CONTENTS IN DETAIL

ACKNOWLEDGMENTS xvii

INTRODUCTION xix

0.1 What Is ML? xix
0.2 The Role of Math in ML Theory and Practice xx
0.3 Why Another ML Book? xx
0.4 Recurring Special Sections xxi
0.5 Background Needed xxi
0.6 The qe*-Series Software xxii
0.7 The Book's Grand Plan xxii
0.8 One More Point xxiii

PART I
PROLOGUE, AND NEIGHBORHOOD-BASED METHODS

1
REGRESSION MODELS 3

1.1 Example: The Bike Sharing Dataset 4
  1.1.1 Loading the Data 5
  1.1.2 A Look Ahead 6
1.2 Machine Learning and Prediction 6
  1.2.1 Predicting Past, Present, and Future 6
  1.2.2 Statistics vs. Machine Learning in Prediction 6
1.3 Introducing the k-Nearest Neighbors Method 7
  1.3.1 Predicting Bike Ridership with k-NN 7
1.4 Dummy Variables and Categorical Variables 9
1.5 Analysis with qeKNN() 10
  1.5.1 Predicting Bike Ridership with qeKNN() 11
1.6 The Regression Function: The Basis of ML 13
1.7 The Bias-Variance Trade-off 15
  1.7.1 Analogy to Election Polls 15
  1.7.2 Back to ML 15
1.8 Example: The mlb Dataset 16
1.9 k-NN and Categorical Features 17
1.10 Scaling 18

1.11 Choosing Hyperparameters 19
1.11.1 Predicting the Training Data 20
1.12 Holdout Sets 21
1.12.1 Loss Functions 21
1.12.2 Holdout Sets in the qe*-Series 21
1.12.3 Motivating Cross-Validation 22
1.12.4 Hyperparameters, Dataset Size, and Number of Features 24
1.13 Pitfall: p-Hacking and Hyperparameter Selection 24
1.14 Pitfall: Long-Term Time Trends 25
1.15 Pitfall: Dirty Data 26
1.16 Pitfall: Missing Data 27
1.17 Direct Access to the regtools k-NN Code 28
1.18 Conclusions 29

2
CLASSIFICATION MODELS 31
2.1 Classification Is a Special Case of Regression 32
2.2 Example: The Telco Churn Dataset 33
2.2.1 Pitfall: Factor Data Read as Non-factor 34
2.2.2 Pitfall: Retaining Useless Features 35
2.2.3 Dealing with NA Values 35
2.2.4 Applying the k-Nearest Neighbors Method 36
2.2.5 Pitfall: Overfitting Due to Features with Many Categories 37
2.3 Example: Vertebrae Data 38
2.3.1 Analysis 38
2.4 Pitfall: Error Rate Improves Only Slightly Using the Features 39
2.5 The Confusion Matrix 41
2.6 Clearing the Confusion: Unbalanced Data 41
2.6.1 Example: The Kaggle Appointments Dataset 42
2.6.2 A Better Approach to Unbalanced Data 44
2.7 Receiver Operating Characteristic and Area Under Curve 46
2.7.1 Details of ROC and AUC 46
2.7.2 The qeROC() Function 47
2.7.3 Example: Telco Churn Data 47
2.7.4 Example: Vertebrae Data 48
2.7.5 Pitfall: Overreliance on AUC 48
2.8 Conclusions 49

3
BIAS, VARIANCE, OVERFITTING, AND CROSS-VALIDATION 51
3.1 Overfitting and Underfitting 52
3.1.1 Intuition Regarding the Number of Features and Overfitting .. 53
3.1.2 Relation to Overall Dataset Size 53
3.1.3 Well Then, What Are the Best Values of k and p? 54

3.2 Cross-Validation 55
3.2.1 K-Fold Cross-Validation 55
3.2.2 Using the replicMeans() Function 56
3.2.3 Example: Programmer and Engineer Data 56
3.2.4 Triple Cross-Validation 58
3.3 Conclusions 59

