Contents

1 Statistical Learning as a Regression Problem 1
1.1 Getting Started 2
1.2 Setting the Regression Context 2
1.3 Revisiting the Ubiquitous Linear Regression Model 8
1.3.1 Problems in Practice 9
1.4 Working with Statistical Models that Are Wrong 11
1.4.1 An Alternative Approach to Regression 15
1.5 The Transition to Statistical Learning 23
1.5.1 Models Versus Algorithms 24
1.6 Some Initial Concepts 28
1.6.1 Overall Goals of Statistical Learning 29
1.6.2 Data Requirements: Training Data, Evaluation Data and Test Data 31
1.6.3 Loss Functions and Related Concepts 35
1.6.4 The Bias-Variance Tradeoff 38
1.6.5 Linear Estimators 39
1.6.6 Degrees of Freedom 40
1.6.7 Basis Functions 42
1.6.8 The Curse of Dimensionality 46
1.7 Statistical Learning in Context 48

2 Splines, Smoothers, and Kernels 55
2.1 Introduction 55
2.2 Regression Splines 55
2.2.1 Applying a Piecewise Linear Basis 56
2.2.2 Polynomial Regression Splines 61
2.2.3 Natural Cubic Splines 63
2.2.4 *B*-Splines 66
2.3 Penalized Smoothing 69
2.3.1 Shrinkage and Regularization 70

2.4 Smoothing Splines . . . 81
2.4.1 A Smoothing Splines Illustration . . . 84
2.5 Locally Weighted Regression as a Smoother . . . 86
2.5.1 Nearest Neighbor Methods . . . 87
2.5.2 Locally Weighted Regression . . . 88
2.6 Smoothers for Multiple Predictors . . . 92
2.6.1 Smoothing in Two Dimensions . . . 93
2.6.2 The Generalized Additive Model . . . 96
2.7 Smoothers with Categorical Variables . . . 103
2.7.1 An Illustration Using the Generalized Additive Model with a Binary Outcome . . . 103
2.8 An Illustration of Statistical Inference After Model Selection . . . 106
2.9 Kernelized Regression . . . 114
2.9.1 Radial Basis Kernel . . . 118
2.9.2 ANOVA Radial Basis Kernel . . . 120
2.9.3 A Kernel Regression Application . . . 120
2.10 Summary and Conclusions . . . 124

3 Classification and Regression Trees (CART) . . . 129
3.1 Introduction . . . 129
3.2 The Basic Ideas . . . 131
3.2.1 Tree Diagrams for Understanding Conditional Relationships . . . 132
3.2.2 Classification and Forecasting with CART . . . 136
3.2.3 Confusion Tables . . . 137
3.2.4 CART as an Adaptive Nearest Neighbor Method . . . 139
3.3 Splitting a Node . . . 140
3.4 Fitted Values . . . 144
3.4.1 Fitted Values in Classification . . . 144
3.4.2 An Illustrative Prison Inmate Risk Assessment Using CART . . . 145
3.5 Classification Errors and Costs . . . 148
3.5.1 Default Costs in CART . . . 149
3.5.2 Prior Probabilities and Relative Misclassification Costs . . . 151
3.6 Pruning . . . 157
3.6.1 Impurity Versus $R_\alpha(T)$ . . . 159
3.7 Missing Data . . . 159
3.7.1 Missing Data with CART . . . 161
3.8 Statistical Inference with CART . . . 163
3.9 From Classification to Forecasting . . . 165
3.10 Varying the Prior and the Complexity Parameter . . . 166
3.11 An Example with Three Response Categories . . . 170
3.12 Some Further Cautions in Interpreting CART Results . . . 173
3.12.1 Model Bias . . . 173

3.12.2 Model Variance . . . 173
3.13 Regression Trees . . . 175
3.13.1 A CART Application for the Correlates of a Student's GPA in High School . . . 177
3.14 Multivariate Adaptive Regression Splines (MARS) . . . 179
3.15 Summary and Conclusions . . . 181

4 Bagging . . . 187
4.1 Introduction . . . 187
4.2 The Bagging Algorithm . . . 188
4.3 Some Bagging Details . . . 189
4.3.1 Revisiting the CART Instability Problem . . . 189
4.3.2 Some Background on Resampling . . . 190
4.3.3 Votes and Probabilities . . . 193
4.3.4 Imputation and Forecasting . . . 193
4.3.5 Margins . . . 193
4.3.6 Using Out-Of-Bag Observations as Test Data . . . 195
4.3.7 Bagging and Bias . . . 195
4.3.8 Level I and Level II Analyses with Bagging . . . 196
4.4 Some Limitations of Bagging . . . 197
4.4.1 Sometimes Bagging Cannot Help . . . 197
4.4.2 Sometimes Bagging Can Make the Bias Worse . . . 197
4.4.3 Sometimes Bagging Can Make the Variance Worse . . . 198
4.5 A Bagging Illustration . . . 199
4.6 Bagging a Quantitative Response Variable . . . 200
4.7 Summary and Conclusions . . . 201

