Contents

Preface to the Second Edition **vii**

Preface to the First Edition **xi**

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 43
3.1 Introduction 43
3.2 Linear Regression Models and Least Squares 44
3.2.1 Example: Prostate Cancer 49
3.2.2 The Gauss–Markov Theorem 51
3.2.3 Multiple Regression from Simple Univariate Regression 52
3.2.4 Multiple Outputs 56
3.3 Subset Selection 57
3.3.1 Best-Subset Selection 57
3.3.2 Forward- and Backward-Stepwise Selection . . . 58
3.3.3 Forward-Stagewise Regression 60
3.3.4 Prostate Cancer Data Example (Continued) . . 61
3.4 Shrinkage Methods 61
3.4.1 Ridge Regression 61
3.4.2 The Lasso 68
3.4.3 Discussion: Subset Selection, Ridge Regression and the Lasso 69
3.4.4 Least Angle Regression 73
3.5 Methods Using Derived Input Directions 79
3.5.1 Principal Components Regression 79
3.5.2 Partial Least Squares 80
3.6 Discussion: A Comparison of the Selection and Shrinkage Methods 82
3.7 Multiple Outcome Shrinkage and Selection 84
3.8 More on the Lasso and Related Path Algorithms 86
3.8.1 Incremental Forward Stagewise Regression . . . 86
3.8.2 Piecewise-Linear Path Algorithms 89
3.8.3 The Dantzig Selector 89
3.8.4 The Grouped Lasso 90
3.8.5 Further Properties of the Lasso 91
3.8.6 Pathwise Coordinate Optimization 92
3.9 Computational Considerations 93
Bibliographic Notes 94
Exercises 94

4 Linear Methods for Classification **101**
4.1 Introduction . . . 101
4.2 Linear Regression of an Indicator Matrix . . . 103
4.3 Linear Discriminant Analysis . . . 106
4.3.1 Regularized Discriminant Analysis . . . 112
4.3.2 Computations for LDA . . . 113
4.3.3 Reduced-Rank Linear Discriminant Analysis . . 113
4.4 Logistic Regression . . . 119
4.4.1 Fitting Logistic Regression Models . . . 120
4.4.2 Example: South African Heart Disease . . . 122
4.4.3 Quadratic Approximations and Inference . . . 124
4.4.4 L_1 Regularized Logistic Regression . . . 125
4.4.5 Logistic Regression or LDA? . . . 127
4.5 Separating Hyperplanes . . . 129
4.5.1 Rosenblatt's Perceptron Learning Algorithm . . 130
4.5.2 Optimal Separating Hyperplanes . . . 132
Bibliographic Notes . . . 135
Exercises . . . 135

5 Basis Expansions and Regularization **139**
5.1 Introduction . . . 139
5.2 Piecewise Polynomials and Splines . . . 141
5.2.1 Natural Cubic Splines . . . 144
5.2.2 Example: South African Heart Disease (Continued) 146
5.2.3 Example: Phoneme Recognition . . . 148
5.3 Filtering and Feature Extraction . . . 150
5.4 Smoothing Splines . . . 151
5.4.1 Degrees of Freedom and Smoother Matrices . . . 153
5.5 Automatic Selection of the Smoothing Parameters . . . 156
5.5.1 Fixing the Degrees of Freedom . . . 158
5.5.2 The Bias–Variance Tradeoff . . . 158
5.6 Nonparametric Logistic Regression . . . 161
5.7 Multidimensional Splines . . . 162
5.8 Regularization and Reproducing Kernel Hilbert Spaces . 167
5.8.1 Spaces of Functions Generated by Kernels . . . 168
5.8.2 Examples of RKHS . . . 170
5.9 Wavelet Smoothing . . . 174
5.9.1 Wavelet Bases and the Wavelet Transform . . . 176
5.9.2 Adaptive Wavelet Filtering . . . 179
Bibliographic Notes . . . 181
Exercises . . . 181
Appendix: Computational Considerations for Splines . . . 186
Appendix: B-splines . . . 186
Appendix: Computations for Smoothing Splines . . . 189

