

CONTENTS

Foreword	xiii
Preface	xv
Acknowledgments	xvii
1 Introduction	1
1.1 What Is Data Mining?	1
1.2 Where Is Data Mining Used?	2
1.3 The Origins of Data Mining	2
1.4 The Rapid Growth of Data Mining	3
1.5 Why are there so many different methods?	4
1.6 Terminology and Notation	6
1.7 Road Maps to This Book	7
2 Overview of the Data Mining Process	9
2.1 Introduction	9
2.2 Core Ideas in Data Mining	9
2.2.1 Classification	9
2.2.2 Prediction	10
2.2.3 Association Rules	10
2.2.4 Predictive Analytics	10
	vii

2.2.5	Data Reduction	10
2.2.6	Data Exploration	10
2.2.7	Data Visualization	11
2.3	Supervised and Unsupervised Learning	11
2.4	The Steps in Data Mining	11
2.5	Preliminary Steps	13
2.5.1	Organization of Datasets	13
2.5.2	Sampling from a Database	13
2.5.3	Oversampling Rare Events	13
2.5.4	Pre-processing and Cleaning the Data	14
2.5.5	Use and Creation of Partitions	19
2.6	Building a Model - An Example with Linear Regression	21
2.6.1	The Boston Housing Data	21
2.6.2	The modeling process	21
2.7	Using Excel For Data Mining	28
	Problems	32
3	Data Exploration and Dimension Reduction	35
3.1	Introduction	35
3.2	Practical Considerations	35
3.3	Data Summaries	36
3.4	Data Visualization	39
3.5	Correlation Analysis	42
3.6	Reducing the Number of Categories in Categorical Variables	42
3.7	Principal Components Analysis	42
3.7.1	Example 2: Breakfast Cereals	42
3.7.2	The Principal Components	47
3.7.3	Normalizing the Data	48
3.7.4	Using Principal Components for Classification and Prediction	50
	Problems	52
4	Evaluating Classification and Predictive Performance	55
4.1	Introduction	55
4.2	Judging Classification Performance	55
4.2.1	Accuracy Measures	55
4.2.2	Cutoff For Classification	59
4.2.3	Performance in Unequal Importance of Classes	62
4.2.4	Asymmetric Misclassification Costs	66
4.2.5	Oversampling and Asymmetric Costs	70
4.2.6	Classification Using a Triage Strategy	75
4.3	Evaluating Predictive Performance	76

Problems	78
5 Multiple Linear Regression	81
5.1 Introduction	81
5.2 Explanatory Vs. Predictive Modeling	82
5.3 Estimating the Regression Equation and Prediction	82
5.3.1 Example: Predicting the Price of Used Toyota Corolla Automobiles	83
5.4 Variable Selection in Linear Regression	87
5.4.1 Reducing the Number of Predictors	87
5.4.2 How to Reduce the Number of Predictors	88
Problems	92
6 Three Simple Classification Methods	97
6.1 Introduction	97
6.1.1 Example 1: Predicting Fraudulent Financial Reporting	97
6.1.2 Example 2: Predicting Delayed Flights	98
6.2 The Naive Rule	99
6.3 Naive Bayes	99
6.3.1 Conditional Probabilities and Pivot Tables	100
6.3.2 A Practical Difficulty	100
6.3.3 A Solution: Naive Bayes	101
6.3.4 Advantages and Shortcomings of the Naive Bayes Classifier	106
6.4 k -Nearest Neighbors (k -NN)	110
6.4.1 Example 3: Riding Mowers	110
6.4.2 Choosing k	111
6.4.3 k -NN for a Quantitative Response	112
6.4.4 Advantages and Shortcomings of k -NN Algorithms	112
Problems	115
7 Classification and Regression Trees	117
7.1 Introduction	117
7.2 Classification Trees	119
7.3 Recursive Partitioning	119
7.4 Example 1: Riding Mowers	119
7.4.1 Measures of Impurity	121
7.5 Evaluating the Performance of a Classification Tree	126
7.5.1 Example 2: Acceptance of Personal Loan	126
7.6 Avoiding Overfitting	129
7.6.1 Stopping Tree Growth: CHAID	129
7.6.2 Pruning the Tree	131
7.7 Classification Rules from Trees	136

