DATA MINING FOR BUSINESS INTELLIGENCE

Concepts, Techniques, and Applications in Microsoft Office Excel® with XLMiner®

GALIT SHMUELI
University of Maryland

NITIN R. PATEL
MIT

PETER C. BRUCE
statistics.com

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

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 4

1.7 Road Maps to This Book 6

2 Overview of the Data Mining Process 9

2.1 Introduction 9

2.2 Core Ideas in Data Mining 9

2.3 Supervised and Unsupervised Learning 11

2.4 The Steps in Data Mining 11

2.5 Preliminary Steps 13

2.6 Building a Model: Example with Linear Regression 21

2.7 Using Excel for Data Mining 27
Problems 31

3 Data Exploration and Dimension Reduction 35

3.1 Introduction 35
3.2 Practical Considerations 35
Example 1: House Prices in Boston 36
3.3 Data Summaries 37
3.4 Data Visualization 38
3.5 Correlation Analysis 40
3.6 Reducing the Number of Categories in Categorical Variables 41
3.7 Principal Components Analysis 41
Example 2: Breakfast Cereals 42
Principal Components 45
Normalizing the Data 46
Using Principal Components for Classification and Prediction 49
Problems 51

4 Evaluating Classification and Predictive Performance 53

4.1 Introduction 53
4.2 Judging Classification Performance 53
Accuracy Measures 53
Cutoff for Classification 56
Performance in Unequal Importance of Classes 60
Asymmetric Misclassification Costs 61
Oversampling and Asymmetric Costs 66
Classification Using a Triage Strategy 72
4.3 Evaluating Predictive Performance 72
Problems 74

5 Multiple Linear Regression 75

5.1 Introduction 75
5.2 Explanatory vs. Predictive Modeling 76
5.3 Estimating the Regression Equation and Prediction 76
Example: Predicting the Price of Used Toyota Corolla Automobiles 77
5.4 Variable Selection in Linear Regression 81
Reducing the Number of Predictors 81
How to Reduce the Number of Predictors 82
Problems 86

6 Three Simple Classification Methods 91

6.1 Introduction 91
Example 1: Predicting Fraudulent Financial Reporting 91
Example 2: Predicting Delayed Flights 92
6.2 The Naive Rule 92
6.3 Naive Bayes 93
Conditional Probabilities and Pivot Tables 94
A Practical Difficulty 94
A Solution: Naive Bayes 95
Advantages and Shortcomings of the naive Bayes Classifier 100
6.4 k-Nearest Neighbors 103
Example 3: Riding Mowers 104
Choosing k 105
k-NN for a Quantitative Response 106
Advantages and Shortcomings of k-NN Algorithms 106
Problems 108

7 Classification and Regression Trees 111

7.1 Introduction 111
7.2 Classification Trees 113
7.3 Recursive Partitioning 113
7.4 Example 1: Riding Mowers 113
Measures of Impurity 115
7.5 Evaluating the Performance of a Classification Tree 120
Example 2: Acceptance of Personal Loan 120
7.6 Avoiding Overfitting 121
Stopping Tree Growth: CHAID 121
Pruning the Tree 125
7.7 Classification Rules from Trees 130
7.8 Regression Trees 130
Prediction 130
Measuring Impurity 131
Evaluating Performance 132
7.9 Advantages, Weaknesses, and Extensions 132
Problems 134

8 Logistic Regression 137

8.1 Introduction 137
8.2 The Logistic Regression Model 138
Example: Acceptance of Personal Loan 139
Model with a Single Predictor 141

Estimating the Logistic Model from Data: Computing Parameter Estimates 143
Interpreting Results in Terms of Odds 144
8.3 Why Linear Regression Is Inappropriate for a Categorical Response 146
8.4 Evaluating Classification Performance 148
Variable Selection 148
8.5 Evaluating Goodness of Fit 150
8.6 Example of Complete Analysis: Predicting Delayed Flights 153
Data Preprocessing 154
Model Fitting and Estimation 155
Model Interpretation 155
Model Performance 155
Goodness of fit 157
Variable Selection 158
8.7 Logistic Regression for More Than Two Classes 160
Ordinal Classes 160
Nominal Classes 161
Problems 163

9 Neural Nets 167
9.1 Introduction 167
9.2 Concept and Structure of a Neural Network 168
9.3 Fitting a Network to Data 168
Example 1: Tiny Dataset 169
Computing Output of Nodes 170
Preprocessing the Data 172
Training the Model 172
Example 2: Classifying Accident Severity 176
Avoiding overfitting 177
Using the Output for Prediction and Classification 181
9.4 Required User Input 181
9.5 Exploring the Relationship Between Predictors and Response 182
9.6 Advantages and Weaknesses of Neural Networks 182
Problems 184

10 Discriminant Analysis 187
10.1 Introduction 187
10.2 Example 1: Riding Mowers 187
10.3 Example 2: Personal Loan Acceptance 188
10.4 Distance of an Observation from a Class 188
10.5 Fisher's Linear Classification Functions 191

10.6 Classification Performance of Discriminant Analysis 194
10.7 Prior Probabilities 195
10.8 Unequal Misclassification Costs 195
10.9 Classifying More Than Two Classes 196
Example 3: Medical Dispatch to Accident Scenes 196
10.10 Advantages and Weaknesses 197
Problems 200

11 Association Rules 203

11.1 Introduction 203
11.2 Discovering Association Rules in Transaction Databases 203
11.3 Example 1: Synthetic Data on Purchases of Phone Faceplates 204
11.4 Generating Candidate Rules 204
The Apriori Algorithm 205
11.5 Selecting Strong Rules 206
Support and Confidence 206
Lift Ratio 207
Data Format 207
The Process of Rule Selection 209
Interpreting the Results 210
Statistical Significance of Rules 211
11.6 Example 2: Rules for Similar Book Purchases 212
11.7 Summary 212
Problems 215

12 Cluster Analysis 219

12.1 Introduction 219
12.2 Example: Public Utilities 220
12.3 Measuring Distance Between Two Records 222
Euclidean Distance 223
Normalizing Numerical Measurements 223
Other Distance Measures for Numerical Data 223
Distance Measures for Categorical Data 226
Distance Measures for Mixed Data 226
12.4 Measuring Distance Between Two Clusters 227
12.5 Hierarchical (Agglomerative) Clustering 228
Minimum Distance (Single Linkage) 229
Maximum Distance (Complete Linkage) 229
Group Average (Average Linkage) 230
Dendrograms: Displaying Clustering Process and Results 230
Validating Clusters 231

Limitations of Hierarchical Clustering 232
12.6 Nonhierarchical Clustering: The k-Means Algorithm 233
Initial Partition into k Clusters 234
Problems 237

13 Cases 241
13.1 Charles Book Club 241
13.2 German Credit 250
13.3 Tayko Software Cataloger 254
13.4 Segmenting Consumers of Bath Soap 258
13.5 Direct-Mail Fundraising 262
13.6 Catalog Cross-Selling 265
13.7 Predicting Bankruptcy 267

References 271

Index 273