

# Data Mining

## Practical Machine Learning Tools and Techniques

Third Edition

**Ian H. Witten**

**Eibe Frank**

**Mark A. Hall**



**ELSEVIER**

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

Morgan Kaufmann Publishers is an imprint of Elsevier



# 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 What's It All About? .....</b>	<b>3</b>
<b>1.1 Data Mining and Machine Learning .....</b>	<b>3</b>
Describing Structural Patterns .....	5
Machine Learning .....	7
Data Mining .....	8
<b>1.2 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 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 Machine Learning and Statistics .....</b>	<b>28</b>
<b>1.5 Generalization as Search .....</b>	<b>29</b>
<b>1.6 Data Mining and Ethics .....</b>	<b>33</b>
Reidentification .....	33
Using Personal Information .....	34
Wider Issues .....	35
<b>1.7 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>	Instance-Based Learning.....	244
	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>	Numeric Prediction with Local Linear Models.....	251
	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>	Bayesian Networks.....	261
	Making Predictions.....	262
	Learning Bayesian Networks.....	266
	Specific Algorithms.....	268
	Data Structures for Fast Learning.....	270
	Discussion.....	273
<b>6.8</b>	Clustering.....	273
	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>	Semisupervised Learning.....	294
	Clustering for Classification .....	294
	Co-training .....	296
	EM and Co-training .....	297
	Discussion .....	297
<b>6.10</b>	Multi-Instance Learning.....	298
	Converting to Single-Instance Learning.....	298
	Upgrading Learning Algorithms.....	300
	Dedicated Multi-Instance Methods.....	301
	Discussion .....	302
<b>6.11</b>	Weka Implementations.....	303
<b>CHAPTER 7</b>	<b>Data Transformations .....</b>	<b>305</b>
<b>7.1</b>	Attribute Selection .....	307
	Scheme-Independent Selection.....	308
	Searching the Attribute Space .....	311
	Scheme-Specific Selection.....	312
<b>7.2</b>	Discretizing Numeric Attributes .....	314
	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>	Projections.....	322
	Principal Components Analysis.....	324
	Random Projections.....	326
	Partial Least-Squares Regression .....	326
	Text to Attribute Vectors.....	328
	Time Series .....	330
<b>7.4</b>	Sampling .....	330
	Reservoir Sampling.....	330
<b>7.5</b>	Cleansing.....	331
	Improving Decision Trees.....	332
	Robust Regression .....	333
	Detecting Anomalies.....	334
	One-Class Learning .....	335
<b>7.6</b>	Transforming Multiple Classes to Binary Ones.....	338
	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</b>	<b>The Explorer</b>	<b>407</b>
11.1	Getting Started	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
11.2	Exploring the Explorer	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
11.3	Filtering Algorithms	432
	Unsupervised Attribute Filters	432
	Unsupervised Instance Filters	441
	Supervised Filters	443
11.4	Learning Algorithms	445
	Bayesian Classifiers	451
	Trees	454
	Rules	457
	Functions	459
	Neural Networks	469
	Lazy Classifiers	472
	Multi-Instance Classifiers	472
	Miscellaneous Classifiers	474
11.5	Metalearning Algorithms	474
	Bagging and Randomization	474
	Boosting	476
	Combining Classifiers	477
	Cost-Sensitive Learning	477
	Optimizing Performance	478
	Retargeting Classifiers for Different Tasks	479
11.6	Clustering Algorithms	480
11.7	Association-Rule Learners	485
11.8	Attribute Selection	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>	Conventions for Implementing Classifiers .....	555
	Capabilities .....	555
<b>CHAPTER 17</b>	<b>Tutorial Exercises for the Weka Explorer .....</b>	<b>559</b>
<b>17.1</b>	Introduction to the Explorer Interface .....	559
	Loading a Dataset .....	559
	The Dataset Editor .....	560
	Applying a Filter .....	561
	The Visualize Panel .....	562
	The Classify Panel .....	562
<b>17.2</b>	Nearest-Neighbor Learning and Decision Trees .....	566
	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>	Classification Boundaries .....	571
	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>	Preprocessing and Parameter Tuning .....	574
	Discretization .....	574
	More on Discretization .....	575
	Automatic Attribute Selection .....	575
	More on Automatic Attribute Selection .....	576
	Automatic Parameter Tuning .....	577
<b>17.5</b>	Document Classification .....	578
	Data with String Attributes .....	579
	Classifying Actual Documents .....	580
	Exploring the <i>StringToWordVector</i> Filter .....	581
<b>17.6</b>	Mining Association Rules .....	582
	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>