

# Contents

LIST OF FIGURES .....	xv
LIST OF TABLES .....	xix
PREFACE .....	xxi
Updated and Revised Content .....	xxv
Second Edition.....	xxv
Third Edition.....	xxvi
ACKNOWLEDGMENTS .....	xxix
ABOUT THE AUTHORS .....	xxxiii

## **PART I INTRODUCTION TO DATA MINING**

---

<b>CHAPTER 1</b>	<b>What's It All About? .....</b>	<b>3</b>
<b>1.1</b>	<b>Data Mining and Machine Learning .....</b>	<b>3</b>
	Describing Structural Patterns .....	5
	Machine Learning .....	7
	Data Mining .....	8
<b>1.2</b>	<b>Simple Examples: The Weather Problem and Others .....</b>	<b>9</b>
	The Weather Problem .....	9
	Contact Lenses: An Idealized Problem .....	12
	Iris: A Classic Numeric Dataset .....	13
	CPU Performance: Introducing Numeric Prediction.....	15
	Labor Negotiations: A More Realistic Example .....	15
	Soybean Classification: A Classic Machine Learning Success....	19
<b>1.3</b>	<b>Fielded Applications .....</b>	<b>21</b>
	Web Mining.....	21
	Decisions Involving Judgment .....	22
	Screening Images .....	23
	Load Forecasting.....	24
	Diagnosis.....	25
	Marketing and Sales .....	26
	Other Applications .....	27
<b>1.4</b>	<b>Machine Learning and Statistics .....</b>	<b>28</b>
<b>1.5</b>	<b>Generalization as Search .....</b>	<b>29</b>
<b>1.6</b>	<b>Data Mining and Ethics .....</b>	<b>33</b>
	Reidentification .....	33
	Using Personal Information.....	34
	Wider Issues .....	35
<b>1.7</b>	<b>Further Reading .....</b>	<b>36</b>

<b>CHAPTER 2</b>	<b>Input: Concepts, Instances, and Attributes</b>	<b>39</b>
2.1	What's a Concept?	40
2.2	What's in an Example?	42
	Relations	43
	Other Example Types	46
2.3	What's in an Attribute?	49
2.4	Preparing the Input	51
	Gathering the Data Together	51
	ARFF Format	52
	Sparse Data	56
	Attribute Types	56
	Missing Values	58
	Inaccurate Values	59
	Getting to Know Your Data	60
2.5	Further Reading	60
<b>CHAPTER 3</b>	<b>Output: Knowledge Representation</b>	<b>61</b>
3.1	Tables	61
3.2	Linear Models	62
3.3	Trees	64
3.4	Rules	67
	Classification Rules	69
	Association Rules	72
	Rules with Exceptions	73
	More Expressive Rules	75
3.5	Instance-Based Representation	78
3.6	Clusters	81
3.7	Further Reading	83
<b>CHAPTER 4</b>	<b>Algorithms: The Basic Methods</b>	<b>85</b>
4.1	Inferring Rudimentary Rules	86
	Missing Values and Numeric Attributes	87
	Discussion	89
4.2	Statistical Modeling	90
	Missing Values and Numeric Attributes	94
	Naïve Bayes for Document Classification	97
	Discussion	99
4.3	Divide-and-Conquer: Constructing Decision Trees	99
	Calculating Information	103
	Highly Branching Attributes	105
	Discussion	107

4.4	Covering Algorithms: Constructing Rules .....	108
	Rules versus Trees .....	109
	A Simple Covering Algorithm.....	110
	Rules versus Decision Lists.....	115
4.5	Mining Association Rules.....	116
	Item Sets.....	116
	Association Rules.....	119
	Generating Rules Efficiently.....	122
	Discussion .....	123
4.6	Linear Models .....	124
	Numeric Prediction: Linear Regression .....	124
	Linear Classification: Logistic Regression.....	125
	Linear Classification Using the Perceptron.....	127
	Linear Classification Using Winnow.....	129
4.7	Instance-Based Learning.....	131
	Distance Function .....	131
	Finding Nearest Neighbors Efficiently.....	132
	Discussion .....	137
4.8	Clustering .....	138
	Iterative Distance-Based Clustering .....	139
	Faster Distance Calculations.....	139
	Discussion .....	141
4.9	Multi-Instance Learning.....	141
	Aggregating the Input .....	142
	Aggregating the Output .....	142
	Discussion .....	142
4.10	Further Reading .....	143
4.11	Weka Implementations.....	145
<b>CHAPTER 5</b>	<b>Credibility: Evaluating What's Been Learned .....</b>	<b>147</b>
5.1	Training and Testing .....	148
5.2	Predicting Performance.....	150
5.3	Cross-Validation.....	152
5.4	Other Estimates.....	154
	Leave-One-Out Cross-Validation.....	154
	The Bootstrap.....	155
5.5	Comparing Data Mining Schemes.....	156
5.6	Predicting Probabilities.....	159
	Quadratic Loss Function.....	160
	Informational Loss Function.....	161
	Discussion .....	162

