

# 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
  - Iris: 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

<b>4.3</b>	<b>Divide and conquer: Constructing decision trees</b>	<b>89</b>
	<i>Calculating information</i>	93
	<i>Highly branching attributes</i>	94
	<i>Discussion</i>	97
<b>4.4</b>	<b>Covering algorithms: Constructing rules</b>	<b>97</b>
	<i>Rules versus trees</i>	98
	<i>A simple covering algorithm</i>	98
	<i>Rules versus decision lists</i>	103
<b>4.5</b>	<b>Mining association rules</b>	<b>104</b>
	<i>Item sets</i>	105
	<i>Association rules</i>	105
	<i>Generating rules efficiently</i>	108
	<i>Discussion</i>	111
<b>4.6</b>	<b>Linear models</b>	<b>112</b>
	<i>Numeric prediction</i>	112
	<i>Classification</i>	113
	<i>Discussion</i>	113
<b>4.7</b>	<b>Instance-based learning</b>	<b>114</b>
	<i>The distance function</i>	114
	<i>Discussion</i>	115
<b>4.8</b>	<b>Further reading</b>	<b>116</b>
<b>5</b>	<b>Credibility: Evaluating what's been learned</b>	<b>119</b>
<b>5.1</b>	<b>Training and testing</b>	<b>120</b>
<b>5.2</b>	<b>Predicting performance</b>	<b>123</b>
<b>5.3</b>	<b>Cross-validation</b>	<b>125</b>
<b>5.4</b>	<b>Other estimates</b>	<b>127</b>
	<i>Leave-one-out</i>	127
	<i>The bootstrap</i>	128
<b>5.5</b>	<b>Comparing data mining schemes</b>	<b>129</b>
<b>5.6</b>	<b>Predicting probabilities</b>	<b>133</b>
	<i>Quadratic loss function</i>	134
	<i>Informational loss function</i>	135
	<i>Discussion</i>	136
<b>5.7</b>	<b>Counting the cost</b>	<b>137</b>
	<i>Lift charts</i>	139
	<i>ROC curves</i>	141

	<i>Cost-sensitive learning</i>	144
	<i>Discussion</i>	145
5.8	<b>Evaluating numeric prediction</b>	147
5.9	<b>The minimum description length principle</b>	150
5.10	<b>Applying MDL to clustering</b>	154
5.11	<b>Further reading</b>	155
<b>6</b>	<b>Implementations: Real machine learning schemes</b>	<b>157</b>
6.1	<b>Decision trees</b>	159
	<i>Numeric attributes</i>	159
	<i>Missing values</i>	161
	<i>Pruning</i>	162
	<i>Estimating error rates</i>	164
	<i>Complexity of decision tree induction</i>	167
	<i>From trees to rules</i>	168
	<i>C4.5: Choices and options</i>	169
	<i>Discussion</i>	169
6.2	<b>Classification rules</b>	170
	<i>Criteria for choosing tests</i>	171
	<i>Missing values, numeric attributes</i>	172
	<i>Good rules and bad rules</i>	173
	<i>Generating good rules</i>	174
	<i>Generating good decision lists</i>	175
	<i>Probability measure for rule evaluation</i>	177
	<i>Evaluating rules using a test set</i>	178
	<i>Obtaining rules from partial decision trees</i>	181
	<i>Rules with exceptions</i>	184
	<i>Discussion</i>	187
6.3	<b>Extending linear classification: Support vector machines</b>	<b>188</b>
	<i>The maximum margin hyperplane</i>	189
	<i>Nonlinear class boundaries</i>	191
	<i>Discussion</i>	193
6.4	<b>Instance-based learning</b>	193
	<i>Reducing the number of exemplars</i>	194
	<i>Pruning noisy exemplars</i>	194
	<i>Weighting attributes</i>	195
	<i>Generalizing exemplars</i>	196
	<i>Distance functions for generalized exemplars</i>	197

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

	<i>Stacking</i>	258
	<i>Error-correcting output codes</i>	260
7.5	<b>Further reading</b>	263
<b>8</b>	<b>Nuts and bolts: Machine learning algorithms in Java</b>	<b>265</b>
8.1	<b>Getting started</b>	267
8.2	<b>Javadoc and the class library</b>	271
	<i>Classes, instances, and packages</i>	272
	<i>The weka.core package</i>	272
	<i>The weka.classifiers package</i>	274
	<i>Other packages</i>	276
	<i>Indexes</i>	277
8.3	<b>Processing datasets using the machine learning programs</b>	277
	<i>Using M5'</i>	277
	<i>Generic options</i>	279
	<i>Scheme-specific options</i>	282
	<i>Classifiers</i>	283
	<i>Meta-learning schemes</i>	286
	<i>Filters</i>	289
	<i>Association rules</i>	294
	<i>Clustering</i>	296
8.4	<b>Embedded machine learning</b>	297
	<i>A simple message classifier</i>	299
8.5	<b>Writing new learning schemes</b>	306
	<i>An example classifier</i>	306
	<i>Conventions for implementing classifiers</i>	314
	<i>Writing filters</i>	314
	<i>An example filter</i>	316
	<i>Conventions for writing filters</i>	317
<b>9</b>	<b>Looking forward</b>	<b>321</b>
9.1	<b>Learning from massive datasets</b>	322
9.2	<b>Visualizing machine learning</b>	325
	<i>Visualizing the input</i>	325
	<i>Visualizing the output</i>	327
9.3	<b>Incorporating domain knowledge</b>	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**