

# CONTENTS

Acknowledgments · xi

## PART 1. CONCEPTS

1. What Is Data Mining? · 3
  - The Goals of This Book* · 6
  - Software and Hardware for Data Mining* · 7
  - Basic Terminology* · 8
2. Contrasts with the Conventional Statistical Approach · 13
  - Predictive Power in Conventional Statistical Modeling* · 13
  - Hypothesis Testing in the Conventional Approach* · 15
  - Heteroscedasticity as a Threat to Validity in Conventional Modeling* · 17
  - The Challenge of Complex and Nonrandom Samples* · 20
  - Bootstrapping and Permutation Tests* · 20
  - Nonlinearity in Conventional Predictive Models* · 24
  - Statistical Interactions in Conventional Models* · 25
  - Conclusion* · 27
3. Some General Strategies Used in Data Mining · 30
  - Cross-Validation* · 30
  - Overfitting* · 32
  - Boosting* · 35

	Calibrating	·	38
	Measuring Fit: The Confusion Matrix and ROC Curves	·	39
	Identifying Statistical Interactions and Effect Heterogeneity in Data Mining	·	43
	Bagging and Random Forests	·	45
	The Limits of Prediction	·	48
	Big Data Is Never Big Enough	·	50
2	4. Important Stages in a Data Mining Project	·	53
	When to Sample Big Data	·	53
	Building a Rich Array of Features	·	54
	Feature Selection	·	56
	Feature Extraction	·	56
	Constructing a Model	·	58

## PART 2. WORKED EXAMPLES

	5. Preparing Training and Test Datasets	·	63
	The Logic of Cross-Validation	·	63
	Cross-Validation Methods: An Overview	·	65
	6. Variable Selection Tools	·	72
	Stepwise Regression	·	73
	The LASSO	·	79
	VIF Regression	·	86
	7. Creating New Variables Using Binning and Trees	·	93
	Discretizing a Continuous Predictor	·	95
	Continuous Outcomes and Continuous Predictors	·	100
	Binning Categorical Predictors	·	105
	Using Partition Trees to Study Interactions	·	108
	8. Extracting Variables	·	116
	Principal Component Analysis	·	116
	Independent Component Analysis	·	125
	9. Classifiers	·	133
	K-Nearest Neighbors	·	134
	Naive Bayes	·	142
	Support Vector Machines	·	147
	Optimizing Prediction across Multiple Classifiers	·	156
	10. Classification Trees	·	162
	Partition Trees	·	162
	Boosted Trees and Random Forests	·	172

11.	Neural Networks	·	185
12.	Clustering	·	196
	<i>Hierarchical Clustering</i>	·	199
	<i>K-Means Clustering</i>	·	203
	<i>Normal Mixtures</i>	·	208
	<i>Self-Organized Maps</i>	·	212
13.	Latent Class Analysis and Mixture Models	·	216
	<i>Latent Class Analysis</i>	·	216
	<i>Latent Class Regression</i>	·	221
	<i>Mixture Models</i>	·	223
14.	Association Rules	·	227
	Conclusion	·	235
	Bibliography	·	239
	Notes	·	245
	Index	·	247