<b>5.7</b>	Counting the Cost .....	163
	Cost-Sensitive Classification .....	166
	Cost-Sensitive Learning .....	167
	Lift Charts .....	168
	ROC Curves .....	172
	Recall–Precision Curves .....	174
	Discussion .....	175
	Cost Curves .....	177
<b>5.8</b>	Evaluating Numeric Prediction.....	180
<b>5.9</b>	Minimum Description Length Principle.....	183
<b>5.10</b>	Applying the MDL Principle to Clustering.....	186
<b>5.11</b>	Further Reading .....	187

## **PART II ADVANCED DATA MINING**

<b>CHAPTER 6</b>	<b>Implementations: Real Machine Learning Schemes.....</b>	<b>191</b>
<b>6.1</b>	Decision Trees.....	192
	Numeric Attributes.....	193
	Missing Values .....	194
	Pruning .....	195
	Estimating Error Rates.....	197
	Complexity of Decision Tree Induction .....	199
	From Trees to Rules.....	200
	C4.5: Choices and Options .....	201
	Cost-Complexity Pruning .....	202
	Discussion .....	202
<b>6.2</b>	Classification Rules.....	203
	Criteria for Choosing Tests.....	203
	Missing Values, Numeric Attributes.....	204
	Generating Good Rules.....	205
	Using Global Optimization.....	208
	Obtaining Rules from Partial Decision Trees.....	208
	Rules with Exceptions .....	212
	Discussion .....	215
<b>6.3</b>	Association Rules.....	216
	Building a Frequent-Pattern Tree .....	216
	Finding Large Item Sets .....	219
	Discussion .....	222
<b>6.4</b>	Extending Linear Models .....	223
	Maximum-Margin Hyperplane .....	224
	Nonlinear Class Boundaries .....	226

	Support Vector Regression.....	227
	Kernel Ridge Regression.....	229
	Kernel Perceptron.....	231
	Multilayer Perceptrons.....	232
	Radial Basis Function Networks.....	241
	Stochastic Gradient Descent.....	242
	Discussion.....	243
<b>6.5</b>	<b>Instance-Based Learning.....</b>	<b>244</b>
	Reducing the Number of Exemplars.....	245
	Pruning Noisy Exemplars.....	245
	Weighting Attributes.....	246
	Generalizing Exemplars.....	247
	Distance Functions for Generalized Exemplars.....	248
	Generalized Distance Functions.....	249
	Discussion.....	250
<b>6.6</b>	<b>Numeric Prediction with Local Linear Models.....</b>	<b>251</b>
	Model Trees.....	252
	Building the Tree.....	253
	Pruning the Tree.....	253
	Nominal Attributes.....	254
	Missing Values.....	254
	Pseudocode for Model Tree Induction.....	255
	Rules from Model Trees.....	259
	Locally Weighted Linear Regression.....	259
	Discussion.....	261
<b>6.7</b>	<b>Bayesian Networks.....</b>	<b>261</b>
	Making Predictions.....	262
	Learning Bayesian Networks.....	266
	Specific Algorithms.....	268
	Data Structures for Fast Learning.....	270
	Discussion.....	273
<b>6.8</b>	<b>Clustering.....</b>	<b>273</b>
	Choosing the Number of Clusters.....	274
	Hierarchical Clustering.....	274
	Example of Hierarchical Clustering.....	276
	Incremental Clustering.....	279
	Category Utility.....	284
	Probability-Based Clustering.....	285
	The EM Algorithm.....	287
	Extending the Mixture Model.....	289

