

Contents

Preface	vii
1 Introduction	1
1.1 What Is Data Mining?	2
1.2 Motivating Challenges	4
1.3 The Origins of Data Mining	6
1.4 Data Mining Tasks	7
1.5 Scope and Organization of the Book	11
1.6 Bibliographic Notes	13
1.7 Exercises	16
2 Data	19
2.1 Types of Data	22
2.1.1 Attributes and Measurement	23
2.1.2 Types of Data Sets	29
2.2 Data Quality	36
2.2.1 Measurement and Data Collection Issues	37
2.2.2 Issues Related to Applications	43
2.3 Data Preprocessing	44
2.3.1 Aggregation	45
2.3.2 Sampling	47
2.3.3 Dimensionality Reduction	50
2.3.4 Feature Subset Selection	52
2.3.5 Feature Creation	55
2.3.6 Discretization and Binarization	57
2.3.7 Variable Transformation	63
2.4 Measures of Similarity and Dissimilarity	65
2.4.1 Basics	66
2.4.2 Similarity and Dissimilarity between Simple Attributes .	67
2.4.3 Dissimilarities between Data Objects	69
2.4.4 Similarities between Data Objects	72

xiv Contents

2.4.5 Examples of Proximity Measures	73
2.4.6 Issues in Proximity Calculation	80
2.4.7 Selecting the Right Proximity Measure	83
2.5 Bibliographic Notes	84
2.6 Exercises	88
3 Exploring Data	97
3.1 The Iris Data Set	98
3.2 Summary Statistics	98
3.2.1 Frequencies and the Mode	99
3.2.2 Percentiles	100
3.2.3 Measures of Location: Mean and Median	101
3.2.4 Measures of Spread: Range and Variance	102
3.2.5 Multivariate Summary Statistics	104
3.2.6 Other Ways to Summarize the Data	105
3.3 Visualization	105
3.3.1 Motivations for Visualization	105
3.3.2 General Concepts	106
3.3.3 Techniques	110
3.3.4 Visualizing Higher-Dimensional Data	124
3.3.5 Do's and Don'ts	130
3.4 OLAP and Multidimensional Data Analysis	131
3.4.1 Representing Iris Data as a Multidimensional Array . .	131
3.4.2 Multidimensional Data: The General Case	133
3.4.3 Analyzing Multidimensional Data	135
3.4.4 Final Comments on Multidimensional Data Analysis .	139
3.5 Bibliographic Notes	139
3.6 Exercises	141
4 Classification:	
Basic Concepts, Decision Trees, and Model Evaluation	145
4.1 Preliminaries	146
4.2 General Approach to Solving a Classification Problem . . .	148
4.3 Decision Tree Induction	150
4.3.1 How a Decision Tree Works	150
4.3.2 How to Build a Decision Tree	151
4.3.3 Methods for Expressing Attribute Test Conditions . .	155
4.3.4 Measures for Selecting the Best Split	158
4.3.5 Algorithm for Decision Tree Induction	164
4.3.6 An Example: Web Robot Detection	166

73	4.3.7 Characteristics of Decision Tree Induction	168
80	4.4 Model Overfitting	172
83	4.4.1 Overfitting Due to Presence of Noise	175
84	4.4.2 Overfitting Due to Lack of Representative Samples . .	177
88	4.4.3 Overfitting and the Multiple Comparison Procedure .	178
97	4.4.4 Estimation of Generalization Errors	179
98	4.4.5 Handling Overfitting in Decision Tree Induction . .	184
98	4.5 Evaluating the Performance of a Classifier	186
99	4.5.1 Holdout Method	186
100	4.5.2 Random Subsampling	187
101	4.5.3 Cross-Validation	187
102	4.5.4 Bootstrap	188
104	4.6 Methods for Comparing Classifiers	188
105	4.6.1 Estimating a Confidence Interval for Accuracy . .	189
105	4.6.2 Comparing the Performance of Two Models	191
105	4.6.3 Comparing the Performance of Two Classifiers . .	192
106	4.7 Bibliographic Notes	193
106	4.8 Exercises	198
110		
124	5 Classification: Alternative Techniques	207
130	5.1 Rule-Based Classifier	207
131	5.1.1 How a Rule-Based Classifier Works	209
131	5.1.2 Rule-Ordering Schemes	211
133	5.1.3 How to Build a Rule-Based Classifier	212
135	5.1.4 Direct Methods for Rule Extraction	213
139	5.1.5 Indirect Methods for Rule Extraction	221
139	5.1.6 Characteristics of Rule-Based Classifiers	223
141	5.2 Nearest-Neighbor classifiers	223
145	5.2.1 Algorithm	225
146	5.2.2 Characteristics of Nearest-Neighbor Classifiers . .	226
146	5.3 Bayesian Classifiers	227
148	5.3.1 Bayes Theorem	228
150	5.3.2 Using the Bayes Theorem for Classification . . .	229
150	5.3.3 Naïve Bayes Classifier	231
151	5.3.4 Bayes Error Rate	238
155	5.3.5 Bayesian Belief Networks	240
158	5.4 Artificial Neural Network (ANN)	246
164	5.4.1 Perceptron	247
164	5.4.2 Multilayer Artificial Neural Network	251
166	5.4.3 Characteristics of ANN	255

