



Contents

Foreword	xvii
Preface to the second edition	xix
Preface to the first edition	xxi
Acknowledgments	xxiii

PART I PRELIMINARIES

CHAPTER 1 Introduction	3
1.1 What Is Data Mining?	3
1.2 Where Is Data Mining Used?	4
1.3 Origins of Data Mining	4
1.4 Rapid Growth of Data Mining	5
1.5 Why Are There So Many Different Methods?	6
1.6 Terminology and Notation	7
1.7 Road Maps to This Book	9
Order of Topics	10
CHAPTER 2 Overview of the Data Mining Process	12
2.1 Introduction	12
2.2 Core Ideas in Data Mining	13
Classification	13
Prediction	13
Association Rules	13
Predictive Analytics	14
Data Reduction	14
Data Exploration	14
Data Visualization	14
2.3 Supervised and Unsupervised Learning	15
2.4 Steps in Data Mining	15
2.5 Preliminary Steps	17

- Organization of Datasets 17
- Sampling from a Database 17
- Oversampling Rare Events 17
- Preprocessing and Cleaning the Data 18
- Use and Creation of Partitions 24
- 2.6 Building a Model: Example with Linear Regression 27
 - Boston Housing Data 27
 - Modeling Process 28
- 2.7 Using Excel for Data Mining 34
- Problems 38

PART II DATA EXPLORATION AND DIMENSION REDUCTION

CHAPTER 3 Data Visualization 43

- 3.1 Uses of Data Visualization 43
- 3.2 Data Examples 45
 - Example 1: Boston Housing Data 45
 - Example 2: Ridership on Amtrak Trains 45
- 3.3 Basic Charts: Bar Charts, Line Graphs, and Scatterplots 45
 - Distribution Plots: Boxplots and Histograms 47
 - Heatmaps: Visualizing Correlations and Missing Values 50
- 3.4 Multidimensional Visualization 52
 - Adding Variables: Color, Size, Shape, Multiple Panels, and Animation 52
 - Manipulations: Rescaling, Aggregation and Hierarchies, Zooming, and Panning, and Filtering 54
 - Reference: Trend Lines and Labels 57
 - Scaling up: Large Datasets 58
 - Multivariate Plot: Parallel Coordinates Plot 59
 - Interactive Visualization 60
- 3.5 Specialized Visualizations 63
 - Visualizing Networked Data 63
 - Visualizing Hierarchical Data: Treemaps 65
 - Visualizing Geographical Data: Map Charts 66
- 3.6 Summary of Major Visualizations and Operations, According to Data Mining Goal 67
 - Prediction 67
 - Classification 67
 - Time Series Forecasting 68
 - Unsupervised Learning 68
- Problems 69

CHAPTER 4	Dimension Reduction	71
4.1	Introduction	71
4.2	Practical Considerations	72
	Example 1: House Prices in Boston	72
4.3	Data Summaries	73
	Summary Statistics	73
	Pivot Tables	75
4.4	Correlation Analysis	76
4.5	Reducing the Number of Categories in Categorical Variables	76
4.6	Converting a Categorical Variable to a Numerical Variable .	78
4.7	Principal Components Analysis	78
	Example 2: Breakfast Cereals	78
	Principal Components	83
	Normalizing the Data	83
	Using Principal Components for Classification and Prediction	87
4.8	Dimension Reduction Using Regression Models	87
4.9	Dimension Reduction Using Classification and Regression Trees	88
	Problems	89

PART III PERFORMANCE EVALUATION

CHAPTER 5	Evaluating Classification and Predictive Performance	93
5.1	Introduction	93
5.2	Judging Classification Performance	94
	Benchmark: The Naive Rule	94
	Class Separation	94
	Classification Matrix	96
	Using the Validation Data	96
	Accuracy Measures	97
	Cutoff for Classification	97
	Performance in Unequal Importance of Classes	100
	Asymmetric Misclassification Costs	105
	Oversampling and Asymmetric Costs	109
	Classification Using a Triage Strategy	114
5.3	Evaluating Predictive Performance	115
	Benchmark: The Average	115
	Prediction Accuracy Measures	115
	Problems	118