	Bayesian Clustering .....	290
	Discussion .....	292
<b>6.9</b>	<b>Semisupervised Learning</b> .....	<b>294</b>
	Clustering for Classification .....	294
	Co-training .....	296
	EM and Co-training .....	297
	Discussion .....	297
<b>6.10</b>	<b>Multi-Instance Learning</b> .....	<b>298</b>
	Converting to Single-Instance Learning .....	298
	Upgrading Learning Algorithms .....	300
	Dedicated Multi-Instance Methods .....	301
	Discussion .....	302
<b>6.11</b>	<b>Weka Implementations</b> .....	<b>303</b>
<b>CHAPTER 7</b>	<b>Data Transformations</b> .....	<b>305</b>
<b>7.1</b>	<b>Attribute Selection</b> .....	<b>307</b>
	Scheme-Independent Selection .....	308
	Searching the Attribute Space .....	311
	Scheme-Specific Selection .....	312
<b>7.2</b>	<b>Discretizing Numeric Attributes</b> .....	<b>314</b>
	Unsupervised Discretization .....	316
	Entropy-Based Discretization .....	316
	Other Discretization Methods .....	320
	Entropy-Based versus Error-Based Discretization .....	320
	Converting Discrete Attributes to Numeric Attributes .....	322
<b>7.3</b>	<b>Projections</b> .....	<b>322</b>
	Principal Components Analysis .....	324
	Random Projections .....	326
	Partial Least-Squares Regression .....	326
	Text to Attribute Vectors .....	328
	Time Series .....	330
<b>7.4</b>	<b>Sampling</b> .....	<b>330</b>
	Reservoir Sampling .....	330
<b>7.5</b>	<b>Cleansing</b> .....	<b>331</b>
	Improving Decision Trees .....	332
	Robust Regression .....	333
	Detecting Anomalies .....	334
	One-Class Learning .....	335
<b>7.6</b>	<b>Transforming Multiple Classes to Binary Ones</b> .....	<b>338</b>
	Simple Methods .....	338
	Error-Correcting Output Codes .....	339
	Ensembles of Nested Dichotomies .....	341

7.7	Calibrating Class Probabilities .....	343
7.8	Further Reading .....	346
7.9	Weka Implementations.....	348
<b>CHAPTER 8</b>	<b>Ensemble Learning .....</b>	<b>351</b>
8.1	Combining Multiple Models.....	351
8.2	Bagging .....	352
	Bias–Variance Decomposition .....	353
	Bagging with Costs.....	355
8.3	Randomization .....	356
	Randomization versus Bagging .....	357
	Rotation Forests .....	357
8.4	Boosting .....	358
	AdaBoost.....	358
	The Power of Boosting .....	361
8.5	Additive Regression.....	362
	Numeric Prediction .....	362
	Additive Logistic Regression .....	364
8.6	Interpretable Ensembles.....	365
	Option Trees.....	365
	Logistic Model Trees .....	368
8.7	Stacking.....	369
8.8	Further Reading .....	371
8.9	Weka Implementations.....	372
<b>Chapter 9</b>	<b>Moving on: Applications and Beyond.....</b>	<b>375</b>
9.1	Applying Data Mining .....	375
9.2	Learning from Massive Datasets .....	378
9.3	Data Stream Learning .....	380
9.4	Incorporating Domain Knowledge .....	384
9.5	Text Mining.....	386
9.6	Web Mining.....	389
9.7	Adversarial Situations.....	393
9.8	Ubiquitous Data Mining .....	395
9.9	Further Reading .....	397
 <b>PART III THE WEKA DATA MINING WORKBENCH</b>		
<b>CHAPTER 10</b>	<b>Introduction to Weka .....</b>	<b>403</b>
10.1	What’s in Weka? .....	403
10.2	How Do You Use It? .....	404
10.3	What Else Can You Do?.....	405
10.4	How Do You Get It?.....	406

