.2 Example: South African Heart Disease 100
4.4.3 Quadratic Approximations and Inference 102
4.4.4 Logistic Regression or LDA? 103
4.5 Separating Hyperplanes 105
4.5.1 Rosenblatt's Perceptron Learning Algorithm 107
4.5.2 Optimal Separating Hyperplanes 108
Bibliographic Notes 111
Exercises 111

5 Basis Expansions and Regularization 115
5.1 Introduction . . . 115
5.2 Piecewise Polynomials and Splines . . . 117
5.2.1 Natural Cubic Splines . . . 120
5.2.2 Example: South African Heart Disease (Continued) . 122
5.2.3 Example: Phoneme Recognition . . . 124
5.3 Filtering and Feature Extraction . . . 126
5.4 Smoothing Splines . . . 127
5.4.1 Degrees of Freedom and Smoother Matrices . . . 129
5.5 Automatic Selection of the Smoothing Parameters . . . 134
5.5.1 Fixing the Degrees of Freedom . . . 134
5.5.2 The Bias–Variance Tradeoff . . . 134
5.6 Nonparametric Logistic Regression . . . 137
5.7 Multidimensional Splines . . . 138
5.8 Regularization and Reproducing Kernel Hilbert Spaces . . . 144
5.8.1 Spaces of Functions Generated by Kernels . . . 144
5.8.2 Examples of RKHS . . . 146
5.9 Wavelet Smoothing . . . 148
5.9.1 Wavelet Bases and the Wavelet Transform . . . 150
5.9.2 Adaptive Wavelet Filtering . . . 153
Bibliographic Notes . . . 155
Exercises . . . 155
Appendix: Computational Considerations for Splines . . . 160
Appendix: B-splines . . . 160
Appendix: Computations for Smoothing Splines . . . 163

6 Kernel Methods 165
6.1 One-Dimensional Kernel Smoothers . . . 165
6.1.1 Local Linear Regression . . . 168
6.1.2 Local Polynomial Regression . . . 171
6.2 Selecting the Width of the Kernel . . . 172
6.3 Local Regression in $\mathbb{R}^p$ . . . 174
6.4 Structured Local Regression Models in $\mathbb{R}^p$ . . . 175
6.4.1 Structured Kernels . . . 177
6.4.2 Structured Regression Functions . . . 177
6.5 Local Likelihood and Other Models . . . 179
6.6 Kernel Density Estimation and Classification . . . 182
6.6.1 Kernel Density Estimation . . . 182
6.6.2 Kernel Density Classification . . . 184
6.6.3 The Naive Bayes Classifier . . . 184
6.7 Radial Basis Functions and Kernels . . . 186
6.8 Mixture Models for Density Estimation and Classification . 188
6.9 Computational Considerations . . . 190
Bibliographic Notes . . . 190
Exercises . . . 190

7 Model Assessment and Selection **193**
7.1 Introduction 193
7.2 Bias, Variance and Model Complexity 193
7.3 The Bias–Variance Decomposition 196
7.3.1 Example: Bias–Variance Tradeoff 198
7.4 Optimism of the Training Error Rate 200
7.5 Estimates of In-Sample Prediction Error 203
7.6 The Effective Number of Parameters 205
7.7 The Bayesian Approach and BIC 206
7.8 Minimum Description Length 208
7.9 Vapnik–Chernovenkis Dimension 210
7.9.1 Example (Continued) 212
7.10 Cross-Validation 214
7.11 Bootstrap Methods 217
7.11.1 Example (Continued) 220
Bibliographic Notes 222
Exercises 222

8 Model Inference and Averaging **225**
8.1 Introduction 225
8.2 The Bootstrap and Maximum Likelihood Methods 225
8.2.1 A Smoothing Example 225
8.2.2 Maximum Likelihood Inference 229
8.2.3 Bootstrap versus Maximum Likelihood 231
8.3 Bayesian Methods 231
8.4 Relationship Between the Bootstrap and Bayesian Inference 235
8.5 The EM Algorithm 236
8.5.1 Two-Component Mixture Model 236
8.5.2 The EM Algorithm in General 240
8.5.3 EM as a Maximization–Maximization Procedure . . 241
8.6 MCMC for Sampling from the Posterior 243
8.7 Bagging 246
8.7.1 Example: Trees with Simulated Data 247
8.8 Model Averaging and Stacking 250
8.9 Stochastic Search: Bumping 253
Bibliographic Notes 254
Exercises 255

9 Additive Models, Trees, and Related Methods **257**
9.1 Generalized Additive Models 257
9.1.1 Fitting Additive Models 259
9.1.2 Example: Additive Logistic Regression 261
9.1.3 Summary 266
9.2 Tree-Based Methods 266

9.2.1 Background . . . 266
9.2.2 Regression Trees . . . 267
9.2.3 Classification Trees . . . 270
9.2.4 Other Issues . . . 272
9.2.5 Spam Example (Continued) . . . 275
9.3 PRIM—Bump Hunting . . . 279
9.3.1 Spam Example (Continued) . . . 282
9.4 MARS: Multivariate Adaptive Regression Splines . . . 283
9.4.1 Spam Example (Continued) . . . 287
9.4.2 Example (Simulated Data) . . . 288
9.4.3 Other Issues . . . 289
9.5 Hierarchical Mixtures of Experts . . . 290
9.6 Missing Data . . . 293
9.7 Computational Considerations . . . 295
Bibliographic Notes . . . 295
Exercises . . . 296

