

Data Mining

Practical Machine Learning Tools and Techniques

Fourth Edition

Ian H. Witten

University of Waikato, Hamilton, New Zealand

Eibe Frank

University of Waikato, Hamilton, New Zealand

Mark A. Hall

University of Waikato, Hamilton, New Zealand

Christopher J. Pal

Polytechnique Montréal, Montreal, Quebec, Canada



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

Morgan Kaufmann is an imprint of Elsevier



Contents

List of Figures.....	xv
List of Tables.....	xxi
Preface	xxiii

PART I INTRODUCTION TO DATA MINING

CHAPTER 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.....	12
Iris: A Classic Numeric Dataset	14
CPU Performance: Introducing Numeric Prediction	16
Labor Negotiations: A More Realistic Example.....	16
Soybean Classification: A Classic Machine Learning Success	19
1.3 Fielded Applications	21
Web Mining.....	21
Decisions Involving Judgment	22
Screening Images.....	23
Load Forecasting	24
Diagnosis.....	25
Marketing and Sales	26
Other Applications.....	27
1.4 The Data Mining Process.....	28
1.5 Machine Learning and Statistics.....	30
1.6 Generalization as Search.....	31
Enumerating the Concept Space	32
Bias	33
1.7 Data Mining and Ethics	35
Reidentification.....	36
Using Personal Information.....	37
Wider Issues.....	38
1.8 Further Reading and Bibliographic Notes	38

CHAPTER 2	Input: concepts, instances, attributes	43
2.1	What's a Concept?	44
2.2	What's in an Example?	46
	Relations	47
	Other Example Types	51
2.3	What's in an Attribute?	53
2.4	Preparing the Input	56
	Gathering the Data Together	56
	ARFF Format	57
	Sparse Data	60
	Attribute Types	61
	Missing Values	62
	Inaccurate Values	63
	Unbalanced Data	64
	Getting to Know Your Data	65
2.5	Further Reading and Bibliographic Notes	65
CHAPTER 3	Output: knowledge representation	67
3.1	Tables	68
3.2	Linear Models	68
3.3	Trees	70
3.4	Rules	75
	Classification Rules	75
	Association Rules	79
	Rules With Exceptions	80
	More Expressive Rules	82
3.5	Instance-Based Representation	84
3.6	Clusters	87
3.7	Further Reading and Bibliographic Notes	88
CHAPTER 4	Algorithms: the basic methods	91
4.1	Inferring Rudimentary Rules	93
	Missing Values and Numeric Attributes	94
4.2	Simple Probabilistic Modeling	96
	Missing Values and Numeric Attributes	100
	Naïve Bayes for Document Classification	103
	Remarks	105
4.3	Divide-and-Conquer: Constructing Decision Trees	105
	Calculating Information	108
	Highly Branching Attributes	110

4.4	Covering Algorithms: Constructing Rules	113
	Rules Versus Trees	114
	A Simple Covering Algorithm	115
	Rules Versus Decision Lists	119
4.5	Mining Association Rules	120
	Item Sets	120
	Association Rules	122
	Generating Rules Efficiently	124
4.6	Linear Models	128
	Numeric Prediction: Linear Regression	128
	Linear Classification: Logistic Regression	129
	Linear Classification Using the Perceptron	131
	Linear Classification Using Winnow	133
4.7	Instance-Based Learning	135
	The Distance Function.....	135
	Finding Nearest Neighbors Efficiently	136
	Remarks	141
4.8	Clustering	141
	Iterative Distance-Based Clustering	142
	Faster Distance Calculations	144
	Choosing the Number of Clusters	146
	Hierarchical Clustering	147
	Example of Hierarchical Clustering.....	148
	Incremental Clustering.....	150
	Category Utility	154
	Remarks	156
4.9	Multi-instance Learning	156
	Aggregating the Input.....	157
	Aggregating the Output	157
4.10	Further Reading and Bibliographic Notes	158
4.11	WEKA Implementations	160
CHAPTER 5	Credibility: evaluating what's been learned	161
5.1	Training and Testing	163
5.2	Predicting Performance	165
5.3	Cross-Validation	167
5.4	Other Estimates	169
	Leave-One-Out	169
	The Bootstrap.....	169
5.5	Hyperparameter Selection	171

5.6	Comparing Data Mining Schemes.....	172
5.7	Predicting Probabilities.....	176
	Quadratic Loss Function.....	177
	Informational Loss Function	178
	Remarks	179
5.8	Counting the Cost	179
	Cost-Sensitive Classification	182
	Cost-Sensitive Learning.....	183
	Lift Charts	183
	ROC Curves.....	186
	Recall-Precision Curves.....	190
	Remarks	190
	Cost Curves.....	192
5.9	Evaluating Numeric Prediction	194
5.10	The MDL Principle.....	197
5.11	Applying the MDL Principle to Clustering.....	200
5.12	Using a Validation Set for Model Selection	201
5.13	Further Reading and Bibliographic Notes.....	202