7.8	Regression Trees	137
7.8.1	Prediction	137
7.8.2	Measuring Impurity	137
7.8.3	Evaluating Performance	139
7.9	Advantages, Weaknesses, and Extensions	139
	Problems	141
8	Logistic Regression	145
8.1	Introduction	145
8.2	The Logistic Regression Model	146
8.2.1	Example: Acceptance of Personal Loan	149
8.2.2	A Model with a Single Predictor	149
8.2.3	Estimating the Logistic Model From Data: Computing Parameter Estimates	151
8.2.4	Interpreting Results in Terms of Odds	153
8.3	Why Linear Regression is Inappropriate for a Categorical Response	154
8.4	Evaluating Classification Performance	156
8.4.1	Variable Selection	158
8.5	Evaluating Goodness-of-Fit	158
8.6	Example of Complete Analysis: Predicting Delayed Flights	159
8.6.1	Data preprocessing	162
8.6.2	Model fitting and estimation	162
8.6.3	Model Interpretation	162
8.6.4	Model Performance	164
8.6.5	Goodness-of-fit	164
8.6.6	Variable Selection	166
8.7	Logistic Regression for More than 2 Classes	168
8.7.1	Ordinal Classes	168
8.7.2	Nominal Classes	169
	Problems	171
9	Neural Nets	175
9.1	Introduction	175
9.2	Concept and Structure of a Neural Network	176
9.3	Fitting a Network to Data	177
9.3.1	Example 1: Tiny Dataset	177
9.3.2	Computing Output of Nodes	177
9.3.3	Preprocessing the Data	180
9.3.4	Training the Model	180
9.3.5	Example 2: Classifying Accident Severity	184
9.3.6	Avoiding overfitting	186

9.3.7	Using the Output for Prediction and Classification	186
9.4	Required User Input	190
9.5	Exploring the Relationship Between Predictors and Response	191
9.6	Advantages and Weaknesses of Neural Networks	191
	Problems	193
10	Discriminant Analysis	195
10.1	Introduction	195
10.2	Example 1: Riding Mowers	195
10.3	Example 2: Personal Loan Acceptance	196
10.4	Distance of an Observation from a Class	198
10.5	Fisher's Linear Classification Functions	199
10.6	Classification Performance of Discriminant Analysis	200
10.7	Prior Probabilities	204
10.8	Unequal Misclassification Costs	204
10.9	Classifying More Than Two Classes	205
10.9.1	Example 3: Medical Dispatch to Accident Scenes	205
10.10	Advantages and Weaknesses	207
	Problems	209
11	Association Rules	213
11.1	Introduction	213
11.2	Discovering Association Rules in Transaction Databases	213
11.3	Example 1: Synthetic Data on Purchases of Phone Faceplates	214
11.4	Generating Candidate Rules	215
11.4.1	The Apriori Algorithm	216
11.5	Selecting Strong Rules	216
11.5.1	Support and Confidence	216
11.5.2	Lift Ratio	217
11.5.3	Data Format	217
11.5.4	The Process of Rule Selection	219
11.5.5	Interpreting the Results	220
11.5.6	Statistical Significance of Rules	220
11.6	Example 2: Rules for Similar Book Purchases	222
11.7	Summary	223
	Problems	225
12	Cluster Analysis	229
12.1	Introduction	229
12.2	Example: Public Utilities	230
12.3	Measuring Distance Between Two Records	233

12.3.1	Euclidean Distance	233
12.3.2	Normalizing Numerical Measurements	233
12.3.3	Other Distance Measures for Numerical Data	234
12.3.4	Distance Measures for Categorical Data	237
12.3.5	Distance Measures for Mixed Data	237
12.4	Measuring Distance Between Two Clusters	237
12.5	Hierarchical (Agglomerative) Clustering	239
12.5.1	Minimum Distance (Single Linkage)	240
12.5.2	Maximum Distance (Complete Linkage)	240
12.5.3	Group Average (Average Linkage)	240
12.5.4	Dendrograms: Displaying Clustering Process and Results	240
12.5.5	Validating Clusters	242
12.5.6	Limitations of Hierarchical Clustering	243
12.6	Non-Hierarchical Clustering: The k -Means Algorithm	243
12.6.1	Initial Partition Into k Clusters	245
	Problems	248
13	Cases	251
13.1	Charles Book Club	251
13.2	German Credit	260
13.3	Tayko Software Cataloger	266
13.4	Segmenting Consumers of Bath Soap	271
13.5	Direct Mail Fundraising	275
13.6	Catalog Cross-Selling	278
13.7	Predicting Bankruptcy	280
	References	283
	Index	285