Contents

Preface VII

1 Statistical Learning as a Regression Problem 1
1.1 Getting Started 1
1.2 Setting the Regression Context 2
1.3 The Transition to Statistical Learning 8
1.3.1 Some Goals of Statistical Learning 9
1.3.2 Statistical Inference 14
1.3.3 Some Initial Cautions 16
1.3.4 A Cartoon Illustration 17
1.3.5 A Taste of Things to Come 20
1.4 Some Initial Concepts and Definitions 22
1.4.1 Overall Goals 23
1.4.2 Loss Functions and Related Concepts 23
1.4.3 Linear Estimators 26
1.4.4 Degrees of Freedom 28
1.4.5 Model Evaluation 29
1.4.6 Model Selection 34
1.4.7 Basis Functions 37
1.5 Some Common Themes 41
1.6 Summary and Conclusions 43

2 Regression Splines and Regression Smoothers 49
2.1 Introduction 49
2.2 Regression Splines 49
2.2.1 Applying a Piecewise Linear Basis 49
2.2.2 Polynomial Regression Splines 53
2.2.3 Natural Cubic Splines 54
2.2.4 B-Splines 57
2.3 Penalized Smoothing 60
2.3.1 Shrinkage 61

2.3.2 Shrinkage and Statistical Inference 68
2.3.3 Shrinkage: So What? 69
2.4 Smoothing Splines 70
2.4.1 An Illustration 72
2.5 Locally Weighted Regression as a Smoother 73
2.5.1 Nearest Neighbor Methods 73
2.5.2 Locally Weighted Regression 75
2.6 Smoothers for Multiple Predictors 80
2.6.1 Smoothing in Two Dimensions 81
2.6.2 The Generalized Additive Model 84
2.7 Smoothers with Categorical Variables 88
2.7.1 An Illustration 88
2.8 Locally Adaptive Smoothers 91
2.9 The Role of Statistical Inference 93
2.9.1 Some Apparent Prerequisites 93
2.9.2 Confidence Intervals 94
2.9.3 Statistical Tests 96
2.9.4 Can Asymptotics Help? 97
2.10 Software Issues 98
2.11 Summary and Conclusions 99

3 Classification and Regression Trees (CART) 103
3.1 Introduction 103
3.2 An Overview of Recursive Partitioning with CART 105
3.2.1 Tree Diagrams 105
3.2.2 Classification and Forecasting with CART 108
3.2.3 Confusion Tables 108
3.2.4 CART as an Adaptive Nearest Neighbor Method 110
3.2.5 What CART Needs to Do 112
3.3 Splitting a Node 113
3.4 More on Classification 117
3.4.1 Fitted Values and Related Terms 117
3.4.2 An Example 119
3.5 Classification Errors and Costs 122
3.5.1 Default Costs in CART 123
3.5.2 Prior Probabilities and Costs 125
3.6 Pruning 128
3.6.1 Impurity Versus $R_\alpha(T)$ 130
3.7 Missing Data 131
3.7.1 Missing Data with CART 132
3.8 Statistical Inference with CART 135
3.9 Classification Versus Forecasting 138
3.10 Varying the Prior, Costs, and the Complexity Penalty 139
3.11 An Example with Three Response Categories 145
3.12 CART with Highly Skewed Response Distributions 149

3.13 Some Cautions in Interpreting CART Results 149
3.13.1 Model Bias 149
3.13.2 Model Variance 150
3.14 Regression Trees 154
3.14.1 An Illustration 155
3.14.2 Some Extensions 157
3.14.3 Multivariate Adaptive Regression Splines (MARS) 158
3.15 Software Issues 160
3.16 Summary and Conclusions 161

4 Bagging 169
4.1 Introduction 169
4.2 Overfitting and Cross-Validation 170
4.3 Bagging as an Algorithm 172
4.3.1 Margins 173
4.3.2 Out-Of-Bag Observations 173
4.4 Some Thinking on Why Bagging Works 174
4.4.1 More on Instability in CART 174
4.4.2 How Bagging Can Help 179
4.4.3 A Somewhat More Formal Explanation 180
4.5 Some Limitations of Bagging 182
4.5.1 Sometimes Bagging Does Not Help 182
4.5.2 Sometimes Bagging Can Make the Bias Worse 183
4.5.3 Sometimes Bagging Can Make the Variance Worse 183
4.5.4 Losing the Trees for the Forest 186
4.5.5 Bagging Is Only an Algorithm 186
4.6 An Example 186
4.7 Bagging a Quantitative Response Variable 187
4.8 Software Considerations 188
4.9 Summary and Conclusions 190