<b>CHAPTER 11 The Explorer</b> .....	<b>407</b>
<b>11.1 Getting Started</b> .....	407
Preparing the Data .....	407
Loading the Data into the Explorer .....	408
Building a Decision Tree .....	410
Examining the Output.....	411
Doing It Again .....	413
Working with Models .....	414
When Things Go Wrong.....	415
<b>11.2 Exploring the Explorer</b> .....	416
Loading and Filtering Files .....	416
Training and Testing Learning Schemes .....	422
Do It Yourself: The User Classifier .....	424
Using a Metalearner.....	427
Clustering and Association Rules.....	429
Attribute Selection .....	430
Visualization.....	430
<b>11.3 Filtering Algorithms</b> .....	432
Unsupervised Attribute Filters.....	432
Unsupervised Instance Filters.....	441
Supervised Filters.....	443
<b>11.4 Learning Algorithms</b> .....	445
Bayesian Classifiers .....	451
Trees .....	454
Rules.....	457
Functions .....	459
Neural Networks .....	469
Lazy Classifiers.....	472
Multi-Instance Classifiers .....	472
Miscellaneous Classifiers.....	474
<b>11.5 Metalearning Algorithms</b> .....	474
Bagging and Randomization.....	474
Boosting .....	476
Combining Classifiers .....	477
Cost-Sensitive Learning.....	477
Optimizing Performance.....	478
Retargeting Classifiers for Different Tasks .....	479
<b>11.6 Clustering Algorithms</b> .....	480
<b>11.7 Association-Rule Learners</b> .....	485
<b>11.8 Attribute Selection</b> .....	487
Attribute Subset Evaluators .....	488



Single-Attribute Evaluators .....	490
Search Methods.....	492
<b>CHAPTER 12 The Knowledge Flow Interface .....</b>	<b>495</b>
12.1 Getting Started .....	495
12.2 Components.....	498
12.3 Configuring and Connecting the Components .....	500
12.4 Incremental Learning.....	502
<b>CHAPTER 13 The Experimenter .....</b>	<b>505</b>
13.1 Getting Started .....	505
Running an Experiment.....	506
Analyzing the Results.....	509
13.2 Simple Setup.....	510
13.3 Advanced Setup.....	511
13.4 The Analyze Panel.....	512
13.5 Distributing Processing over Several Machines.....	515
<b>CHAPTER 14 The Command-Line Interface.....</b>	<b>519</b>
14.1 Getting Started .....	519
14.2 The Structure of Weka .....	519
Classes, Instances, and Packages.....	520
The <i>weka.core</i> Package.....	520
The <i>weka.classifiers</i> Package.....	523
Other Packages.....	525
Javadoc Indexes .....	525
14.3 Command-Line Options.....	526
Generic Options .....	526
Scheme-Specific Options .....	529
<b>CHAPTER 15 Embedded Machine Learning .....</b>	<b>531</b>
15.1 A Simple Data Mining Application.....	531
<i>MessageClassifier()</i> .....	536
<i>updateData()</i> .....	536
<i>classifyMessage()</i> .....	537
<b>CHAPTER 16 Writing New Learning Schemes .....</b>	<b>539</b>
16.1 An Example Classifier .....	539
<i>buildClassifier()</i> .....	540
<i>makeTree()</i> .....	540
<i>computeInfoGain()</i> .....	549
<i>classifyInstance()</i> .....	549

	<i>toSource()</i> .....	550
	<i>main()</i> .....	553
<b>16.2</b>	<b>Conventions for Implementing Classifiers</b> .....	<b>555</b>
	Capabilities.....	555
<b>CHAPTER 17</b>	<b>Tutorial Exercises for the Weka Explorer</b> .....	<b>559</b>
<b>17.1</b>	<b>Introduction to the Explorer Interface</b> .....	<b>559</b>
	Loading a Dataset .....	559
	The Dataset Editor .....	560
	Applying a Filter.....	561
	The Visualize Panel .....	562
	The Classify Panel .....	562
<b>17.2</b>	<b>Nearest-Neighbor Learning and Decision Trees</b> .....	<b>566</b>
	The Glass Dataset .....	566
	Attribute Selection .....	567
	Class Noise and Nearest-Neighbor Learning .....	568
	Varying the Amount of Training Data.....	569
	Interactive Decision Tree Construction .....	569
<b>17.3</b>	<b>Classification Boundaries</b> .....	<b>571</b>
	Visualizing 1R.....	571
	Visualizing Nearest-Neighbor Learning .....	572
	Visualizing Naïve Bayes.....	573
	Visualizing Decision Trees and Rule Sets.....	573
	Messing with the Data.....	574
<b>17.4</b>	<b>Preprocessing and Parameter Tuning</b> .....	<b>574</b>
	Discretization .....	574
	More on Discretization .....	575
	Automatic Attribute Selection .....	575
	More on Automatic Attribute Selection .....	576
	Automatic Parameter Tuning.....	577
<b>17.5</b>	<b>Document Classification</b> .....	<b>578</b>
	Data with String Attributes .....	579
	Classifying Actual Documents .....	580
	Exploring the <i>StringToWordVector</i> Filter .....	581
<b>17.6</b>	<b>Mining Association Rules</b> .....	<b>582</b>
	Association-Rule Mining.....	582
	Mining a Real-World Dataset.....	584
	Market Basket Analysis .....	584
	<b>REFERENCES</b> .....	<b>587</b>
	<b>INDEX</b> .....	<b>607</b>

