Data Mining

A TUTORIAL-BASED PRIMER

Richard J. Roiger Michael W. Geatz

Contents

Preface *iii*
Acknowledgments *xiv*

Part I: Data Mining Fundamentals 1

Chapter 1 Data Mining: A First View 3

1.1. Data Mining: A Definition 4
1.2 What Can Computers Learn? 5
Three Concept Views 6
Supervised Learning 7
Supervised Learning: A Decision Tree Example 9
Unsupervised Clustering 11
1.3. Is Data Mining Appropriate for My Problem? 14
Data Mining or Data Query? 14
Data Mining vs. Data Query: An Example 16
1.4 Expert Systems or Data Mining? 17
1.5 A Simple Data Mining Process Model 19
Assembling the Data 19
The Data Warehouse 20
Relational Databases and Flat Files 20
Mining the Data 21
Interpreting the Results 21
Result Application 21
1.6 Why Not Simple Search? 22
1.7 Data Mining Applications 23
Example Applications 24

Customer Intrinsic Value 25
1.8 Chapter Summary 27
1.9 Key Terms 28
1.10 Exercises 30

Chapter 2 Data Mining: A Closer Look 33

2.1 Data Mining Strategies 34
Classification 34
Estimation 36
Prediction 36
Unsupervised Clustering 40
Market Basket Analysis 41
2.2 Supervised Data Mining Techniques 42
The Credit Card Promotion Database 42
Production Rules 43
Neural Networks 43
Statistical Regression 47
2.3 Association Rules 49
2.4 Clustering Techniques 50
2.5 Evaluating Performance 52
Evaluating Supervised Learner Models 52
Two Class Error Analysis 54
Evaluating Numeric Output 55
Comparing Models by Measuring Lift 56
Unsupervised Model Evaluation 58
2.6 Chapter Summary 59
2.7 Key Terms 60
2.8 Exercises 62

Chapter 3 Basic Data Mining Techniques 67

3.1 Decision Trees 68
An Algorithm for Building Decision Trees 68
Decision Trees for the Credit Card Promotion Database 73
Decision Tree Rules 76
Other Methods for Building Decision Trees 77
General Considerations 77
3.2 Generating Association Rules 78
Confidence and Support 78
Mining Association Rules: An Example 80
General Considerations 83
3.3 The K-Means Algorithm 84
An Example Using K-Means 84
General Considerations 88

