Table of Contents

Foreword p. xiii

Preface p. xv

Acknowledgments p. xvii

1 Introduction p. 1

1.1 What Is Data Mining? p. 1

1.2 Where Is Data Mining Used? p. 2

1.3 The Origins of Data Mining p. 2

1.4 The Rapid Growth of Data Mining p. 3

1.5 Why Are There So Many Different Methods? p. 4

1.6 Terminology and Notation p. 4

1.7 Road Maps to This Book p. 6

2 Overview of the Data Mining Process p. 9

2.1 Introduction p. 9

2.2 Core Ideas in Data Mining p. 9

2.3 Supervised and Unsupervised Learning p. 11

2.4 The Steps in Data Mining p. 11

2.5 Preliminary Steps p. 13

2.6 Building a Model: Example with Linear Regression p. 21

2.7 Using Excel for Data Mining p. 27

Problems p. 31

3 Data Exploration and Dimension Reduction p. 35

3.1 Introduction p. 35

3.2 Practical Considerations p. 35

Example 1 House Prices in Boston p. 36

3.3 Data Summaries p. 37

3.4 Data Visualization p. 38

3.5 Correlation Analysis p. 40

3.6 Reducing the Number of Categories in Categorical Variables p. 41

3.7 Principal Components Analysis p. 41

Example 2 Breakfast Cereals p. 42

Principal Components p. 45

Normalizing the Data p. 46

Using Principal Components for Classification and Prediction p. 49

Problems p. 51

4 Evaluating Classification and Predictive Performance p. 53

4.1 Introduction p. 53

4.2 Judging Classification Performance p. 53

Accuracy Measures p. 53

Cutoff for Classification p. 56

Performance in Unequal Importance of Classes p. 60

Asymmetric Misclassification Costs p. 61

Oversampling and Asymmetric Costs p. 66

Classification Using a Triage Strategy p. 72

4.3 Evaluating Predictive Performance p. 72

Problems p. 74

5 Multiple Linear Regression p. 75

5.1 Introduction p. 75

5.2 Explanatory vs. Predictive Modeling p. 76

5.3 Estimating the Regression Equation and Prediction p. 76

Example: Predicting the Price of Used Toyota Corolla Automobiles p. 77

5.4 Variable Selection in Linear Regression p. 81

Reducing the Number of Predictors p. 81

How to Reduce the Number of Predictors p. 82

Problems p. 86

6 Three Simple Classification Methods p. 91

6.1 Introduction p. 91

Example 1 Predicting Fraudulent Financial Reporting p. 91

Example 2 Predicting Delayed Flights p. 92

6.2 The Naive Rule p. 92

6.3 Naive Bayes p. 93

Conditional Probabilities and Pivot Tables p. 94

A Practical Difficulty p. 94

A Solution: Naive Bayes p. 95

Advantages and Shortcomings of the naive Bayes Classifier p. 100

6.4 k-Nearest Neighbors p. 103

Example 3 Riding Mowers p. 104

Choosing k p. 105

k-NN for a Quantitative Response p. 106

Advantages and Shortcomings of k-NN Algorithms p. 106

Problems p. 108

7 Classification and Regression Trees p. 111

7.1 Introduction p. 111

7.2 Classification Trees p. 113

7.3 Recursive Partitioning   p. 113

7.4 Example 1: Riding Mowers   p. 113

Measures of Impurity   p. 115

7.5 Evaluating the Performance of a Classification Tree   p. 120

Example 2 Acceptance of Personal Loan   p. 120

7.6 Avoiding Overfitting   p. 121

Stopping Tree Growth: CHAID   p. 121

Pruning the Tree   p. 125

7.7 Classification Rules from Trees   p. 130

7.8 Regression Trees   p. 130

Prediction   p. 130

Measuring Impurity   p. 131

Evaluating Performance   p. 132

7.9 Advantages, Weaknesses, and Extensions   p. 132

Problems   p. 134

8 Logistic Regression   p. 137

8.1 Introduction   p. 137

8.2 The Logistic Regression Model   p. 138

Example: Acceptance of Personal Loan   p. 139

Model with a Single Predictor   p. 141

