Contents

Preface xi
Organization—How to Use This Book xiii
Acknowledgments xvii
About the Companion Website xix

1 Introduction—Examples from Real Life 1

2 The Problem of Learning 3
   2.1 Domain 4
   2.2 Range 4
   2.3 Data 4
   2.4 Loss 6
   2.5 Risk 8
   2.6 The Reality of the Unknown Function 12
   2.7 Training and Selection of Models, and Purposes of Learning 12
   2.8 Notation 13

3 Regression 15
   3.1 General Framework 16
   3.2 Loss 17
   3.3 Estimating the Model Parameters 17
   3.4 Properties of Fitted Values 19
   3.5 Estimating the Variance 22
   3.6 A Normality Assumption 23
   3.7 Computation 24
   3.8 Categorical Features 25
   3.9 Feature Transformations, Expansions, and Interactions 27
   3.10 Variations in Linear Regression 28
   3.11 Nonparametric Regression 32
CONTENTS

4 Survey of Classification Techniques 33
  4.1 The Bayes Classifier 34
  4.2 Introduction to Classifiers 37
  4.3 A Running Example 38
  4.4 Likelihood Methods 40
    4.4.1 Quadratic Discriminant Analysis 41
    4.4.2 Linear Discriminant Analysis 43
    4.4.3 Gaussian Mixture Models 45
    4.4.4 Kernel Density Estimation 47
    4.4.5 Histograms 51
    4.4.6 The Naive Bayes Classifier 54
  4.5 Prototype Methods 54
    4.5.1 $k$-Nearest-Neighbor 55
    4.5.2 Condensed $k$-Nearest-Neighbor 56
    4.5.3 Nearest-Cluster 56
    4.5.4 Learning Vector Quantization 58
  4.6 Logistic Regression 59
  4.7 Neural Networks 62
    4.7.1 Activation Functions 62
    4.7.2 Neurons 64
    4.7.3 Neural Networks 65
    4.7.4 Logistic Regression and Neural Networks 73
  4.8 Classification Trees 74
    4.8.1 Classification of Data by Leaves (Terminal Nodes) 74
    4.8.2 Impurity of Nodes and Trees 75
    4.8.3 Growing Trees 76
    4.8.4 Pruning Trees 79
    4.8.5 Regression Trees 81
  4.9 Support Vector Machines 81
    4.9.1 Support Vector Machine Classifiers 81
    4.9.2 Kernelization 88
    4.9.3 Proximal Support Vector Machine Classifiers 92
  4.10 Postscript: Example Problem Revisited 93

5 Bias–Variance Trade-off 97
  5.1 Squared-Error Loss 98
  5.2 Arbitrary Loss 101

6 Combining Classifiers 107
  6.1 Ensembles 107
  6.2 Ensemble Design 110
  6.3 Bootstrap Aggregation (Bagging) 112
CONTENTS

6.4 Bumping 115
6.5 Random Forests 116
6.6 Boosting 118
6.7 Arcing 121
6.8 Stacking and Mixture of Experts 121

7 Risk Estimation and Model Selection 127
7.1 Risk Estimation via Training Data 128
7.2 Risk Estimation via Validation or Test Data 128
  7.2.1 Training, Validation, and Test Data 128
  7.2.2 Risk Estimation 129
  7.2.3 Size of Training, Validation, and Test Sets 130
  7.2.4 Testing Hypotheses About Risk 131
  7.2.5 Example of Use of Training, Validation, and Test Sets 132
7.3 Cross-Validation 133
7.4 Improvements on Cross-Validation 135
7.5 Out-of-Bag Risk Estimation 137
7.6 Akaike’s Information Criterion 138
7.7 Schwartz’s Bayesian Information Criterion 138
7.8 Rissanen’s Minimum Description Length Criterion 139
7.9 $R^2$ and Adjusted $R^2$ 140
7.10 Stepwise Model Selection 141
7.11 Occam’s Razor 142

8 Consistency 143
8.1 Convergence of Sequences of Random Variables 144
8.2 Consistency for Parameter Estimation 144
8.3 Consistency for Prediction 145
8.4 There Are Consistent and Universally Consistent Classifiers 145
8.5 Convergence to Asymptopia Is Not Uniform and May Be Slow 147

9 Clustering 149
9.1 Gaussian Mixture Models 150
9.2 $k$-Means 150
9.3 Clustering by Mode-Hunting in a Density Estimate 151
9.4 Using Classifiers to Cluster 152
9.5 Dissimilarity 153
9.6 $k$-Medoids 153
9.7 Agglomerative Hierarchical Clustering 154
9.8 Divisive Hierarchical Clustering 155
9.9 How Many Clusters Are There? Interpretation of Clustering 155
9.10 An Impossibility Theorem 157
CONTENTS