# List of Figures

Figure 1.1 Rules for the contact lens data.	12
Figure 1.2 Decision tree for the contact lens data.	13
Figure 1.3 Decision trees for the labor negotiations data.	18
Figure 2.1 A family tree and two ways of expressing the sister-of relation.	43
Figure 2.2 ARFF file for the weather data.	53
Figure 2.3 Multi-instance ARFF file for the weather data.	55
Figure 3.1 A linear regression function for the CPU performance data.	62
Figure 3.2 A linear decision boundary separating <i>Iris setosas</i> from <i>Iris versicolors</i> .	63
Figure 3.3 Constructing a decision tree interactively.	66
Figure 3.4 Models for the CPU performance data.	68
Figure 3.5 Decision tree for a simple disjunction.	69
Figure 3.6 The exclusive-or problem.	70
Figure 3.7 Decision tree with a replicated subtree.	71
Figure 3.8 Rules for the iris data.	74
Figure 3.9 The shapes problem.	76
Figure 3.10 Different ways of partitioning the instance space.	80
Figure 3.11 Different ways of representing clusters.	82
Figure 4.1 Pseudocode for 1R.	86
Figure 4.2 Tree stumps for the weather data.	100
Figure 4.3 Expanded tree stumps for the weather data.	102
Figure 4.4 Decision tree for the weather data.	103
Figure 4.5 Tree stump for the <i>ID code</i> attribute.	105
Figure 4.6 Covering algorithm.	109
Figure 4.7 The instance space during operation of a covering algorithm.	110
Figure 4.8 Pseudocode for a basic rule learner.	114
Figure 4.9 Logistic regression.	127
Figure 4.10 The perceptron.	129
Figure 4.11 The Winnow algorithm.	130
Figure 4.12 A <i>kD</i> -tree for four training instances.	133
Figure 4.13 Using a <i>kD</i> -tree to find the nearest neighbor of the star.	134
Figure 4.14 Ball tree for 16 training instances.	136
Figure 4.15 Ruling out an entire ball (gray) based on a target point (star) and its current nearest neighbor.	137
Figure 4.16 A ball tree.	141
Figure 5.1 A hypothetical lift chart.	170
Figure 5.2 Analyzing the expected benefit of a mailing campaign.	171
Figure 5.3 A sample ROC curve.	173
Figure 5.4 ROC curves for two learning schemes.	174
Figure 5.5 Effect of varying the probability threshold.	178
Figure 6.1 Example of subtree raising.	196