xvi Contents

5.5	Support Vector Machine (SVM)	256
5.5.1	Maximum Margin Hyperplanes	256
5.5.2	Linear SVM: Separable Case	259
5.5.3	Linear SVM: Nonseparable Case	266
5.5.4	Nonlinear SVM	270
5.5.5	Characteristics of SVM	276
5.6	Ensemble Methods	276
5.6.1	Rationale for Ensemble Method	277
5.6.2	Methods for Constructing an Ensemble Classifier	278
5.6.3	Bias-Variance Decomposition	281
5.6.4	Bagging	283
5.6.5	Boosting	285
5.6.6	Random Forests	290
5.6.7	Empirical Comparison among Ensemble Methods	294
5.7	Class Imbalance Problem	294
5.7.1	Alternative Metrics	295
5.7.2	The Receiver Operating Characteristic Curve	298
5.7.3	Cost-Sensitive Learning	302
5.7.4	Sampling-Based Approaches	305
5.8	Multiclass Problem	306
5.9	Bibliographic Notes	309
5.10	Exercises	315
6	Association Analysis: Basic Concepts and Algorithms	327
6.1	Problem Definition	328
6.2	Frequent Itemset Generation	332
6.2.1	The <i>Apriori</i> Principle	333
6.2.2	Frequent Itemset Generation in the <i>Apriori</i> Algorithm	335
6.2.3	Candidate Generation and Pruning	338
6.2.4	Support Counting	342
6.2.5	Computational Complexity	345
6.3	Rule Generation	349
6.3.1	Confidence-Based Pruning	350
6.3.2	Rule Generation in <i>Apriori</i> Algorithm	350
6.3.3	An Example: Congressional Voting Records	352
6.4	Compact Representation of Frequent Itemsets	353
6.4.1	Maximal Frequent Itemsets	354
6.4.2	Closed Frequent Itemsets	355
6.5	Alternative Methods for Generating Frequent Itemsets	359
6.6	FP-Growth Algorithm	363

56	6.6.1	FP-Tree Representation	363
56	6.6.2	Frequent Itemset Generation in FP-Growth Algorithm	366
59	6.7	Evaluation of Association Patterns	370
66	6.7.1	Objective Measures of Interestingness	371
70	6.7.2	Measures beyond Pairs of Binary Variables	382
76	6.7.3	Simpson's Paradox	384
76	6.8	Effect of Skewed Support Distribution	386
77	6.9	Bibliographic Notes	390
78	6.10	Exercises	404
81			
83	7	Association Analysis: Advanced Concepts	415
85	7.1	Handling Categorical Attributes	415
90	7.2	Handling Continuous Attributes	418
94	7.2.1	Discretization-Based Methods	418
94	7.2.2	Statistics-Based Methods	422
95	7.2.3	Non-discretization Methods	424
98	7.3	Handling a Concept Hierarchy	426
02	7.4	Sequential Patterns	429
05	7.4.1	Problem Formulation	429
06	7.4.2	Sequential Pattern Discovery	431
09	7.4.3	Timing Constraints	436
15	7.4.4	Alternative Counting Schemes	439
27	7.5	Subgraph Patterns	442
28	7.5.1	Graphs and Subgraphs	443
32	7.5.2	Frequent Subgraph Mining	444
33	7.5.3	<i>Apriori</i> -like Method	447
35	7.5.4	Candidate Generation	448
38	7.5.5	Candidate Pruning	453
42	7.5.6	Support Counting	457
45	7.6	Infrequent Patterns	457
49	7.6.1	Negative Patterns	458
50	7.6.2	Negatively Correlated Patterns	458
52	7.6.3	Comparisons among Infrequent Patterns, Negative Patterns, and Negatively Correlated Patterns	460
53	7.6.4	Techniques for Mining Interesting Infrequent Patterns	461
54	7.6.5	Techniques Based on Mining Negative Patterns	463
55	7.6.6	Techniques Based on Support Expectation	465
59	7.7	Bibliographic Notes	469
63	7.8	Exercises	473