10 Optimization 159

10.1 Quasi-Newton Methods 160

10.1.1 Newton’s Method for Finding Zeros 160

10.1.2 Newton’s Method for Optimization 161

10.1.3 Gradient Descent 161

10.1.4 The BFGS Algorithm 162

10.1.5 Modifications to Quasi-Newton Methods 162

10.1.6 Gradients for Logistic Regression and Neural Networks 163

10.2 The Nelder–Mead Algorithm 166

10.3 Simulated Annealing 168

10.4 Genetic Algorithms 168

10.5 Particle Swarm Optimization 169

10.6 General Remarks on Optimization 170

10.6.1 Imperfectly Known Objective Functions 170

10.6.2 Objective Functions Which Are Sums 171

10.6.3 Optimization from Multiple Starting Points 172

10.7 The Expectation-Maximization Algorithm 173

10.7.1 The General Algorithm 173

10.7.2 EM Climbs the Marginal Likelihood of the Observations 173

10.7.3 Example—Fitting a Gaussian Mixture Model Via EM 176

10.7.4 Example—The Expectation Step 177

10.7.5 Example—The Maximization Step 178

11 High-Dimensional Data 179

11.1 The Curse of Dimensionality 180

11.2 Two Running Examples 187

11.2.1 Example 1: Equilateral Simplex 187

11.2.2 Example 2: Text 187

11.3 Reducing Dimension While Preserving Information 190

11.3.1 The Geometry of Means and Covariances of Real Features 190

11.3.2 Principal Component Analysis 192

11.3.3 Working in “Dissimilarity Space” 193

11.3.4 Linear Multidimensional Scaling 195

11.3.5 The Singular Value Decomposition and Low-Rank Approximation 197

11.3.6 Stress-Minimizing Multidimensional Scaling 199

11.3.7 Projection Pursuit 199

11.3.8 Feature Selection 201

11.3.9 Clustering 202
CONTENTS

11.3.10 Manifold Learning 202
11.3.11 Autoencoders 205
11.4 Model Regularization 209
  11.4.1 Duality and the Geometry of Parameter Penalization 212
  11.4.2 Parameter Penalization as Prior Information 213

12 Communication with Clients 217
  12.1 Binary Classification and Hypothesis Testing 218
  12.2 Terminology for Binary Decisions 219
  12.3 ROC Curves 219
  12.4 One-Dimensional Measures of Performance 224
  12.5 Confusion Matrices 225
  12.6 Multiple Testing
    12.6.1 Control the Familywise Error 226
    12.6.2 Control the False Discovery Rate 227
  12.7 Expert Systems 228

13 Current Challenges in Machine Learning 231
  13.1 Streaming Data 231
  13.2 Distributed Data 231
  13.3 Semi-supervised Learning 232
  13.4 Active Learning 232
  13.5 Feature Construction via Deep Neural Networks 233
  13.6 Transfer Learning 233
  13.7 Interpretability of Complex Models 233

14 R Source Code 235
  14.1 Author’s Biases 236
  14.2 Libraries 236
  14.3 The Running Example (Section 4.3) 237
  14.4 The Bayes Classifier (Section 4.1) 241
  14.5 Quadratic Discriminant Analysis (Section 4.4.1) 243
  14.6 Linear Discriminant Analysis (Section 4.4.2) 243
  14.7 Gaussian Mixture Models (Section 4.4.3) 244
  14.8 Kernel Density Estimation (Section 4.4.4) 245
  14.9 Histograms (Section 4.4.5) 248
  14.10 The Naive Bayes Classifier (Section 4.4.6) 253
  14.11 $k$-Nearest-Neighbor (Section 4.5.1) 255
  14.12 Learning Vector Quantization (Section 4.5.4) 257
  14.13 Logistic Regression (Section 4.6) 259
  14.14 Neural Networks (Section 4.7) 260
  14.15 Classification Trees (Section 4.8) 263
CONTENTS

14.16 Support Vector Machines (Section 4.9) 267
14.17 Bootstrap Aggregation (Section 6.3) 272
14.18 Boosting (Section 6.6) 274
14.19 Arcing (Section 6.7) 275
14.20 Random Forests (Section 6.5) 275

A List of Symbols 277
B Solutions to Selected Exercises 279
C Converting Between Normal Parameters and Level-Curve Ellipsoids 299
D Training Data and Fitted Parameters 301

References 305
Index 315