4
DEALING WITH LARGE NUMBERS OF FEATURES 61

4.1 Pitfall: Computational Issues in Large Datasets 61
4.2 Introduction to Dimension Reduction 62
4.2.1 Example: The Million Song Dataset 63
4.2.2 The Need for Dimension Reduction 64
4.3 Methods for Dimension Reduction 65
4.3.1 Consolidation and Embedding 65
4.3.2 The All Possible Subsets Method 66
4.3.3 Principal Components Analysis 66
4.3.4 But Now We Have Two Hyperparameters 70
4.3.5 Using the qePCA() Wrapper 71
4.3.6 PCs and the Bias-Variance Trade-off 73
4.4 The Curse of Dimensionality 74
4.5 Other Methods of Dimension Reduction 75
4.5.1 Feature Ordering by Conditional Independence 75
4.5.2 Uniform Manifold Approximation and Projection 77
4.6 Going Further Computationally 77
4.7 Conclusions 77

PART II
TREE-BASED METHODS

5
A STEP BEYOND K-NN: DECISION TREES 81

5.1 Basics of Decision Trees 81
5.2 The qeDT() Function 82
5.2.1 Looking at the Plot 83
5.3 Example: New York City Taxi Data 85
5.3.1 Pitfall: Too Many Combinations of Factor Levels 86
5.3.2 Tree-Based Analysis 86
5.4 Example: Forest Cover Data 88
5.5 Decision Tree Hyperparameters: How to Split? 90
5.6 Hyperparameters in the qeDT() Function 91
5.7 Conclusions 93

6 TWEAKING THE TREES 95

6.1 Bias vs. Variance, Bagging, and Boosting 96
6.2 Bagging: Generating New Trees by Resampling 96
6.2.1 Random Forests 97
6.2.2 The qeRF() Function 97
6.2.3 Example: Vertebrae Data 98
6.2.4 Example: Remote-Sensing Soil Analysis 99
6.3 Boosting: Repeatedly Tweaking a Tree 100
6.3.1 Implementation: AdaBoost 101
6.3.2 Gradient Boosting 102
6.3.3 Example: Call Network Monitoring 103
6.3.4 Example: Vertebrae Data 105
6.3.5 Bias vs. Variance in Boosting 106
6.3.6 Computational Speed 106
6.3.7 Further Hyperparameters 106
6.3.8 The Learning Rate 106
6.4 Pitfall: No Free Lunch 109

7 FINDING A GOOD SET OF HYPERPARAMETERS 111

7.1 Combinations of Hyperparameters 111
7.2 Grid Searching with qeFT() 112
7.2.1 How to Call qeFT() 112
7.3 Example: Programmer and Engineer Data 113
7.3.1 Confidence Intervals 114
7.3.2 The Takeaway on Grid Searching 116
7.4 Example: Programmer and Engineer Data 116
7.5 Example: Phoneme Data 117
7.6 Conclusions 119

PART III METHODS BASED ON LINEAR RELATIONSHIPS

8 PARAMETRIC METHODS 123

8.1 Motivating Example: The Baseball Player Data 124
8.1.1 A Graph to Guide Our Intuition 124
8.1.2 View as Dimension Reduction 126
8.2 The lm() Function 126
8.3 Wrapper for lm() in the qe*-Series: qeLin() 127

8.4 Use of Multiple Features 127
8.4.1 Example: Baseball Player, Continued 127
8.4.2 Beta Notation 128
8.4.3 Example: Airbnb Data 128
8.4.4 Applying the Linear Model 130
8.5 Dimension Reduction 130
8.5.1 Which Features Are Important? 131
8.5.2 Statistical Significance and Dimension Reduction 131
8.6 Least Squares and Residuals 135
8.7 Diagnostics: Is the Linear Model Valid? 136
8.7.1 Exactness? 137
8.7.2 Diagnostic Methods 137
8.8 The R-Squared Value(s) 137
8.9 Classification Applications: The Logistic Model 138
8.9.1 The glm() and qeLogit() Functions 139
8.9.2 Example: Telco Churn Data 139
8.9.3 Multiclass Case 140
8.9.4 Example: Fall Detection Data 141
8.10 Bias and Variance in Linear/Generalized Linear Models 142
8.10.1 Example: Bike Sharing Data 143
8.11 Polynomial Models 144
8.11.1 Motivation 144
8.11.2 Modeling Nonlinearity with a Linear Model 145
8.11.3 Polynomial Logistic Regression 147
8.11.4 Example: Programmer and Engineer Wages 147
8.12 Blending the Linear Model with Other Methods 148
8.13 The qeCompare() Function 149
8.13.1 Need for Caution Regarding Polynomial Models 149
8.14 What's Next 150