PART II MORE ADVANCED MACHINE LEARNING SCHEMES

CHAPTER 6	Trees and rules	209
6.1	Decision Trees.....	210
	Numeric Attributes	210
	Missing Values	212
	Pruning.....	213
	Estimating Error Rates	215
	Complexity of Decision Tree Induction.....	217
	From Trees to Rules	219
	C4.5: Choices and Options.....	219
	Cost-Complexity Pruning	220
	Discussion	221
6.2	Classification Rules.....	221
	Criteria for Choosing Tests	222
	Missing Values, Numeric Attributes	223
	Generating Good Rules	224
	Using Global Optimization.....	226
	Obtaining Rules From Partial Decision Trees	227
	Rules With Exceptions	231
	Discussion	233

6.3	Association Rules.....	234
	Building a Frequent Pattern Tree.....	235
	Finding Large Item Sets.....	240
	Discussion.....	241
6.4	WEKA Implementations.....	242
CHAPTER 7	Extending instance-based and linear models	243
7.1	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.....	250
	Discussion.....	250
7.2	Extending Linear Models.....	252
	The Maximum Margin Hyperplane.....	253
	Nonlinear Class Boundaries.....	254
	Support Vector Regression.....	256
	Kernel Ridge Regression.....	258
	The Kernel Perceptron.....	260
	Multilayer Perceptrons.....	261
	Radial Basis Function Networks.....	270
	Stochastic Gradient Descent.....	270
	Discussion.....	272
7.3	Numeric Prediction With Local Linear Models.....	273
	Model Trees.....	274
	Building the Tree.....	275
	Pruning the Tree.....	275
	Nominal Attributes.....	276
	Missing Values.....	276
	Pseudocode for Model Tree Induction.....	277
	Rules From Model Trees.....	281
	Locally Weighted Linear Regression.....	281
	Discussion.....	283
7.4	WEKA Implementations.....	284
CHAPTER 8	Data transformations	285
8.1	Attribute Selection.....	288
	Scheme-Independent Selection.....	289
	Searching the Attribute Space.....	292
	Scheme-Specific Selection.....	293

8.2	Discretizing Numeric Attributes	296
	Unsupervised Discretization	297
	Entropy-Based Discretization	298
	Other Discretization Methods	301
	Entropy-Based Versus Error-Based Discretization	302
	Converting Discrete to Numeric Attributes	303
8.3	Projections	304
	Principal Component Analysis	305
	Random Projections	307
	Partial Least Squares Regression	307
	Independent Component Analysis	309
	Linear Discriminant Analysis	310
	Quadratic Discriminant Analysis	310
	Fisher's Linear Discriminant Analysis	311
	Text to Attribute Vectors	313
	Time Series	314
8.4	Sampling	315
	Reservoir Sampling	315
8.5	Cleansing	316
	Improving Decision Trees	316
	Robust Regression	317
	Detecting Anomalies	318
	One-Class Learning	319
	Outlier Detection	320
	Generating Artificial Data	321
8.6	Transforming Multiple Classes to Binary Ones	322
	Simple Methods	323
	Error-Correcting Output Codes	324
	Ensembles of Nested Dichotomies	326
8.7	Calibrating Class Probabilities	328
8.8	Further Reading and Bibliographic Notes	331
8.9	WEKA Implementations	334
CHAPTER 9	Probabilistic methods	335
9.1	Foundations	336
	Maximum Likelihood Estimation	338
	Maximum a Posteriori Parameter Estimation	339
9.2	Bayesian Networks	339
	Making Predictions	340

Learning Bayesian Networks	344
Specific Algorithms	347
Data Structures for Fast Learning	349
9.3 Clustering and Probability Density Estimation	352
The Expectation Maximization Algorithm for a Mixture of Gaussians	353
Extending the Mixture Model	356
Clustering Using Prior Distributions	358
Clustering With Correlated Attributes	359
Kernel Density Estimation	361
Comparing Parametric, Semiparametric and Nonparametric Density Models for Classification	362
9.4 Hidden Variable Models	363
Expected Log-Likelihoods and Expected Gradients.....	364
The Expectation Maximization Algorithm	365
Applying the Expectation Maximization Algorithm to Bayesian Networks	366
9.5 Bayesian Estimation and Prediction	367
Probabilistic Inference Methods.....	368
9.6 Graphical Models and Factor Graphs	370
Graphical Models and Plate Notation	371
Probabilistic Principal Component Analysis.....	372
Latent Semantic Analysis	376
Using Principal Component Analysis for Dimensionality Reduction	377
Probabilistic LSA.....	378
Latent Dirichlet Allocation.....	379
Factor Graphs.....	382
Markov Random Fields	385
Computing Using the Sum-Product and Max-Product Algorithms	386
9.7 Conditional Probability Models	392
Linear and Polynomial Regression as Probability Models.....	392
Using Priors on Parameters	393
Multiclass Logistic Regression.....	396
Gradient Descent and Second-Order Methods.....	400
Generalized Linear Models	400
Making Predictions for Ordered Classes.....	402
Conditional Probabilistic Models Using Kernels.....	402