xviii Contents

8 Cluster Analysis: Basic Concepts and Algorithms	487
8.1 Overview	490
8.1.1 What Is Cluster Analysis?	490
8.1.2 Different Types of Clusterings	491
8.1.3 Different Types of Clusters	493
8.2 K-means	496
8.2.1 The Basic K-means Algorithm	497
8.2.2 K-means: Additional Issues	506
8.2.3 Bisecting K-means	508
8.2.4 K-means and Different Types of Clusters	510
8.2.5 Strengths and Weaknesses	510
8.2.6 K-means as an Optimization Problem	513
8.3 Agglomerative Hierarchical Clustering	515
8.3.1 Basic Agglomerative Hierarchical Clustering Algorithm	516
8.3.2 Specific Techniques	518
8.3.3 The Lance-Williams Formula for Cluster Proximity . .	524
8.3.4 Key Issues in Hierarchical Clustering	524
8.3.5 Strengths and Weaknesses	526
8.4 DBSCAN	526
8.4.1 Traditional Density: Center-Based Approach	527
8.4.2 The DBSCAN Algorithm	528
8.4.3 Strengths and Weaknesses	530
8.5 Cluster Evaluation	532
8.5.1 Overview	533
8.5.2 Unsupervised Cluster Evaluation Using Cohesion and Separation	536
8.5.3 Unsupervised Cluster Evaluation Using the Proximity Matrix	542
8.5.4 Unsupervised Evaluation of Hierarchical Clustering . .	544
8.5.5 Determining the Correct Number of Clusters	546
8.5.6 Clustering Tendency	547
8.5.7 Supervised Measures of Cluster Validity	548
8.5.8 Assessing the Significance of Cluster Validity Measures .	553
8.6 Bibliographic Notes	555
8.7 Exercises	559
9 Cluster Analysis: Additional Issues and Algorithms	569
9.1 Characteristics of Data, Clusters, and Clustering Algorithms .	570
9.1.1 Example: Comparing K-means and DBSCAN	570
9.1.2 Data Characteristics	571

187	9.1.3 Cluster Characteristics	573
190	9.1.4 General Characteristics of Clustering Algorithms	575
190	9.2 Prototype-Based Clustering	577
191	9.2.1 Fuzzy Clustering	577
193	9.2.2 Clustering Using Mixture Models	583
196	9.2.3 Self-Organizing Maps (SOM)	594
197	9.3 Density-Based Clustering	600
506	9.3.1 Grid-Based Clustering	601
508	9.3.2 Subspace Clustering	604
510	9.3.3 DENCLUE: A Kernel-Based Scheme for Density-Based Clustering	608
513	9.4 Graph-Based Clustering	612
515	9.4.1 Sparsification	613
516	9.4.2 Minimum Spanning Tree (MST) Clustering	614
518	9.4.3 OPOSSUM: Optimal Partitioning of Sparse Similarities Using METIS	616
524	9.4.4 Chameleon: Hierarchical Clustering with Dynamic Modeling	616
526	9.4.5 Shared Nearest Neighbor Similarity	622
527	9.4.6 The Jarvis-Patrick Clustering Algorithm	625
528	9.4.7 SNN Density	627
530	9.4.8 SNN Density-Based Clustering	629
532	9.5 Scalable Clustering Algorithms	630
533	9.5.1 Scalability: General Issues and Approaches	630
536	9.5.2 BIRCH	633
536	9.5.3 CURE	635
542	9.6 Which Clustering Algorithm?	639
544	9.7 Bibliographic Notes	643
546	9.8 Exercises	647
547	10 Anomaly Detection	651
548	10.1 Preliminaries	653
553	10.1.1 Causes of Anomalies	653
555	10.1.2 Approaches to Anomaly Detection	654
559	10.1.3 The Use of Class Labels	655
669	10.1.4 Issues	656
570	10.2 Statistical Approaches	658
570	10.2.1 Detecting Outliers in a Univariate Normal Distribution	659
571	10.2.2 Outliers in a Multivariate Normal Distribution	661
571	10.2.3 A Mixture Model Approach for Anomaly Detection	662