9
CUTTING THINGS DOWN TO SIZE: REGULARIZATION 151

9.1 Motivation 151
9.2 Size of a Vector 152
9.3 Ridge Regression and the LASSO 153
9.3.1 How They Work 153
9.3.2 The Bias-Variance Trade-off, Avoiding Overfitting 154
9.3.3 Relation Between λ, n, and p 154
9.3.4 Comparison, Ridge vs. LASSO 154
9.4 Software 155
9.5 Example: NYC Taxi Data 155
9.6 Example: Airbnb Data 157
9.7 Example: African Soil Data 159
9.7.1 LASSO Analysis 159
9.8 Optional Section: The Famous LASSO Picture 161
9.9 Coming Up 162

PART IV
METHODS BASED ON SEPARATING LINES AND PLANES

10
A BOUNDARY APPROACH: SUPPORT VECTOR MACHINES 165

10.1 Motivation 166
10.1.1 Example: The Forest Cover Dataset 166
10.2 Lines, Planes, and Hyperplanes 170
10.3 Math Notation 170
10.3.1 Vector Expressions 170
10.3.2 Dot Products 171
10.3.3 SVM as a Parametric Model 172
10.4 SVM: The Basic Ideas—Separable Case 172
10.4.1 Example: The Anderson Iris Dataset 173
10.4.2 Optimizing Criterion 174
10.5 Major Problem: Lack of Linear Separability 177
10.5.1 Applying a "Kernel" 178
10.5.2 Soft Margin 181
10.6 Example: Forest Cover Data 182
10.7 And What About That Kernel Trick? 183
10.8 "Warning: Maximum Number of Iterations Reached" 183
10.9 Summary 184

11
LINEAR MODELS ON STEROIDS: NEURAL NETWORKS 185

11.1 Overview 186
11.2 Working on Top of a Complex Infrastructure 187
11.3 Example: Vertebrae Data 188
11.4 Neural Network Hyperparameters 189
11.5 Activation Functions 189
11.6 Regularization 190
11.6.1 L1 and L2 Regularization 191
11.6.2 Regularization by Dropout 191
11.7 Example: Fall Detection Data 191
11.8 Pitfall: Convergence Problems 193
11.9 Close Relation to Polynomial Regression 194
11.10 Bias vs. Variance in Neural Networks 195
11.11 Discussion 195

PART V
APPLICATIONS

12
IMAGE CLASSIFICATION 199

12.1 Example: The Fashion MNIST Data . . . 200
12.1.1 A First Try Using a Logit Model . . . 200
12.1.2 Refinement via PCA . . . 201
12.2 Convolutional Models . . . 201
12.2.1 Need for Recognition of Locality . . . 201
12.2.2 Overview of Convolutional Methods . . . 202
12.2.3 Image Tiling . . . 203
12.2.4 The Convolution Operation . . . 204
12.2.5 The Pooling Operation . . . 206
12.2.6 Shape Evolution Across Layers . . . 207
12.2.7 Dropout . . . 207
12.2.8 Summary of Shape Evolution . . . 208
12.2.9 Translation Invariance . . . 209
12.3 Tricks of the Trade . . . 209
12.3.1 Data Augmentation . . . 209
12.3.2 Pretrained Networks . . . 210
12.4 So, What About the Overfitting Issue? . . . 210
12.5 Conclusions . . . 210

13
HANDLING TIME SERIES AND TEXT DATA 211

13.1 Converting Time Series Data to Rectangular Form . . . 212
13.1.1 Toy Example . . . 212
13.1.2 The regtools Function TStoX() . . . 213
13.2 The qeTS() Function . . . 214
13.3 Example: Weather Data . . . 214
13.4 Bias vs. Variance . . . 216
13.5 Text Applications . . . 217
13.5.1 The Bag-of-Words Model . . . 217
13.5.2 The qeText() Function . . . 218
13.5.3 Example: Quiz Data . . . 218
13.5.4 Example: AG News Dataset . . . 219
13.6 Summary . . . 220

A
LIST OF ACRONYMS AND SYMBOLS **221**

B
STATISTICS AND ML TERMINOLOGY CORRESPONDENCE **223**

C
MATRICES, DATA FRAMES, AND FACTOR CONVERSIONS **225**

C.1 Matrices 225
C.2 Conversions: Between R Factors and Dummy Variables, Between Data Frames and Matrices 226

D
PITFALL: BEWARE OF "P-HACKING"! **229**

INDEX **233**