Contents

Foreword v

Preface xxiii

Updated and revised content xxvii

Acknowledgments xxix

Part I Machine learning tools and techniques 1

1 What's it all about? 3

1.1 Data mining and machine learning 4

Describing structural patterns 6

Machine learning 7

Data mining 9

1.2 Simple examples: The weather problem and others 9

The weather problem 10

Contact lenses: An idealized problem 13

Irises: A classic numeric dataset 15

CPU performance: Introducing numeric prediction 16

Labor negotiations: A more realistic example 17

Soybean classification: A classic machine learning success 18

1.3 Fielded applications 22

Decisions involving judgment 22

Screening images 23

Load forecasting 24

Diagnosis 25

Marketing and sales 26

Other applications 28

1.4 Machine learning and statistics 29

1.5 Generalization as search 30

Enumerating the concept space 31

Bias 32

1.6 Data mining and ethics 35

1.7 Further reading 37

2 Input: Concepts, instances, and attributes 41

2.1 What's a concept? 42

2.2 What's in an example? 45

2.3 What's in an attribute? 49

2.4 Preparing the input 52

Gathering the data together 52

ARFF format 53

Sparse data 55

Attribute types 56

Missing values 58

Inaccurate values 59

Getting to know your data 60

2.5 Further reading 60

3 Output: Knowledge representation 61

3.1 Decision tables 62

3.2 Decision trees 62

3.3 Classification rules 65

3.4 Association rules 69

3.5 Rules with exceptions 70

3.6 Rules involving relations 73

3.7 Trees for numeric prediction 76

3.8 Instance-based representation 76

3.9 Clusters 81

3.10 Further reading 82

4 Algorithms: The basic methods 83

4.1 Inferring rudimentary rules 84
Missing values and numeric attributes 86
Discussion 88

4.2 Statistical modeling 88
Missing values and numeric attributes 92
Bayesian models for document classification 94
Discussion 96

4.3 Divide-and-conquer: Constructing decision trees 97
Calculating information 100
Highly branching attributes 102
Discussion 105

4.4 Covering algorithms: Constructing rules 105
Rules versus trees 107
A simple covering algorithm 107
Rules versus decision lists 111

4.5 Mining association rules 112
Item sets 113
Association rules 113
Generating rules efficiently 117
Discussion 118

4.6 Linear models 119
Numeric prediction: Linear regression 119
Linear classification: Logistic regression 121
Linear classification using the perceptron 124
Linear classification using Winnow 126

4.7 Instance-based learning 128
The distance function 128
Finding nearest neighbors efficiently 129
Discussion 135

4.8 Clustering 136
Iterative distance-based clustering 137
Faster distance calculations 138
Discussion 139

4.9 Further reading 139

5 Credibility: Evaluating what's been learned 143

5.1 Training and testing 144

5.2 Predicting performance 146

5.3 Cross-validation 149

5.4 Other estimates 151
Leave-one-out 151
The bootstrap 152

5.5 Comparing data mining methods 153

5.6 Predicting probabilities 157
Quadratic loss function 158
Informational loss function 159
Discussion 160

5.7 Counting the cost 161
Cost-sensitive classification 164
Cost-sensitive learning 165
Lift charts 166
ROC curves 168
Recall–precision curves 171
Discussion 172
Cost curves 173

5.8 Evaluating numeric prediction 176

5.9 The minimum description length principle 179

5.10 Applying the MDL principle to clustering 183

5.11 Further reading 184

6 Implementations: Real machine learning schemes 187

6.1 Decision trees 189
Numeric attributes 189
Missing values 191
Pruning 192
Estimating error rates 193
Complexity of decision tree induction 196
From trees to rules 198
C4.5: Choices and options 198
Discussion 199

6.2 Classification rules 200
Criteria for choosing tests 200
Missing values, numeric attributes 201

Generating good rules 202
Using global optimization 205
Obtaining rules from partial decision trees 207
Rules with exceptions 210
Discussion 213

6.3 Extending linear models 214
The maximum margin hyperplane 215
Nonlinear class boundaries 217
Support vector regression 219
The kernel perceptron 222
Multilayer perceptrons 223
Discussion 235

6.4 Instance-based learning 235
Reducing the number of exemplars 236
Pruning noisy exemplars 236
Weighting attributes 237
Generalizing exemplars 238
Distance functions for generalized exemplars 239
Generalized distance functions 241
Discussion 242