Figure 6.2 Pruning the labor negotiations decision tree.	200
Figure 6.3 Algorithm for forming rules by incremental reduced-error pruning.	207
Figure 6.4 RIPPER.	209
Figure 6.5 Algorithm for expanding examples into a partial tree.	210
Figure 6.6 Example of building a partial tree.	211
Figure 6.7 Rules with exceptions for the iris data.	213
Figure 6.8 Extended prefix trees for the weather data.	220
Figure 6.9 A maximum-margin hyperplane.	225
Figure 6.10 Support vector regression.	228
Figure 6.11 Example datasets and corresponding perceptrons.	233
Figure 6.12 Step versus sigmoid.	240
Figure 6.13 Gradient descent using the error function $w^2 + 1$ .	240
Figure 6.14 Multilayer perceptron with a hidden layer.	241
Figure 6.15 Hinge, squared, and 0 – 1 loss functions.	242
Figure 6.16 A boundary between two rectangular classes.	248
Figure 6.17 Pseudocode for model tree induction.	255
Figure 6.18 Model tree for a dataset with nominal attributes.	256
Figure 6.19 A simple Bayesian network for the weather data.	262
Figure 6.20 Another Bayesian network for the weather data.	264
Figure 6.21 The weather data.	270
Figure 6.22 Hierarchical clustering displays.	276
Figure 6.23 Clustering the weather data.	279
Figure 6.24 Hierarchical clusterings of the iris data.	281
Figure 6.25 A two-class mixture model.	285
Figure 6.26 <i>DensiTree</i> showing possible hierarchical clusterings of a given dataset.	291
Figure 7.1 Attribute space for the weather dataset.	311
Figure 7.2 Discretizing the <i>temperature</i> attribute using the entropy method.	318
Figure 7.3 The result of discretizing the <i>temperature</i> attribute.	318
Figure 7.4 Class distribution for a two-class, two-attribute problem.	321
Figure 7.5 Principal components transform of a dataset.	325
Figure 7.6 Number of international phone calls from Belgium, 1950–1973.	333
Figure 7.7 Overoptimistic probability estimation for a two-class problem.	344
Figure 8.1 Algorithm for bagging.	355
Figure 8.2 Algorithm for boosting.	359
Figure 8.3 Algorithm for additive logistic regression.	365
Figure 8.4 Simple option tree for the weather data.	366
Figure 8.5 Alternating decision tree for the weather data.	367
Figure 9.1 A tangled “web.”	391
Figure 11.1 The Explorer interface.	408
Figure 11.2 Weather data.	409
Figure 11.3 The Weka Explorer.	410

Figure 11.4 Using <i>J4.8</i> .	411
Figure 11.5 Output from the <i>J4.8</i> decision tree learner.	412
Figure 11.6 Visualizing the result of <i>J4.8</i> on the iris dataset.	415
Figure 11.7 Generic Object Editor.	417
Figure 11.8 The SQLViewer tool.	418
Figure 11.9 Choosing a filter.	420
Figure 11.10 The weather data with two attributes removed.	422
Figure 11.11 Processing the CPU performance data with <i>M5'</i> .	423
Figure 11.12 Output from the <i>M5'</i> program for numeric prediction.	425
Figure 11.13 Visualizing the errors.	426
Figure 11.14 Working on the segment-challenge data with the User Classifier.	428
Figure 11.15 Configuring a metalearner for boosting decision stumps.	429
Figure 11.16 Output from the <i>Apriori</i> program for association rules.	430
Figure 11.17 Visualizing the iris dataset.	431
Figure 11.18 Using Weka's metalearner for discretization.	443
Figure 11.19 Output of <i>NaiveBayes</i> on the weather data.	452
Figure 11.20 Visualizing a Bayesian network for the weather data (nominal version).	454
Figure 11.21 Changing the parameters for <i>J4.8</i> .	455
Figure 11.22 Output of <i>OneR</i> on the labor negotiations data.	458
Figure 11.23 Output of <i>PART</i> for the labor negotiations data.	460
Figure 11.24 Output of <i>SimpleLinearRegression</i> for the CPU performance data.	461
Figure 11.25 Output of <i>SMO</i> on the iris data.	463
Figure 11.26 Output of <i>SMO</i> with a nonlinear kernel on the iris data.	465
Figure 11.27 Output of <i>Logistic</i> on the iris data.	468
Figure 11.28 Using Weka's neural-network graphical user interface.	470
Figure 11.29 Output of <i>SimpleKMeans</i> on the weather data.	481
Figure 11.30 Output of <i>EM</i> on the weather data.	482
Figure 11.31 Clusters formed by <i>DBScan</i> on the iris data.	484
Figure 11.32 <i>OPTICS</i> visualization for the iris data.	485
Figure 11.33 Attribute selection: specifying an evaluator and a search method.	488
Figure 12.1 The Knowledge Flow interface.	496
Figure 12.2 Configuring a data source.	497
Figure 12.3 Status area after executing the configuration shown in Figure 12.1.	497
Figure 12.4 Operations on the Knowledge Flow components.	500
Figure 12.5 A Knowledge Flow that operates incrementally.	503
Figure 13.1 An experiment.	506
Figure 13.2 Statistical test results for the experiment in Figure 13.1.	509
Figure 13.3 Setting up an experiment in advanced mode.	511
Figure 13.4 An experiment in clustering.	513