5 Random Forests . . . 205
5.1 Introduction and Overview . . . 205
5.1.1 Unpacking How Random Forests Works . . . 206
5.2 An Initial Random Forests Illustration . . . 208
5.3 A Few Technical Formalities . . . 210
5.3.1 What Is a Random Forest? . . . 211
5.3.2 Margins and Generalization Error for Classifiers in General . . . 211
5.3.3 Generalization Error for Random Forests . . . 212
5.3.4 The Strength of a Random Forest . . . 214
5.3.5 Dependence . . . 214
5.3.6 Implications . . . 214
5.3.7 Putting It All Together . . . 215
5.4 Random Forests and Adaptive Nearest Neighbor Methods . . . 217
5.5 Introducing Misclassification Costs . . . 221
5.5.1 A Brief Illustration Using Asymmetric Costs . . . 222
5.6 Determining the Importance of the Predictors . . . 224
5.6.1 Contributions to the Fit . . . 224

5.6.2 Contributions to Prediction . . . 225
5.7 Input Response Functions . . . 230
5.7.1 Partial Dependence Plot Examples . . . 234
5.8 Classification and the Proximity Matrix . . . 237
5.8.1 Clustering by Proximity Values . . . 238
5.9 Empirical Margins . . . 242
5.10 Quantitative Response Variables . . . 243
5.11 A Random Forest Illustration Using a Quantitative Response Variable . . . 245
5.12 Statistical Inference with Random Forests . . . 250
5.13 Software and Tuning Parameters . . . 252
5.14 Summary and Conclusions . . . 255
5.14.1 Problem Set 2 . . . 256
5.14.2 Problem Set 3 . . . 257

6 Boosting . . . 259
6.1 Introduction . . . 259
6.2 Adaboost . . . 260
6.2.1 A Toy Numerical Example of Adaboost.M1 . . . 261
6.2.2 Why Does Boosting Work so Well for Classification? . . . 263
6.3 Stochastic Gradient Boosting . . . 266
6.3.1 Tuning Parameters . . . 271
6.3.2 Output . . . 273
6.4 Asymmetric Costs . . . 274
6.5 Boosting, Estimation, and Consistency . . . 276
6.6 A Binomial Example . . . 276
6.7 A Quantile Regression Example . . . 281
6.8 Summary and Conclusions . . . 286

7 Support Vector Machines . . . 291
7.1 Support Vector Machines in Pictures . . . 292
7.1.1 The Support Vector Classifier . . . 292
7.1.2 Support Vector Machines . . . 295
7.2 Support Vector Machines More Formally . . . 295
7.2.1 The Support Vector Classifier Again: The Separable Case . . . 296
7.2.2 The Nonseparable Case . . . 297
7.2.3 Support Vector Machines . . . 299
7.2.4 SVM for Regression . . . 301
7.2.5 Statistical Inference for Support Vector Machines . . . 301
7.3 A Classification Example . . . 302
7.4 Summary and Conclusions . . . 308

8 Some Other Procedures Briefly 311
8.1 Neural Networks. 311
8.2 Bayesian Additive Regression Trees (BART). 316
8.3 Reinforcement Learning and Genetic Algorithms 320
8.3.1 Genetic Algorithms 320

9 Broader Implications and a Bit of Craft Lore 325
9.1 Some Integrating Themes 325
9.2 Some Practical Suggestions 326
9.2.1 Choose the Right Procedure 326
9.2.2 Get to Know Your Software 328
9.2.3 Do Not Forget the Basics 329
9.2.4 Getting Good Data 330
9.2.5 Match Your Goals to What You Can Credibly Do 331
9.3 Some Concluding Observations 331

Erratum to: Statistical Learning from a Regression Perspective E1

References 333

Index 343

http://www.springer.com/978-3-319-44047-7

Statistical Learning from a Regression Perspective

Berk, R.A.
2016, XXV, 347 p. 120 illus., 91 illus. in color.,
Hardcover
ISBN: 978-3-319-44047-7