5 Random Forests 193
5.1 Introduction and Overview 193
5.1.1 Unpacking How Random Forests Works 194
5.2 An Initial Illustration 198
5.3 A Few Formalities 199
5.3.1 What Is a Random Forest? 199
5.3.2 Margins and Generalization Error for Classifiers in General 200
5.3.3 Generalization Error for Random Forests 201
5.3.4 The Strength of a Random Forest 202
5.3.5 Dependence 203
5.3.6 Implications 203
5.4 Random Forests and Adaptive Nearest Neighbor Methods 204
5.5 Taking Costs into Account in Random Forests 210

5.5.1 A Brief Illustration 212
5.6 Determining the Importance of the Predictors 213
5.6.1 Contributions to the Fit 213
5.6.2 Contributions to Forecasting Skill 214
5.7 Response Functions 222
5.7.1 An Example 226
5.8 The Proximity Matrix 229
5.8.1 Clustering by Proximity Values 231
5.8.2 Using Proximity Values to Impute Missing Data 231
5.8.3 Using Proximities to Detect Outliers 232
5.9 Quantitative Response Variables 233
5.10 Tuning Parameters 234
5.11 An Illustration Using a Binary Response Variable 236
5.12 An Illustration Using a Quantitative Response Variable 242
5.13 Software Considerations 249
5.14 Summary and Conclusions 252
5.14.1 Problem Set 1 252
5.14.2 Problem Set 2 253
5.14.3 Problem Set 3 254

6 Boosting 257
6.1 Introduction 257
6.2 Adaboost 258
6.2.1 A Toy Numerical Example of Adaboost 259
6.2.2 A Statistical Perspective on Adaboost 261
6.3 Why Does Adaboost Work So Well? 263
6.3.1 Least Angle Regression (LARS) 264
6.4 Stochastic Gradient Boosting 266
6.4.1 Tuning Parameters 271
6.4.2 Output 273
6.5 Some Problems and Some Possible Solutions 274
6.5.1 Some Potential Problems 274
6.5.2 Some Potential Solutions 275
6.6 Some Examples 277
6.6.1 A Garden Variety Data Analysis 277
6.6.2 Inmate Misconduct Again 281
6.6.3 Homicides and the Impact of Executions 286
6.6.4 Imputing the Number of Homeless 290
6.6.5 Estimating Conditional Probabilities 293
6.7 Software Considerations 295
6.8 Summary and Conclusions 296

7 Support Vector Machines . . . 301
7.1 A Simple Didactic Illustration . . . 301
7.2 Support Vector Machines in Pictures . . . 303
7.2.1 Support Vector Classifiers . . . 303
7.2.2 Support Vector Machines . . . 307
7.3 Support Vector Machines in Statistical Notation . . . 309
7.3.1 Support Vector Classifiers . . . 309
7.3.2 Support Vector Machines . . . 312
7.3.3 SVM for Regression . . . 315
7.4 A Classification Example . . . 315
7.4.1 SVM Analysis with a Linear Kernel . . . 316
7.4.2 SVM Analysis with a Radial Kernel . . . 318
7.4.3 Varying Tuning Parameters . . . 318
7.4.4 Taking the Costs of Classification Errors into Account . 321
7.4.5 Comparisons to Logistic Regression . . . 322
7.5 Software Considerations . . . 324
7.6 Summary and Conclusions . . . 326

8 Broader Implications and a Bit of Craft Lore . . . 329
8.1 Some Fundamental Limitations of Statistical Learning . . . 329
8.2 Some Assets of Statistical Learning . . . 330
8.2.1 The Attitude Adjustment . . . 331
8.2.2 Selectively Better Performance . . . 332
8.2.3 Improving Other Procedures . . . 334
8.3 Some Practical Suggestions . . . 335
8.3.1 Matching Tools to Jobs . . . 335
8.3.2 Getting to Know Your Software . . . 337
8.3.3 Not Forgetting the Basics . . . 337
8.3.4 Getting Good Data . . . 338
8.3.5 Being Sensitive to Overtuning . . . 339
8.3.6 Matching Your Goals to What You Can Credibly Do . . 339
8.4 Some Concluding Observations . . . 340

References . . . 343

Index . . . 355