Estimating the Logistic Model from Data: Computing Parameter Estimates   p. 143

Interpreting Results in Terms of Odds   p. 144

8.3 Why Linear Regression Is Inappropriate for a Categorical Response   p. 146

8.4 Evaluating Classification Performance   p. 148

Variable Selection   p. 148

8.5 Evaluating Goodness of Fit p. 150

8.6 Example of Complete Analysis: Predicting Delayed Flights p. 153

Data Preprocessing p. 154

Model Fitting and Estimation p. 155

Model Interpretation p. 155

Model Performance p. 155

Goodness of fit p. 157

Variable Selection p. 158

8.7 Logistic Regression for More Than Two Classes p. 160

Ordinal Classes p. 160

Nominal Classes p. 161

Problems p. 163

9 Neural Nets p. 167

9.1 Introduction p. 167

9.2 Concept and Structure of a Neural Network p. 168

9.3 Fitting a Network to Data p. 168

Example 1 Tiny Dataset p. 169

Computing Output of Nodes p. 170

Preprocessing the Data p. 172

Training the Model p. 172

Example 2 Classifying Accident Severity p. 176

Avoiding overfitting p. 177

Using the Output for Prediction and Classification p. 181

9.4 Required User Input p. 181

9.5 Exploring the Relationship Between Predictors and Response p. 182

9.6 Advantages and Weaknesses of Neural Networks p. 182

Problems p. 184

10 Discriminant Analysis p. 187

10.1 Introduction p. 187

10.2 Example 1: Riding Mowers p. 187

10.3 Example 2: Personal Loan Acceptance p. 188

10.4 Distance of an Observation from a Class p. 188

10.5 Fisher's Linear Classification Functions p. 191

10.6 Classification Performance of Discriminant Analysis p. 194

10.7 Prior Probabilities p. 195

10.8 Unequal Misclassification Costs p. 195

10.9 Classifying More Than Two Classes p. 196

Example 3 Medical Dispatch to Accident Scenes p. 196

10.10 Advantages and Weaknesses p. 197

Problems p. 200

11 Association Rules p. 203

11.1 Introduction p. 203

11.2 Discovering Association Rules in Transaction Databases p. 203

11.3 Example 1: Synthetic Data on Purchases of Phone Faceplates p. 204

11.4 Generating Candidate Rules p. 204

The Apriori Algorithm p. 205

11.5 Selecting Strong Rules p. 206

Support and Confidence p. 206

Lift Ratio p. 207

Data Format p. 207

The Process of Rule Selection p. 209

Interpreting the Results p. 210

Statistical Significance of Rules p. 211

11.6 Example 2: Rules for Similar Book Purchases p. 212

11.7 Summary p. 212

Problems p. 215

12 Cluster Analysis p. 219

12.1 Introduction p. 219

12.2 Example: Public Utilities p. 220

12.3 Measuring Distance Between Two Records p. 222

Euclidean Distance p. 223

Normalizing Numerical Measurements p. 223

Other Distance Measures for Numerical Data p. 223

Distance Measures for Categorical Data p. 226

Distance Measures for Mixed Data p. 226

12.4 Measuring Distance Between Two Clusters p. 227

12.5 Hierarchical (Agglomerative) Clustering p. 228

Minimum Distance (Single Linkage) p. 229

Maximum Distance (Complete Linkage) p. 229

Group Average (Average Linkage) p. 230

Dendrograms: Displaying Clustering Process and Results p. 230

Validating Clusters p. 231

Limitations of Hierarchical Clustering p. 232

12.6 Nonhierarchical Clustering: The k-Means Algorithm p. 233

Initial Partition into k Clusters p. 234

Problems p. 237

13 Cases p. 241

13.1 Charles Book Club p. 241

13.2 German Credit p. 250

13.3 Tayko Software Cataloger p. 254

13.4 Segmenting Consumers of Bath Soap p. 258

13.5 Direct-Mail Fundraising p. 262

13.6 Catalog Cross-Selling p. 265

13.7 Predicting Bankruptcy p. 267

References p. 271

Index p. 273