10 Boosting and Additive Trees 299
10.1 Boosting Methods . . . 299
10.1.1 Outline of this Chapter . . . 302
10.2 Boosting Fits an Additive Model . . . 303
10.3 Forward Stagewise Additive Modeling . . . 304
10.4 Exponential Loss and AdaBoost . . . 305
10.5 Why Exponential Loss? . . . 306
10.6 Loss Functions and Robustness . . . 308
10.7 "Off-the-Shelf" Procedures for Data Mining . . . 312
10.8 Example—Spam Data . . . 314
10.9 Boosting Trees . . . 316
10.10 Numerical Optimization . . . 319
10.10.1 Steepest Descent . . . 320
10.10.2 Gradient Boosting . . . 320
10.10.3 MART . . . 322
10.11 Right-Sized Trees for Boosting . . . 323
10.12 Regularization . . . 324
10.12.1 Shrinkage . . . 326
10.12.2 Penalized Regression . . . 328
10.12.3 Virtues of the L_1 Penalty (Lasso) over L_2 . . . 330
10.13 Interpretation . . . 331
10.13.1 Relative Importance of Predictor Variables . . . 331
10.13.2 Partial Dependence Plots . . . 333
10.14 Illustrations . . . 335
10.14.1 California Housing . . . 335
10.14.2 Demographics Data . . . 339
Bibliographic Notes . . . 340
Exercises . . . 344

11 Neural Networks 347

11.1 Introduction 347
11.2 Projection Pursuit Regression 347
11.3 Neural Networks 350
11.4 Fitting Neural Networks 353
11.5 Some Issues in Training Neural Networks 355
11.5.1 Starting Values 355
11.5.2 Overfitting 356
11.5.3 Scaling of the Inputs 358
11.5.4 Number of Hidden Units and Layers 358
11.5.5 Multiple Minima 359
11.6 Example: Simulated Data 359
11.7 Example: ZIP Code Data 362
11.8 Discussion 366
11.9 Computational Considerations 367
Bibliographic Notes 367
Exercises 369

12 Support Vector Machines and Flexible Discriminants 371

12.1 Introduction 371
12.2 The Support Vector Classifier 371
12.2.1 Computing the Support Vector Classifier 373
12.2.2 Mixture Example (Continued) 375
12.3 Support Vector Machines 377
12.3.1 Computing the SVM for Classification 377
12.3.2 The SVM as a Penalization Method 380
12.3.3 Function Estimation and Reproducing Kernels 381
12.3.4 SVMs and the Curse of Dimensionality 384
12.3.5 Support Vector Machines for Regression 385
12.3.6 Regression and Kernels 387
12.3.7 Discussion 389
12.4 Generalizing Linear Discriminant Analysis 390
12.5 Flexible Discriminant Analysis 391
12.5.1 Computing the FDA Estimates 394
12.6 Penalized Discriminant Analysis 397
12.7 Mixture Discriminant Analysis 399
12.7.1 Example: Waveform Data 402
Bibliographic Notes 406
Exercises 406

13 Prototype Methods and Nearest-Neighbors 411
13.1 Introduction 411
13.2 Prototype Methods 411
13.2.1 K-means Clustering 412
13.2.2 Learning Vector Quantization 414
13.2.3 Gaussian Mixtures 415
13.3 k-Nearest-Neighbor Classifiers 415
13.3.1 Example: A Comparative Study 420
13.3.2 Example: k-Nearest-Neighbors and Image Scene Classification 422
13.3.3 Invariant Metrics and Tangent Distance 423
13.4 Adaptive Nearest-Neighbor Methods 427
13.4.1 Example 430
13.4.2 Global Dimension Reduction for Nearest-Neighbors . 431
13.5 Computational Considerations 432
Bibliographic Notes 433
Exercises 433

14 Unsupervised Learning 437
14.1 Introduction 437
14.2 Association Rules 439
14.2.1 Market Basket Analysis 440
14.2.2 The Apriori Algorithm 441
14.2.3 Example: Market Basket Analysis 444
14.2.4 Unsupervised as Supervised Learning 447
14.2.5 Generalized Association Rules 449
14.2.6 Choice of Supervised Learning Method 451
14.2.7 Example: Market Basket Analysis (Continued) . . . 451
14.3 Cluster Analysis 453
14.3.1 Proximity Matrices 455
14.3.2 Dissimilarities Based on Attributes 455
14.3.3 Object Dissimilarity 457
14.3.4 Clustering Algorithms 459
14.3.5 Combinatorial Algorithms 460
14.3.6 K-means 461
14.3.7 Gaussian Mixtures as Soft K-means Clustering . . . 463
14.3.8 Example: Human Tumor Microarray Data 463
14.3.9 Vector Quantization 466
14.3.10 K-medoids 468
14.3.11 Practical Issues 470
14.3.12 Hierarchical Clustering 472
14.4 Self-Organizing Maps 480
14.5 Principal Components, Curves and Surfaces 485
14.5.1 Principal Components 485
14.5.2 Principal Curves and Surfaces 491

14.6 Independent Component Analysis and Exploratory Projection Pursuit 494
14.6.1 Latent Variables and Factor Analysis 494
14.6.2 Independent Component Analysis 496
14.6.3 Exploratory Projection Pursuit 500
14.6.4 A Different Approach to ICA 500
14.7 Multidimensional Scaling 502
Bibliographic Notes 503
Exercises 504

References **509**

Author Index **523**

Index **527**