6 Kernel Smoothing Methods **191**
6.1 One-Dimensional Kernel Smoothers 192
6.1.1 Local Linear Regression 194
6.1.2 Local Polynomial Regression 197
6.2 Selecting the Width of the Kernel 198
6.3 Local Regression in $\mathbb{R}^p$ 200
6.4 Structured Local Regression Models in $\mathbb{R}^p$ 201
6.4.1 Structured Kernels 203
6.4.2 Structured Regression Functions 203
6.5 Local Likelihood and Other Models 205
6.6 Kernel Density Estimation and Classification 208
6.6.1 Kernel Density Estimation 208
6.6.2 Kernel Density Classification 210
6.6.3 The Naive Bayes Classifier 210
6.7 Radial Basis Functions and Kernels 212
6.8 Mixture Models for Density Estimation and Classification 214
6.9 Computational Considerations 216
Bibliographic Notes 216
Exercises 216

7 Model Assessment and Selection **219**
7.1 Introduction 219
7.2 Bias, Variance and Model Complexity 219
7.3 The Bias–Variance Decomposition 223
7.3.1 Example: Bias–Variance Tradeoff 226
7.4 Optimism of the Training Error Rate 228
7.5 Estimates of In-Sample Prediction Error 230
7.6 The Effective Number of Parameters 232
7.7 The Bayesian Approach and BIC 233
7.8 Minimum Description Length 235
7.9 Vapnik–Chervonenkis Dimension 237
7.9.1 Example (Continued) 239
7.10 Cross-Validation 241
7.10.1 K-Fold Cross-Validation 241
7.10.2 The Wrong and Right Way to Do Cross-validation 245
7.10.3 Does Cross-Validation Really Work? 247
7.11 Bootstrap Methods 249
7.11.1 Example (Continued) 252
7.12 Conditional or Expected Test Error? 254
Bibliographic Notes 257
Exercises 257

8 Model Inference and Averaging **261**
8.1 Introduction 261

8.2 The Bootstrap and Maximum Likelihood Methods 261
8.2.1 A Smoothing Example 261
8.2.2 Maximum Likelihood Inference 265
8.2.3 Bootstrap versus Maximum Likelihood 267
8.3 Bayesian Methods 267
8.4 Relationship Between the Bootstrap and Bayesian Inference 271
8.5 The EM Algorithm 272
8.5.1 Two-Component Mixture Model 272
8.5.2 The EM Algorithm in General 276
8.5.3 EM as a Maximization–Maximization Procedure 277
8.6 MCMC for Sampling from the Posterior 279
8.7 Bagging 282
8.7.1 Example: Trees with Simulated Data 283
8.8 Model Averaging and Stacking 288
8.9 Stochastic Search: Bumping 290
Bibliographic Notes 292
Exercises 293

9 Additive Models, Trees, and Related Methods 295
9.1 Generalized Additive Models 295
9.1.1 Fitting Additive Models 297
9.1.2 Example: Additive Logistic Regression 299
9.1.3 Summary 304
9.2 Tree-Based Methods 305
9.2.1 Background 305
9.2.2 Regression Trees 307
9.2.3 Classification Trees 308
9.2.4 Other Issues 310
9.2.5 Spam Example (Continued) 313
9.3 PRIM: Bump Hunting 317
9.3.1 Spam Example (Continued) 320
9.4 MARS: Multivariate Adaptive Regression Splines 321
9.4.1 Spam Example (Continued) 326
9.4.2 Example (Simulated Data) 327
9.4.3 Other Issues 328
9.5 Hierarchical Mixtures of Experts 329
9.6 Missing Data 332
9.7 Computational Considerations 334
Bibliographic Notes 334
Exercises 335

10 Boosting and Additive Trees 337
10.1 Boosting Methods 337
10.1.1 Outline of This Chapter 340

10.2 Boosting Fits an Additive Model . . . 341
10.3 Forward Stagewise Additive Modeling . . . 342
10.4 Exponential Loss and AdaBoost . . . 343
10.5 Why Exponential Loss? . . . 345
10.6 Loss Functions and Robustness . . . 346
10.7 "Off-the-Shelf" Procedures for Data Mining . . . 350
10.8 Example: Spam Data . . . 352
10.9 Boosting Trees . . . 353
10.10 Numerical Optimization via Gradient Boosting . . . 358
10.10.1 Steepest Descent . . . 358
10.10.2 Gradient Boosting . . . 359
10.10.3 Implementations of Gradient Boosting . . . 360
10.11 Right-Sized Trees for Boosting . . . 361
10.12 Regularization . . . 364
10.12.1 Shrinkage . . . 364
10.12.2 Subsampling . . . 365
10.13 Interpretation . . . 367
10.13.1 Relative Importance of Predictor Variables . . . 367
10.13.2 Partial Dependence Plots . . . 369
10.14 Illustrations . . . 371
10.14.1 California Housing . . . 371
10.14.2 New Zealand Fish . . . 375
10.14.3 Demographics Data . . . 379
Bibliographic Notes . . . 380
Exercises . . . 384

