



Data Mining

Practical Machine Learning Tools and Techniques,
Second Edition

V

Ian H. Witten

Department of Computer Science
University of Waikato

Eibe Frank

Department of Computer Science
University of Waikato

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

ELSEVIER

NI01-tfrAN K-V, l'N-iANN i'i BI,i'iIFRS IS || IMPRINT Or l l l -J | !TR

MORGAN KAUFMANN PUBLISHERS

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

CONTENTS

1.4	Machine learning and statistics	29
1.5	Generalization as search	30
	<i>Enumerating the concept space</i>	<i>31</i>
	<i>Bias</i>	<i>32</i>
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
	<i>Gathering the data together</i>	<i>52</i>
	<i>ARFF format</i>	<i>53</i>
	<i>Sparse data</i>	<i>55</i>
	<i>Attribute types</i>	<i>56</i>
	<i>Missing values</i>	<i>58</i>
	<i>Inaccurate values</i>	<i>59</i>
	<i>Getting to know your data</i>	<i>60</i>
2.5	Further reading	60
6	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
	<i>Missing values and numeric attributes</i>	86
	<i>Discussion</i>	88
4.2	Statistical modeling	88
	<i>Missing values and numeric attributes</i>	92
	<i>Bayesian models for document classification</i>	94
	<i>Discussion</i>	96
4.3	Divide-and-conquer: Constructing decision trees	97
	<i>Calculating information</i>	100
	<i>Highly branching attributes</i>	102
	<i>Discussion</i>	105
4.4	Covering algorithms: Constructing rules	105
	<i>Rules versus trees</i>	107
	<i>A simple covering algorithm</i>	107
	<i>Rules versus decision lists</i>	111
4.5	Mining association rules	112
	<i>Item sets</i>	113
	<i>Association rules</i>	113
	<i>Generating rules efficiently</i>	117
	<i>Discussion</i>	118
4.6	Linear models	119
	<i>Numeric prediction: Linear regression</i>	119
	<i>Linear classification: Logistic regression</i>	121
	<i>Linear classification using the perceptron</i>	124
	<i>Linear classification using Winnow</i>	126
4.7	Instance-based learning	128
	<i>The distance function</i>	128
	<i>Finding nearest neighbors efficiently</i>	129
	<i>Discussion</i>	135
4.8	Clustering	136
	<i>Iterative distance-based clustering</i>	137
	<i>Faster distance calculations</i>	138
	<i>Discussion</i>	139
4.9	Further reading	139

CONTENTS

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
	<i>Leave-one-out</i>	<i>151</i>
	<i>The bootstrap</i>	<i>152</i>
5.5	Comparing data mining methods	153
5.6	Predicting probabilities	157
	<i>Quadratic loss function</i>	<i>158</i>
	<i>Informational loss function</i>	<i>159</i>
	<i>Discussion</i>	<i>160</i>
5.7	Counting the cost	161
	<i>Cost-sensitive classification</i>	<i>164</i>
	<i>Cost-sensitive learning</i>	<i>165</i>
	<i>Lift charts</i>	<i>166</i>
	<i>ROC curves</i>	<i>168</i>
	<i>Recall-precision curves</i>	<i>171</i>
	<i>Discussion</i>	<i>172</i>
	<i>Cost curves</i>	<i>173</i>
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
	<i>Numeric attributes</i>	<i>189</i>
	<i>Missing values</i>	<i>191</i>
	<i>Pruning</i>	<i>192</i>
	<i>Estimating error rates</i>	<i>193</i>
	<i>Complexity of decision tree induction</i>	<i>196</i>
	<i>From trees to rules</i>	<i>198</i>
	<i>C4.5: Choices and options</i>	<i>198</i>
	<i>Discussion</i>	<i>199</i>
6.2	Classification rules	200
	<i>Criteria for choosing tests</i>	<i>200</i>
	<i>Missing values, numeric attributes</i>	<i>201</i>