3.4 Genetic Learning 89

Genetic Algorithms and Supervised Learning 90

Genetic Algorithms and Unsupervised Clustering 95

General Consideration 97

3.5 Choosing a Data Mining Technique 98

3.6 Chapter Summary 100

3.7 Key Terms 101

3.8 Exercises 102

Chapter 4 An Excel-Based Data Mining Tool 105

4.1 The iData Analyzer 106

Installing iDA 107

Limitations 110

4.2 ESX: A Multipurpose Tool for Data Mining 110

4.3 iDAV Format for Data Mining 112

4.4 A Five-Step Approach for Unsupervised Clustering 114

Applying the Five-Step Approach for Unsupervised Clustering 114

Step 1: Enter the Data to be Mined 114

Step 2: Perform a Data Mining Session 115

Step 3: Read and Interpret Summary Results 117

Step 4: Read and Interpret Individual Class Results 121

Step 5: Visualize Individual Class Rules 124

4.5 A Six-Step Approach for Supervised Learning 126

Applying the Six-Step Approach for Supervised Learning 127

Step 1: Choose an Output Attribute 127

Step 2: Perform the Mining Session 127

Step 3: Read and Interpret Summary Results 127

Step 4: Read and Interpret Test Set Results 128

Step 5: Read and Interpret Results for Individual Classes 129

Step 6: Visualize and Interpret Class Rules 130

4.6 Techniques for Generating Rules 131

4.7 Instance Typicality 133

4.8 Special Considerations and Features 135

Avoid Mining Delays 135

The Quick Mine Feature 136

Applying the Quick Mine Feature to the Titanic Dataset 136

Erroneous and Missing Data 138

4.9 Chapter Summary 139

4.10 Key Terms 139

4.11 Exercises 140

Part II: Tools for Knowledge Discovery 145

Chapter 5 Knowledge Discovery in Databases 147

5.1 A KDD Process Model 148
5.2 Step 1: Goal Identification 150
5.3 Step 2: Creating a Target Dataset 151
5.4 Step 3: Data Preprocessing 153
Noisy Data 153
Missing Data 154
5.5 Step 4: Data Transformation 155
Data Normalization 155
Data Type Conversion 156
Attribute and Instance Selection 157
5.6 Step 5: Data Mining 161
5.7 Step 6: Interpretation and Evaluation 161
5.8 Step 7: Taking Action 162
5.9 The CRISP-DM Process Model 163
5.10 Experimenting with ESX 164
A Four-Step Process Model for Knowledge Discovery 164
Experiment 1: Attribute Evaluation 165
Applying the Four-Step Process Model to the Credit Card Screening Dataset 166
Experiment 2: Parameter Evaluation 169
Applying the Four-Step Process Model to the Satellite Image Dataset 169
5.11 Chapter Summary 172
5.12 Key Terms 173
5.13 Exercises 174

Chapter 6 The Data Warehouse 179

6.1 Operational Databases 180
Data Modeling and Normalization 180
The Relational Model 182
6.2 Data Warehouse Design 184
Entering Data into the Warehouse 185
Structuring the Data Warehouse: The Star Schema 186
Decision Support: Analyzing the Warehouse Data 190
6.3 On-line Analytical Processing 192
OLAP: An Example 194
General Considerations 197
6.4 Excel Pivot Tables for Data Analysis 197
Creating a Simple Pivot Table 198
Pivot Tables for Hypothesis Testing 201

Creating a Multidimensional Pivot Table 205
6.5 Chapter Summary 208
6.6 Key Terms 209
6.7 Exercises 210

Chapter 7 Formal Evaluation Techniques 213
7.1 What Should Be Evaluated? 214
7.2 Tools for Evaluation 216
Single-Valued Summary Statistics 217
The Normal Distribution 217
Normal Distributions and Sample Means 219
A Classical Model for Hypothesis Testing 221
7.3 Computing Test Set Confidence Intervals 222
7.4 Comparing Supervised Learner Models 225
7.5 Attribute Evaluation 227
Locating Redundant Attributes with Excel 228
Computing Attribute Correaltions with MS Excel 228
Creating a Scatterplot Diagram with MS Excel 231
Hypothesis Testing for Numerical Attribute Significance 232
7.6 Unsupervised Evaluation Techniques 232
Unsupervised Clustering for Supervised Evaluation 233
Supervised Evaluation for Unsupervised Clustering 234
Additional Methods for Evaluating an Unsupervised Clustering 234
7.7 Evaluating Supervised Models with Numeric Output 235
7.8 Chapter Summary 236
7.9 Key Terms 238
7.10 Exercises 239

Part III: Advanced Data Mining Techniques 243

Chapter 8 Neural Networks 245
8.1 Feed-Forward Neural Networks 246
Neural Network Input Format 247
Neural Network Output Format 248
The Sigmoid Function 249
8.2 Neural Network Training: A Conceptual View 250
Supervised Learning with Feed-Forward Networks 251
Unsupervised Clustering with Self-Organizing Maps 253
8.3 Neural Network Explanation 254
8.4 General Considerations 256
8.5 Neural Network Training: A Detailed Overview 257
The Backpropagation Algorithms: An Example 257

Kohonen Self-Organizing Maps: An Example 260
8.6 Chapter Summary 262
8.7 Key Terms 263
8.8 Exercises 264

Chapter 9 Building Neural Networks with iDA 265
9.1 A Four-Step Approach for Backpropagation Learning 266
Example 1: Modeling the Exclusive-OR Function 266
Applying the Four-Step Approach to the XOR Function 268
Example 2: The Satellite Image Dataset 274
Applying the Four-Step Approach to the Satellite Image Dataset 275
General Considerations 278
9.2 A Four-Step Approach for Neural Network Clustering 278
An Example: The Deer Hunter Dataset 278
Applying the Four-Step Approach to the Deer Hunter Dataset 279
General Considerations 283
9.3 ESX for Neural Network Cluster Analysis 284
Applying Supervised Learning to Evaluate Neural Network Clustering 284
9.4 Chapter Summary 285
9.5 Key Terms 286
9.6 Exercises 286

Chapter 10 Statistical Techniques 291
10.1 Linear Regression Analysis 292
Simple Linear Regression 292
Multiple Linear Regression with Excel 293
Applying the LINEST Function 293
Regression Trees 297
10.2 Logistic Regression 298
Transforming the Linear Regression Model 299
Logistic Regression: An Example 300
10.3 Bayes Classifier 302
Bayes Classifier: An Example 303
Zero-Valued Attribute Counts 305
Missing Data 306
Numeric Data 306
10.4 Clustering Algorithms 308
Agglomerative Clustering 309
Conceptual Clustering 312
Expectation Maximization 315
10.5 Heuristics or Statistics? 318
Query and Visualization Techniques 318