PART IV PREDICTION AND CLASSIFICATION METHODS

CHAPTER 6 Multiple Linear Regression 121

- 6.1 Introduction 121
- 6.2 Explanatory versus Predictive Modeling 122
- 6.3 Estimating the Regression Equation and Prediction 123
 - Example: Predicting the Price of Used Toyota Corolla Automobiles 124
- 6.4 Variable Selection in Linear Regression 127
 - Reducing the Number of Predictors 127
 - How to Reduce the Number of Predictors 128
- Problems 133

CHAPTER 7 *k*-Nearest Neighbors (*k*-NN) 137

- 7.1 *k*-NN Classifier (Categorical Outcome) 137
 - Determining Neighbors 138
 - Classification Rule 138
 - Example: Riding Mowers 139
 - Choosing *k* 140
 - Setting the Cutoff Value 141
 - k*-NN with More Than Two Classes 142
- 7.2 *k*-NN for a Numerical Response 142
- 7.3 Advantages and Shortcomings of *k*-NN Algorithms 144
- Problems 146

CHAPTER 8 Naive Bayes 148

- 8.1 Introduction 148
 - Example 1: Predicting Fraudulent Financial Reporting 149
- 8.2 Applying the Full (Exact) Bayesian Classifier 150
 - Practical Difficulty with the Complete (Exact) Bayes Procedure 151
 - Solution: Naive Bayes 152
 - Example 2: Predicting Fraudulent Financial Reports, Two Predictors 153
 - Example 3: Predicting Delayed Flights 155
- 8.3 Advantages and Shortcomings of the Naive Bayes Classifier 159
- Problems 162

CHAPTER 9 Classification and Regression Trees 164

- 9.1 Introduction 164
- 9.2 Classification Trees 166
 - Recursive Partitioning 166
 - Example 1: Riding Mowers 166

9.3	Measures of Impurity	169
	Tree Structure	172
	Classifying a New Observation	173
9.4	Evaluating the Performance of a Classification Tree	173
	Example 2: Acceptance of Personal Loan	174
9.5	Avoiding Overfitting	179
	Stopping Tree Growth: CHAID	179
	Pruning the Tree	180
9.6	Classification Rules from Trees	183
9.7	Classification Trees for More Than Two Classes	185
9.8	Regression Trees	185
	Prediction	186
	Measuring Impurity	187
	Evaluating Performance	187
9.9	Advantages, Weaknesses, and Extensions	187
	Problems	189
CHAPTER 10 Logistic Regression		192
10.1	Introduction	192
10.2	Logistic Regression Model	194
	Example: Acceptance of Personal Loan	196
	Model with a Single Predictor	197
	Estimating the Logistic Model from Data: Computing Parameter Estimates	199
	Interpreting Results in Terms of Odds	201
10.3	Evaluating Classification Performance	202
	Variable Selection	203
	Impact of Single Predictors	205
10.4	Example of Complete Analysis: Predicting Delayed Flights	206
	Data Preprocessing	208
	Model Fitting and Estimation	208
	Model Interpretation	208
	Model Performance	209
	Variable Selection	211
10.5	Appendix: Logistic Regression for Profiling	211
	Appendix A: Why Linear Regression Is Inappropriate for a Categorical Response	212
	Appendix B: Evaluating Goodness of Fit	214
	Appendix C: Logistic Regression for More Than Two Classes	215
	Problems	219

CHAPTER 11 Neural Nets 222

- 11.1 Introduction 222
- 11.2 Concept and Structure of a Neural Network 223
- 11.3 Fitting a Network to Data 223
 - Example 1: Tiny Dataset 224
 - Computing Output of Nodes 225
 - Preprocessing the Data 228
 - Training the Model 228
 - Example 2: Classifying Accident Severity 232
 - Avoiding Overfitting 236
 - Using the Output for Prediction and Classification 237
- 11.4 Required User Input 237
- 11.5 Exploring the Relationship Between Predictors and Response 239
- 11.6 Advantages and Weaknesses of Neural Networks 239
- Problems 241

CHAPTER 12 Discriminant Analysis 243

- 12.1 Introduction 243
 - Example 1: Riding Mowers 244
 - Example 2: Personal Loan Acceptance 244
- 12.2 Distance of an Observation from a Class 246
- 12.3 Fisher’s Linear Classification Functions 247
- 12.4 Classification Performance of Discriminant Analysis 251
- 12.5 Prior Probabilities 252
- 12.6 Unequal Misclassification Costs 252
- 12.7 Classifying More Than Two Classes 253
 - Example 3: Medical Dispatch to Accident Scenes 253
- 12.8 Advantages and Weaknesses 254
- Problems 258