Figure 13.5	Rows and columns of Figure 13.2.	514
Figure 14.1	Using Javadoc.	521
Figure 14.2	<i>DecisionStump</i> , a class of the <i>weka.classifiers.trees</i> package.	524
Figure 15.1	Source code for the message classifier.	532
Figure 16.1	Source code for the ID3 decision tree learner.	541
Figure 16.2	Source code produced by <i>weka.classifiers.trees.Id3</i> for the weather data.	551
Figure 16.3	Javadoc for the <i>Capability</i> enumeration.	556
Figure 17.1	The data viewer.	560
Figure 17.2	Output after building and testing the classifier.	564
Figure 17.3	The decision tree that has been built.	565

# List of Tables

Table 1.1 Contact Lens Data	6
Table 1.2 Weather Data	10
Table 1.3 Weather Data with Some Numeric Attributes	11
Table 1.4 Iris Data	14
Table 1.5 CPU Performance Data	16
Table 1.6 Labor Negotiations Data	17
Table 1.7 Soybean Data	20
Table 2.1 Iris Data as a Clustering Problem	41
Table 2.2 Weather Data with a Numeric Class	42
Table 2.3 Family Tree	44
Table 2.4 Sister-of Relation	45
Table 2.5 Another Relation	47
Table 3.1 New Iris Flower	73
Table 3.2 Training Data for the Shapes Problem	76
Table 4.1 Evaluating Attributes in the Weather Data	87
Table 4.2 Weather Data with Counts and Probabilities	91
Table 4.3 A New Day	92
Table 4.4 Numeric Weather Data with Summary Statistics	95
Table 4.5 Another New Day	96
Table 4.6 Weather Data with Identification Codes	106
Table 4.7 Gain Ratio Calculations for Figure 4.2 Tree Stumps	107
Table 4.8 Part of Contact Lens Data for which <i>astigmatism = yes</i>	112
Table 4.9 Part of Contact Lens Data for which <i>astigmatism = yes</i> and <i>tear production rate = normal</i>	113
Table 4.10 Item Sets for Weather Data with Coverage 2 or Greater	117
Table 4.11 Association Rules for Weather Data	120
Table 5.1 Confidence Limits for Normal Distribution	152
Table 5.2 Confidence Limits for Student's Distribution with 9 Degrees of Freedom	159
Table 5.3 Different Outcomes of a Two-Class Prediction	164
Table 5.4 Different Outcomes of a Three-Class Prediction	165
Table 5.5 Default Cost Matrixes	166
Table 5.6 Data for a Lift Chart	169
Table 5.7 Different Measures Used to Evaluate the False Positive versus False Negative Trade-Off	176
Table 5.8 Performance Measures for Numeric Prediction	180
Table 5.9 Performance Measures for Four Numeric Prediction Models	182
Table 6.1 Preparing Weather Data for Insertion into an <i>FP</i> -Tree	217
Table 6.2 Linear Models in the Model Tree	257
Table 7.1 First Five Instances from CPU Performance Data	327
Table 7.2 Transforming a Multiclass Problem into a Two-Class One	340

Table 7.3 Nested Dichotomy in the Form of a Code Matrix	342
Table 9.1 Top 10 Algorithms in Data Mining	376
Table 11.1 Unsupervised Attribute Filters	433
Table 11.2 Unsupervised Instance Filters	441
Table 11.3 Supervised Attribute Filters	444
Table 11.4 Supervised Instance Filters	444
Table 11.5 Classifier Algorithms in Weka	446
Table 11.6 Metalearning Algorithms in Weka	475
Table 11.7 Clustering Algorithms	480
Table 11.8 Association-Rule Learners	486
Table 11.9 Attribute Evaluation Methods for Attribute Selection	489
Table 11.10 Search Methods for Attribute Selection	490
Table 12.1 Visualization and Evaluation Components	499
Table 14.1 Generic Options for Learning Schemes	527
Table 14.2 Scheme-Specific Options for the J4.8 Decision Tree Learner	528
Table 16.1 Simple Learning Schemes in Weka	540
Table 17.1 Accuracy Obtained Using <i>IBk</i> , for Different Attribute Subsets	568
Table 17.2 Effect of Class Noise on <i>IBk</i> , for Different Neighborhood Sizes	569
Table 17.3 Effect of Training Set Size on <i>IBk</i> and <i>J48</i>	570
Table 17.4 Training Documents	580
Table 17.5 Test Documents	580
Table 17.6 Number of Rules for Different Values of Minimum Confidence and Support	584