Machine Learning and Statistical Techniques 319
10.6 Chapter Summary 320
10.7 Key Terms 321
10.8 Exercises 322

Chapter 11 Specialized Techniques 327
11.1 Time-Series Analysis 328
An Example with Linear Regression 329
A Neural Network Example 331
Categorical Attribute Prediction 332
General Considerations 333
11.2 Mining the Web 334
Web-Based Mining: General Issues 335
Data Mining for Web Site Evaluation 339
Data Mining for Personalization 340
Data Mining for Web Site Adaptation 341
11.3 Mining Textual Data 342
11.4 Improving Performance 343
Bagging 343
Boosting 344
Instance Typicality 345
11.5 Chapter Summary 347
11.6 Key Terms 348
11.7 Exercises 349

Part IV: Intelligent Systems **CD-1

Chapter 12 Rule-Based Systems CD-3
12.1 Exploring Artificial Intelligence CD-4
12.2 Problem Solving as a State Space Search CD-7
The Water Jug Problem CD-8
Backward Chaining CD-12
Data Mining and State Space Search CD-16
12.3 Expert Systems CD-18
The Components of an Expert System CD-19
Separating Knowledge and Reasoning CD-19
Developing an Expert System CD-21
12.4 Structuring a Rule-Based System CD-23
Example 1: Form 1040 Tax Dependency CD-24
Example 2: Choosing a Data Mining Technique CD-29
12.5 Chapter Summary CD-32

12.6 Key Terms CD-33
12.7 Exercises CD-34

Chapter 13 Managing Uncertainty in Rule-Based Systems CD-37
13.1 Uncertainty: Sources and Solutions CD-38
Sources of Uncertainty CD-38
General Methods for Dealing with Uncertainty CD-39
13.2 Fuzzy Rule-Based Systems CD-40
Fuzzy Sets CD-41
Fuzzy Reasoning: An Example CD-42
13.3 A Probability-Based Approach to Uncertainty CD-48
Multiple Evidence with Bayes Theorem CD-50
Likelihood Ratios: Necessity and Sufficiency CD-52
13.4 Chapter Summary CD-55
13.5 Key Terms CD-56
13.6 Exercises CD-57

Chapter 14 Intelligent Agents CD-59
14.1 Characteristics of Intelligent Agents CD-60
14.2 Types of Agents CD-61
14.3 Integrating Data Mining, Expert Systems, and Intelligent Agents CD-62
14.4 Chapter Summary CD-64
14.5 Key Terms CD-64
14.6 Exercises CD-65

*** Part IV (Chapters 12, 13, and 14) are available on the CD that accompanies this book.*

Appendixes

Appendix A The iDA Software A-1
A.1 Software Installation A-2
Installing iDA A-2
Installing Java A-3
Reinstalling iDA A-3
A.2 Uninstalling iDA A-3
A.3 Software Limitations A-4
A.4 General Guidelines for Software Usage A-4
A.5 Troubleshooting A-5
Cannot Initiate a Data Mining Session A-5

Error Message Says There Is a Column Containing Data but There Is No Attribute Name A-5

Running a Data Mining Session Causes an Illegal Operation Error or Excel Closes A-5

Summary Statistics Are Given for Real-Valued Attributes Declared as Unused A-6

A.6 Software Support A-6

Appendix B Datasets for Data Mining B-1

B.1 The iDA Dataset Package B-2

B.2 Web Sites Containing Datasets for Mining B-6

DASL B-7

MLnet Online Information Service B-7

KDNuggets B-7

UCI Machine Learning Repository B-7

JSE Data Archive B-8

Appendix C Decision Tree Attribute Selection **CD-67

Appendix D Statistics for Performance Evaluation **CD-71

D.1 Single-Valued Summary Statistics CD-72

D.2 The Normal Distribution CD-72

D.3 Comparing Supervised Learner Models CD-73

Comparing Models with Independent Test Data CD-73

Pairwise Comparison with a Single Test Dataset CD-74

D.4 Confidence Intervals for Numeric Output CD-75

D.5 Comparing Models with Numeric Output CD-76

Appendix E Excel Pivot Tables: Office 97 **CD-79

Creating a Simple Pivot Table CD-80

Pivot Tables for Hypothesis Testing CD-83

Creating a Multidimensional Pivot Table CD-86

** *Appendices C, D, and E are available on the CD that accompanies this book.*

Bibliography BIB-1

Index I-1