PART V MINING RELATIONSHIPS AMONG RECORDS

CHAPTER 13 Association Rules 263

- 13.1 Introduction 263
- 13.2 Discovering Association Rules in Transaction Databases 263
 - Example 1: Synthetic Data on Purchases of Phone Faceplates 265
- 13.3 Generating Candidate Rules 265
 - The Apriori Algorithm 266
- 13.4 Selecting Strong Rules 267
 - Support and Confidence 267

Lift Ratio	268
Data Format	269
Process of Rule Selection	270
Interpreting the Results	271
Statistical Significance of Rules	272
Example 2: Rules for Similar Book Purchases	274
13.5 Summary	275
Problems	277
CHAPTER 14 Cluster Analysis	279
14.1 Introduction	279
Example: Public Utilities	281
14.2 Measuring Distance Between Two Records	283
Euclidean Distance	283
Normalizing Numerical Measurements	283
Other Distance Measures for Numerical Data	284
Distance Measures for Categorical Data	286
Distance Measures for Mixed Data	287
14.3 Measuring Distance Between Two Clusters	287
14.4 Hierarchical (Agglomerative) Clustering	290
Minimum Distance (Single Linkage)	290
Maximum Distance (Complete Linkage)	291
Average Distance (Average Linkage)	291
Centroid Distance (Average Group Linkage)	291
Ward's Method	291
Dendrograms: Displaying Clustering Process and Results	292
Validating Clusters	293
Limitations of Hierarchical Clustering	295
14.5 Nonhierarchical Clustering: The <i>k</i> -Means Algorithm	295
Initial Partition into <i>k</i> Clusters	297
Problems	300
PART VI FORECASTING TIME SERIES	
CHAPTER 15 Handling Time Series	305
15.1 Introduction	305
15.2 Explanatory versus Predictive Modeling	306
15.3 Popular Forecasting Methods in Business	307
Combining Methods	307
15.4 Time Series Components	308
Example: Ridership on Amtrak Trains	308
15.5 Data Partitioning	312
Problems	314

CHAPTER 16 Regression-Based Forecasting 317

- 16.1 Model with Trend 317
 - Linear Trend 317
 - Exponential Trend 319
 - Polynomial Trend 321
- 16.2 Model with Seasonality 322
- 16.3 Model with Trend and Seasonality 324
- 16.4 Autocorrelation and ARIMA Models 324
 - Computing Autocorrelation 325
 - Improving Forecasts by Integrating Autocorrelation Information 328
 - Evaluating Predictability 331
- Problems 334

CHAPTER 17 Smoothing Methods 344

- 17.1 Introduction 344
- 17.2 Moving Average 345
 - Centered Moving Average for Visualization 345
 - Trailing Moving Average for Forecasting 346
 - Choosing Window Width (w) 350
- 17.3 Simple Exponential Smoothing 350
 - Choosing Smoothing Parameter α 351
 - Relation between Moving Average and Simple Exponential Smoothing 352
- 17.4 Advanced Exponential Smoothing 353
 - Series with a Trend 353
 - Series with a Trend and Seasonality 354
 - Series with Seasonality (No Trend) 354
- Problems 356

PART VII CASES

CHAPTER 18 Cases 367

- 18.1 Charles Book Club 367
 - The Book Industry 367
 - Database Marketing at Charles 368
 - Data Mining Techniques 370
 - Assignment 373
- 18.2 German Credit 375
 - Assignment 379
- 18.3 Tayko Software Cataloger 379

Background	379
The Mailing Experiment	380
Data	380
Assignment	380
18.4 Segmenting Consumers of Bath Soap	383
Business Situation	383
Key Problems	384
Data	384
Measuring Brand Loyalty	386
Assignment	386
Appendix	386
18.5 Direct-Mail Fundraising	387
Background	387
Data	387
Assignment	387
18.6 Catalog Cross Selling	389
Background	389
Assignment	390
18.7 Predicting Bankruptcy	390
Predicting Corporate Bankruptcy	391
Assignment	393
18.8 Time Series Case: Forecasting Public Transportation	
Demand	393
Background	393
Problem Description	393
Available Data	394
Assignment Goal	394
Assignment	394
Tips and Suggested Steps	394
References	397
Index	399

