Data Mining

Practical Machine Learning Tools and Techniques with Java Implementations

Ian H. Witten

Department of Computer Science
University of Waikato

Eibe Frank

Department of Computer Science
University of Waikato

MORGAN KAUFMANN PUBLISHERS
AN IMPRINT OF ACADEMIC PRESS
A Harcourt Science and Technology Company
SAN FRANCISCO SAN DIEGO NEW YORK BOSTON
LONDON SYDNEY TOKYO

Technische Universität Darmstadt
FACHBEREICH INFORMATIK
BIBLIOTHEK
Inventar-Nr.: T02-00384
Sachgebiete:
Standort:

Contents

Foreword v

Preface xix

1 What's it all about? 1

1.1 **Data mining and machine learning** 2
Describing structural patterns 4
Machine learning 5
Data mining 7

1.2 **Simple examples: The weather problem and others** 8
The weather problem 8
Contact lenses: An idealized problem 11
Irises: A classic numeric dataset 13
CPU performance: Introducing numeric prediction 15
Labor negotiations: A more realistic example 16
Soybean classification: A classic machine learning success 17

1.3 **Fielded applications** 20
Decisions involving judgment 21
Screening images 22
Load forecasting 23
Diagnosis 24
Marketing and sales 25

1.4 **Machine learning and statistics** 26

1.5 **Generalization as search** 27
Enumerating the concept space 28
Bias 29

1.6 Data mining and ethics 32

1.7 Further reading 34

2 Input: Concepts, instances, attributes 37

2.1 What's a concept? 38

2.2 What's in an example? 41

2.3 What's in an attribute? 45

2.4 Preparing the input 48

Gathering the data together 48

ARFF format 49

Attribute types 51

Missing values 52

Inaccurate values 53

Getting to know your data 54

2.5 Further reading 55

3 Output: Knowledge representation 57

3.1 Decision tables 58

✓ 3.2 Decision trees 58

✓ 3.3 Classification rules 59

✓ 3.4 Association rules 63

✓ 3.5 Rules with exceptions 64

3.6 Rules involving relations 67

3.7 Trees for numeric prediction 70

✓ 3.8 Instance-based representation 72

3.9 Clusters 75

3.10 Further reading 76

4 Algorithms: The basic methods 77

4.1 Inferring rudimentary rules 78

Missing values and numeric attributes 80

Discussion 81

✓ 4.2 Statistical modeling 82

Missing values and numeric attributes 85

Discussion 88

✓ **4.3 Divide and conquer: Constructing decision trees 89**
Calculating information 93
Highly branching attributes 94
Discussion 97

✓ **4.4 Covering algorithms: Constructing rules 97**
Rules versus trees 98
A simple covering algorithm 98
Rules versus decision lists 103

4.5 Mining association rules 104
Item sets 105
Association rules 105
Generating rules efficiently 108
Discussion 111

4.6 Linear models 112
Numeric prediction 112
Classification 113
Discussion 113

✓ **4.7 Instance-based learning 114**
The distance function 114
Discussion 115

4.8 Further reading 116

5 Credibility: Evaluating what's been learned 119

✓ **5.1 Training and testing 120**

✓ **5.2 Predicting performance 123**

✓ **5.3 Cross-validation 125**

✓ **5.4 Other estimates 127**
Leave-one-out 127
The bootstrap 128

✓ **5.5 Comparing data mining schemes 129**

5.6 Predicting probabilities 133
Quadratic loss function 134
Informational loss function 135
Discussion 136

✓ **5.7 Counting the cost 137**
Lift charts 139
ROC curves 141

Cost-sensitive learning 144
Discussion 145
5.8 **Evaluating numeric prediction 147**
5.9 **The minimum description length principle 150**
5.10 **Applying MDL to clustering 154**
5.11 **Further reading 155**

6 Implementations: Real machine learning schemes 157

✓ 6.1 **Decision trees 159**
Numeric attributes 159
Missing values 161
Pruning 162
Estimating error rates 164
Complexity of decision tree induction 167
From trees to rules 168
C4.5: Choices and options 169
Discussion 169
✓ 6.2 **Classification rules 170**
Criteria for choosing tests 171
Missing values, numeric attributes 172
Good rules and bad rules 173
Generating good rules 174
Generating good decision lists 175
Probability measure for rule evaluation 177
Evaluating rules using a test set 178
Obtaining rules from partial decision trees 181
Rules with exceptions 184
Discussion 187
6.3 **Extending linear classification: Support vector machines 188**
The maximum margin hyperplane 189
Nonlinear class boundaries 191
Discussion 193
✓ 6.4 **Instance-based learning 193**
Reducing the number of exemplars 194
Pruning noisy exemplars 194
Weighting attributes 195
Generalizing exemplars 196
Distance functions for generalized exemplars 197

Generalized distance functions 199
Discussion 200
6.5 **Numeric prediction** **201**
Model trees 202
Building the tree 202
Pruning the tree 203
Nominal attributes 204
Missing values 204
Pseudo-code for model tree induction 205
Locally weighted linear regression 208
Discussion 209
6.6 **Clustering** **210**
Iterative distance-based clustering 211
Incremental clustering 212
Category utility 217
Probability-based clustering 218
The EM algorithm 221
Extending the mixture model 223
Bayesian clustering 225
Discussion 226
7 Moving on: Engineering the input and output 229
7.1 **Attribute selection** **232**
Scheme-independent selection 233
Searching the attribute space 235
Scheme-specific selection 236
7.2 **Discretizing numeric attributes** **238**
Unsupervised discretization 239
Entropy-based discretization 240
Other discretization methods 243
Entropy-based versus error-based discretization 244
Converting discrete to numeric attributes 246
7.3 **Automatic data cleansing** **247**
Improving decision trees 247
Robust regression 248
Detecting anomalies 249
7.4 **Combining multiple models** **250**
Bagging 251
Boosting 254

Stacking *258*
Error-correcting output codes *260*
7.5 Further reading 263

8 Nuts and bolts: Machine learning algorithms in Java 265

8.1 Getting started 267
8.2 Javadoc and the class library 271
Classes, instances, and packages *272*
The weka.core package *272*
The weka.classifiers package *274*
Other packages *276*
Indexes *277*
8.3 Processing datasets using the machine learning programs 277
Using 'M5' *277*
Generic options *279*
Scheme-specific options *282*
Classifiers *283*
Meta-learning schemes *286*
Filters *289*
Association rules *294*
Clustering *296*
8.4 Embedded machine learning 297
A simple message classifier *299*
8.5 Writing new learning schemes 306
An example classifier *306*
Conventions for implementing classifiers *314*
Writing filters *314*
An example filter *316*
Conventions for writing filters *317*

9 Looking forward 321

9.1 Learning from massive datasets 322
9.2 Visualizing machine learning 325
Visualizing the input *325*
Visualizing the output *327*
9.3 Incorporating domain knowledge 329

9.4 **Text mining** 331
Finding key phrases for documents 331
Finding information in running text 333
Soft parsing 334
9.5 **Mining the World Wide Web** 335
9.6 **Further reading** 336

References 339

Index 351

About the authors 371