11 Neural Networks 389
11.1 Introduction . . . 389
11.2 Projection Pursuit Regression . . . 389
11.3 Neural Networks . . . 392
11.4 Fitting Neural Networks . . . 395
11.5 Some Issues in Training Neural Networks . . . 397
11.5.1 Starting Values . . . 397
11.5.2 Overfitting . . . 398
11.5.3 Scaling of the Inputs . . . 398
11.5.4 Number of Hidden Units and Layers . . . 400
11.5.5 Multiple Minima . . . 400
11.6 Example: Simulated Data . . . 401
11.7 Example: ZIP Code Data . . . 404
11.8 Discussion . . . 408
11.9 Bayesian Neural Nets and the NIPS 2003 Challenge . . . 409
11.9.1 Bayes, Boosting and Bagging . . . 410
11.9.2 Performance Comparisons . . . 412
11.10 Computational Considerations . . . 414
Bibliographic Notes . . . 415

Exercises 415

12 Support Vector Machines and Flexible Discriminants 417
12.1 Introduction 417
12.2 The Support Vector Classifier 417
12.2.1 Computing the Support Vector Classifier 420
12.2.2 Mixture Example (Continued) 421
12.3 Support Vector Machines and Kernels 423
12.3.1 Computing the SVM for Classification 423
12.3.2 The SVM as a Penalization Method 426
12.3.3 Function Estimation and Reproducing Kernels . 428
12.3.4 SVMs and the Curse of Dimensionality 431
12.3.5 A Path Algorithm for the SVM Classifier 432
12.3.6 Support Vector Machines for Regression 434
12.3.7 Regression and Kernels 436
12.3.8 Discussion 438
12.4 Generalizing Linear Discriminant Analysis 438
12.5 Flexible Discriminant Analysis 440
12.5.1 Computing the FDA Estimates 444
12.6 Penalized Discriminant Analysis 446
12.7 Mixture Discriminant Analysis 449
12.7.1 Example: Waveform Data 451
Bibliographic Notes 455
Exercises 455

13 Prototype Methods and Nearest-Neighbors 459
13.1 Introduction 459
13.2 Prototype Methods 459
13.2.1 K-means Clustering 460
13.2.2 Learning Vector Quantization 462
13.2.3 Gaussian Mixtures 463
13.3 k-Nearest-Neighbor Classifiers 463
13.3.1 Example: A Comparative Study 468
13.3.2 Example: k-Nearest-Neighbors and Image Scene Classification 470
13.3.3 Invariant Metrics and Tangent Distance 471
13.4 Adaptive Nearest-Neighbor Methods 475
13.4.1 Example 478
13.4.2 Global Dimension Reduction for Nearest-Neighbors 479
13.5 Computational Considerations 480
Bibliographic Notes 481
Exercises 481

14 Unsupervised Learning **485**
14.1 Introduction . . . 485
14.2 Association Rules . . . 487
14.2.1 Market Basket Analysis . . . 488
14.2.2 The Apriori Algorithm . . . 489
14.2.3 Example: Market Basket Analysis . . . 492
14.2.4 Unsupervised as Supervised Learning . . . 495
14.2.5 Generalized Association Rules . . . 497
14.2.6 Choice of Supervised Learning Method . . . 499
14.2.7 Example: Market Basket Analysis (Continued) . 499
14.3 Cluster Analysis . . . 501
14.3.1 Proximity Matrices . . . 503
14.3.2 Dissimilarities Based on Attributes . . . 503
14.3.3 Object Dissimilarity . . . 505
14.3.4 Clustering Algorithms . . . 507
14.3.5 Combinatorial Algorithms . . . 507
14.3.6 K-means . . . 509
14.3.7 Gaussian Mixtures as Soft K-means Clustering . 510
14.3.8 Example: Human Tumor Microarray Data . . . 512
14.3.9 Vector Quantization . . . 514
14.3.10 K-medoids . . . 515
14.3.11 Practical Issues . . . 518
14.3.12 Hierarchical Clustering . . . 520
14.4 Self-Organizing Maps . . . 528
14.5 Principal Components, Curves and Surfaces . . . 534
14.5.1 Principal Components . . . 534
14.5.2 Principal Curves and Surfaces . . . 541
14.5.3 Spectral Clustering . . . 544
14.5.4 Kernel Principal Components . . . 547
14.5.5 Sparse Principal Components . . . 550
14.6 Non-negative Matrix Factorization . . . 553
14.6.1 Archetypal Analysis . . . 554
14.7 Independent Component Analysis and Exploratory Projection Pursuit . . . 557
14.7.1 Latent Variables and Factor Analysis . . . 558
14.7.2 Independent Component Analysis . . . 560
14.7.3 Exploratory Projection Pursuit . . . 565
14.7.4 A Direct Approach to ICA . . . 565
14.8 Multidimensional Scaling . . . 570
14.9 Nonlinear Dimension Reduction and Local Multidimensional Scaling . . . 572
14.10 The Google PageRank Algorithm . . . 576
Bibliographic Notes . . . 578
Exercises . . . 579