6.5 Numeric prediction 243
Model trees 244
Building the tree 245
Pruning the tree 245
Nominal attributes 246
Missing values 246
Pseudocode for model tree induction 247
Rules from model trees 250
Locally weighted linear regression 251
Discussion 253

6.6 Clustering 254
Choosing the number of clusters 254
Incremental clustering 255
Category utility 260
Probability-based clustering 262
The EM algorithm 265
Extending the mixture model 266
Bayesian clustering 268
Discussion 270

6.7 Bayesian networks 271
Making predictions 272
Learning Bayesian networks 276

Specific algorithms 278
Data structures for fast learning 280
Discussion 283

7 Transformations: Engineering the input and output 285

7.1 Attribute selection 288
Scheme-independent selection 290
Searching the attribute space 292
Scheme-specific selection 294

7.2 Discretizing numeric attributes 296
Unsupervised discretization 297
Entropy-based discretization 298
Other discretization methods 302
Entropy-based versus error-based discretization 302
Converting discrete to numeric attributes 304

7.3 Some useful transformations 305
Principal components analysis 306
Random projections 309
Text to attribute vectors 309
Time series 311

7.4 Automatic data cleansing 312
Improving decision trees 312
Robust regression 313
Detecting anomalies 314

7.5 Combining multiple models 315
Bagging 316
Bagging with costs 319
Randomization 320
Boosting 321
Additive regression 325
Additive logistic regression 327
Option trees 328
Logistic model trees 331
Stacking 332
Error-correcting output codes 334

7.6 Using unlabeled data 337
Clustering for classification 337
Co-training 339
EM and co-training 340

7.7 Further reading 341

8 Moving on: Extensions and applications 345

8.1 **Learning from massive datasets 346**

8.2 **Incorporating domain knowledge 349**

8.3 **Text and Web mining 351**

8.4 **Adversarial situations 356**

8.5 **Ubiquitous data mining 358**

8.6 **Further reading 361**

Part II The Weka machine learning workbench 363

9 Introduction to Weka 365

9.1 **What's in Weka? 366**

9.2 **How do you use it? 367**

9.3 **What else can you do? 368**

9.4 **How do you get it? 368**

10 The Explorer 369

10.1 **Getting started 369**

Preparing the data 370

Loading the data into the Explorer 370

Building a decision tree 373

Examining the output 373

Doing it again 377

Working with models 377

When things go wrong 378

10.2 **Exploring the Explorer 380**

Loading and filtering files 380

Training and testing learning schemes 384

Do it yourself: The User Classifier 388

Using a metalearner 389

Clustering and association rules 391

Attribute selection 392

Visualization 393

10.3 **Filtering algorithms 393**

Unsupervised attribute filters 395

Unsupervised instance filters 400

Supervised filters 401

10.4 **Learning algorithms 403**
Bayesian classifiers 403
Trees 406
Rules 408
Functions 409
Lazy classifiers 413
Miscellaneous classifiers 414
10.5 **Metalearning algorithms 414**
Bagging and randomization 414
Boosting 416
Combining classifiers 417
Cost-sensitive learning 417
Optimizing performance 417
Retargeting classifiers for different tasks 418
10.6 **Clustering algorithms 418**
10.7 **Association-rule learners 419**
10.8 **Attribute selection 420**
Attribute subset evaluators 422
Single-attribute evaluators 422
Search methods 423

11 The Knowledge Flow interface 427
11.1 **Getting started 427**
11.2 **The Knowledge Flow components 430**
11.3 **Configuring and connecting the components 431**
11.4 **Incremental learning 433**

12 The Experimenter 437
12.1 **Getting started 438**
Running an experiment 439
Analyzing the results 440
12.2 **Simple setup 441**
12.3 **Advanced setup 442**
12.4 **The Analyze panel 443**
12.5 **Distributing processing over several machines 445**

13 The command-line interface 449

13.1 Getting started 449

13.2 The structure of Weka 450

Classes, instances, and packages 450

The weka.core package 451

The weka.classifiers package 453

Other packages 455

Javadoc indices 456

13.3 Command-line options 456

Generic options 456

Scheme-specific options 458

14 Embedded machine learning 461

14.1 A simple data mining application 461

14.2 Going through the code 462

main() 462

MessageClassifier() 462

updateData() 468

classifyMessage() 468

15 Writing new learning schemes 471

15.1 An example classifier 471

buildClassifier() 472

makeTree() 472

computeInfoGain() 480

classifyInstance() 480

main() 481

15.2 Conventions for implementing classifiers 483

References 485

Index 505

About the authors 525