	<i>Generating good rules</i>	202
	<i>Using global optimization</i>	205
	<i>Obtaining rules from partial decision trees</i>	207
	<i>Rules with exceptions</i>	210
	<i>Discussion</i>	213
6.3	Extending linear models	214
	<i>The maximum margin hyperplane</i>	215
	<i>Nonlinear class boundaries</i>	217
	<i>Support vector regression</i>	219
	<i>The kernel perceptron</i>	222
	<i>Multilayer perceptrons</i>	223
	<i>Discussion</i>	235
6.4	Instance-based learning	235
	<i>Reducing the number of exemplars</i>	236
	<i>Pruning noisy exemplars</i>	236
	<i>Weighting attributes</i>	237
	<i>Generalizing exemplars</i>	238
	<i>Distance functions for generalized exemplars</i>	239
	<i>Generalized distance functions</i>	241
	<i>Discussion</i>	242
6.5	Numeric prediction	243
	<i>Model trees</i>	244
	<i>Building the tree</i>	245
	<i>Pruning the tree</i>	245
	<i>Nominal attributes</i>	246
	<i>Missing values</i>	246
	<i>Pseudocode for model tree induction</i>	247
	<i>Rules from model trees</i>	250
	<i>Locally weighted linear regression</i>	251
	<i>Discussion</i>	253
6.6	Clustering	254
	<i>Choosing the number of clusters</i>	254
	<i>Incremental clustering</i>	255
	<i>Category utility</i>	260
	<i>Probability-based clustering</i>	262
	<i>The EM algorithm</i>	265
	<i>Extending the mixture model</i>	266
	<i>Bayesian clustering</i>	268
	<i>Discussion</i>	270
6.7	Bayesian networks	271
	<i>Making predictions</i>	272
	<i>Learning Bayesian networks</i>	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

CONTENTS

10.4	Learning algorithms	403
	<i>Bayesian classifiers</i>	403
	<i>Trees</i>	406
	<i>Rules</i>	408
	<i>Functions</i>	409
	<i>Lazy classifiers</i>	413
	<i>Miscellaneous classifiers</i>	414
10.5	Metalearning algorithms	414
	<i>Bagging and randomization</i>	414
	<i>Boosting</i>	416
	<i>Combining classifiers</i>	417
	<i>Cost-sensitive learning</i>	417
	<i>Optimizing performance</i>	417
	<i>Retargeting classifiers for different tasks</i>	418
10.6	Clustering algorithms	418
10.7	Association-rule learners	419
10.8	Attribute selection	420
	<i>Attribute subset evaluators</i>	422
	<i>Single-attribute evaluators</i>	422
	<i>Search methods</i>	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
	<i>Running an experiment</i>	439
	<i>Analyzing the results</i>	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
	<i>Classes, instances, and packages</i>	<i>450</i>
	<i>The weka.core package</i>	<i>451</i>
	<i>The weka.classifiers package</i>	<i>453</i>
	<i>Other packages</i>	<i>455</i>
	<i>Javadoc indices</i>	<i>456</i>
13.3	Command-line options	456
	<i>Generic options</i>	<i>456</i>
	<i>Scheme-specific options</i>	<i>458</i>
14	Embedded machine learning	461
14.1	A simple data mining application	461
14.2	Going through the code	462
	<i>main()</i>	<i>462</i>
	<i>MessageClassifier()</i>	<i>462</i>
	<i>updateData()</i>	<i>468</i>
	<i>classifyMessage()</i>	<i>468</i>
15	Writing new learning schemes	471
15.1	An example classifier	471
	<i>buildClassifier()</i>	<i>472</i>
	<i>makeTree()</i>	<i>472</i>
	<i>computeInfoGain()</i>	<i>480</i>
	<i>classifyInstance()</i>	<i>480</i>
	<i>main()</i>	<i>481</i>
15.2	Conventions for implementing classifiers	483
	References	485
	Index	505
	About the authors	525