15 Random Forests **587**
15.1 Introduction . . . 587
15.2 Definition of Random Forests . . . 587
15.3 Details of Random Forests . . . 592
15.3.1 Out of Bag Samples . . . 592
15.3.2 Variable Importance . . . 593
15.3.3 Proximity Plots . . . 595
15.3.4 Random Forests and Overfitting . . . 596
15.4 Analysis of Random Forests . . . 597
15.4.1 Variance and the De-Correlation Effect . . . 597
15.4.2 Bias . . . 600
15.4.3 Adaptive Nearest Neighbors . . . 601
Bibliographic Notes . . . 602
Exercises . . . 603

16 Ensemble Learning **605**
16.1 Introduction . . . 605
16.2 Boosting and Regularization Paths . . . 607
16.2.1 Penalized Regression . . . 607
16.2.2 The "Bet on Sparsity" Principle . . . 610
16.2.3 Regularization Paths, Over-fitting and Margins . 613
16.3 Learning Ensembles . . . 616
16.3.1 Learning a Good Ensemble . . . 617
16.3.2 Rule Ensembles . . . 622
Bibliographic Notes . . . 623
Exercises . . . 624

17 Undirected Graphical Models **625**
17.1 Introduction . . . 625
17.2 Markov Graphs and Their Properties . . . 627
17.3 Undirected Graphical Models for Continuous Variables . 630
17.3.1 Estimation of the Parameters when the Graph Structure is Known . . . 631
17.3.2 Estimation of the Graph Structure . . . 635
17.4 Undirected Graphical Models for Discrete Variables . . . 638
17.4.1 Estimation of the Parameters when the Graph Structure is Known . . . 639
17.4.2 Hidden Nodes . . . 641
17.4.3 Estimation of the Graph Structure . . . 642
17.4.4 Restricted Boltzmann Machines . . . 643
Exercises . . . 645

18 High-Dimensional Problems: $p \gg N$ **649**
18.1 When p is Much Bigger than N . . . 649

18.2 Diagonal Linear Discriminant Analysis and Nearest Shrunken Centroids 651
18.3 Linear Classifiers with Quadratic Regularization 654
18.3.1 Regularized Discriminant Analysis 656
18.3.2 Logistic Regression with Quadratic Regularization 657
18.3.3 The Support Vector Classifier 657
18.3.4 Feature Selection 658
18.3.5 Computational Shortcuts When $p \gg N$ 659
18.4 Linear Classifiers with L_1 Regularization 661
18.4.1 Application of Lasso to Protein Mass Spectroscopy 664
18.4.2 The Fused Lasso for Functional Data 666
18.5 Classification When Features are Unavailable 668
18.5.1 Example: String Kernels and Protein Classification 668
18.5.2 Classification and Other Models Using Inner-Product Kernels and Pairwise Distances . 670
18.5.3 Example: Abstracts Classification 672
18.6 High-Dimensional Regression: Supervised Principal Components 674
18.6.1 Connection to Latent-Variable Modeling 678
18.6.2 Relationship with Partial Least Squares 680
18.6.3 Pre-Conditioning for Feature Selection 681
18.7 Feature Assessment and the Multiple-Testing Problem . . 683
18.7.1 The False Discovery Rate 687
18.7.2 Asymmetric Cutpoints and the SAM Procedure 690
18.7.3 A Bayesian Interpretation of the FDR 692
18.8 Bibliographic Notes 693
Exercises 694

References **699**

Author Index **729**

Index **737**