9.8	Sequential and Temporal Models	403
	Markov Models and <i>N</i> -gram Methods	403
	Hidden Markov Models.....	404
	Conditional Random Fields	406
9.9	Further Reading and Bibliographic Notes.....	410
	Software Packages and Implementations.....	414
9.10	WEKA Implementations.....	416
CHAPTER 10	Deep learning	417
10.1	Deep Feedforward Networks	420
	The MNIST Evaluation	421
	Losses and Regularization	422
	Deep Layered Network Architecture	423
	Activation Functions.....	424
	Backpropagation Revisited.....	426
	Computation Graphs and Complex Network Structures	429
	Checking Backpropagation Implementations	430
10.2	Training and Evaluating Deep Networks	431
	Early Stopping	431
	Validation, Cross-Validation, and Hyperparameter Tuning ...	432
	Mini-Batch-Based Stochastic Gradient Descent.....	433
	Pseudocode for Mini-Batch Based Stochastic Gradient Descent.....	434
	Learning Rates and Schedules.....	434
	Regularization With Priors on Parameters.....	435
	Dropout	436
	Batch Normalization.....	436
	Parameter Initialization.....	436
	Unsupervised Pretraining.....	437
	Data Augmentation and Synthetic Transformations	437
10.3	Convolutional Neural Networks	437
	The ImageNet Evaluation and Very Deep Convolutional Networks	438
	From Image Filtering to Learnable Convolutional Layers.....	439
	Convolutional Layers and Gradients.....	443
	Pooling and Subsampling Layers and Gradients	444
	Implementation	445
10.4	Autoencoders.....	445
	Pretraining Deep Autoencoders With RBMs.....	448
	Denoising Autoencoders and Layerwise Training.....	448
	Combining Reconstructive and Discriminative Learning.....	449

10.5	Stochastic Deep Networks	449
	Boltzmann Machines	449
	Restricted Boltzmann Machines.....	451
	Contrastive Divergence	452
	Categorical and Continuous Variables.....	452
	Deep Boltzmann Machines.....	453
	Deep Belief Networks	455
10.6	Recurrent Neural Networks	456
	Exploding and Vanishing Gradients	457
	Other Recurrent Network Architectures	459
10.7	Further Reading and Bibliographic Notes.....	461
10.8	Deep Learning Software and Network Implementations.....	464
	Theano.....	464
	Tensor Flow	464
	Torch.....	465
	Computational Network Toolkit.....	465
	Caffe.....	465
	Deeplearning4j.....	465
	Other Packages: Lasagne, Keras, and cuDNN.....	465
10.9	WEKA Implementations.....	466
CHAPTER 11	Beyond supervised and unsupervised learning	467
11.1	Semisupervised Learning.....	468
	Clustering for Classification.....	468
	Cotraining	470
	EM and Cotraining	471
	Neural Network Approaches	471
11.2	Multi-instance Learning.....	472
	Converting to Single-Instance Learning	472
	Upgrading Learning Algorithms	475
	Dedicated Multi-instance Methods.....	475
11.3	Further Reading and Bibliographic Notes.....	477
11.4	WEKA Implementations.....	478
CHAPTER 12	Ensemble learning.....	479
12.1	Combining Multiple Models.....	480
12.2	Bagging	481
	Bias–Variance Decomposition	482
	Bagging With Costs.....	483
12.3	Randomization	484
	Randomization Versus Bagging.....	485
	Rotation Forests	486

12.4	Boosting	486
	AdaBoost.....	487
	The Power of Boosting.....	489
12.5	Additive Regression.....	490
	Numeric Prediction.....	491
	Additive Logistic Regression	492
12.6	Interpretable Ensembles.....	493
	Option Trees	494
	Logistic Model Trees.....	496
12.7	Stacking.....	497
12.8	Further Reading and Bibliographic Notes.....	499
12.9	WEKA Implementations.....	501
CHAPTER 13	Moving on: applications and beyond.....	503
13.1	Applying Machine Learning.....	504
13.2	Learning From Massive Datasets.....	506
13.3	Data Stream Learning.....	509
13.4	Incorporating Domain Knowledge.....	512
13.5	Text Mining	515
	Document Classification and Clustering.....	516
	Information Extraction.....	517
	Natural Language Processing	518
13.6	Web Mining	519
	Wrapper Induction.....	519
	Page Rank	520
13.7	Images and Speech	522
	Images.....	523
	Speech.....	524
13.8	Adversarial Situations.....	524
13.9	Ubiquitous Data Mining.....	527
13.10	Further Reading and Bibliographic Notes	529
13.11	WEKA Implementations	532
	Appendix A: Theoretical foundations.....	533
	Appendix B: The WEKA workbench.....	553
	References.....	573
	Index	601