xx **Contents**

10.2.4	Strengths and Weaknesses	665
10.3	Proximity-Based Outlier Detection	666
10.3.1	Strengths and Weaknesses	666
10.4	Density-Based Outlier Detection	668
10.4.1	Detection of Outliers Using Relative Density	669
10.4.2	Strengths and Weaknesses	670
10.5	Clustering-Based Techniques	671
10.5.1	Assessing the Extent to Which an Object Belongs to a Cluster	672
10.5.2	Impact of Outliers on the Initial Clustering	674
10.5.3	The Number of Clusters to Use	674
10.5.4	Strengths and Weaknesses	674
10.6	Bibliographic Notes	675
10.7	Exercises	680
Appendix A Linear Algebra		685
A.1	Vectors	685
A.1.1	Definition	685
A.1.2	Vector Addition and Multiplication by a Scalar	685
A.1.3	Vector Spaces	687
A.1.4	The Dot Product, Orthogonality, and Orthogonal Projections	688
A.1.5	Vectors and Data Analysis	690
A.2	Matrices	691
A.2.1	Matrices: Definitions	691
A.2.2	Matrices: Addition and Multiplication by a Scalar	692
A.2.3	Matrices: Multiplication	693
A.2.4	Linear Transformations and Inverse Matrices	695
A.2.5	Eigenvalue and Singular Value Decomposition	697
A.2.6	Matrices and Data Analysis	699
A.3	Bibliographic Notes	700
Appendix B Dimensionality Reduction		701
B.1	PCA and SVD	701
B.1.1	Principal Components Analysis (PCA)	701
B.1.2	SVD	706
B.2	Other Dimensionality Reduction Techniques	708
B.2.1	Factor Analysis	708
B.2.2	Locally Linear Embedding (LLE)	710
B.2.3	Multidimensional Scaling, FastMap, and ISOMAP	712

365	B.2.4 Common Issues	715
366	B.3 Bibliographic Notes	716
366		
368	Appendix C Probability and Statistics	719
369	C.1 Probability	719
370	C.1.1 Expected Values	722
371	C.2 Statistics	723
372	C.2.1 Point Estimation	724
374	C.2.2 Central Limit Theorem	724
374	C.2.3 Interval Estimation	725
374	C.3 Hypothesis Testing	726
374		
375	Appendix D Regression	729
380	D.1 Preliminaries	729
385	D.2 Simple Linear Regression	730
685	D.2.1 Least Square Method	731
685	D.2.2 Analyzing Regression Errors	733
685	D.2.3 Analyzing Goodness of Fit	735
685	D.3 Multivariate Linear Regression	736
687	D.4 Alternative Least-Square Regression Methods	737
688		
690	Appendix E Optimization	739
691	E.1 Unconstrained Optimization	739
691	E.1.1 Numerical Methods	742
692	E.2 Constrained Optimization	746
692	E.2.1 Equality Constraints	746
693	E.2.2 Inequality Constraints	747
695		
697	Author Index	750
699		
700	Subject Index	758
701		
701	Copyright Permissions	769
701		
701		
706		
708		
708		
710		
712		

Subject Index

- accuracy, 149, 208
 - activation function, 248
 - AdaBoost, 288
 - adjacency matrix, 448
 - aggregation, 45–47
 - anomaly detection
 - applications, 651–652
 - approaches, 655–656
 - semi-supervised, *see* anomaly detection, semi-supervised
 - supervised, *see* anomaly detection, supervised
 - unsupervised, *see* anomaly detection, unsupervised
 - causes of, 653–654
 - clustering-based, 671–675
 - definition, 671
 - example, 672
 - impact of outliers, 674
 - membership in a cluster, 672
 - number of clusters, 674
 - strengths and weaknesses, 674
 - definitions, 654–655
 - density-based, 655
 - distance-based, 655
 - model-based, 654
 - density-based, 668–670
 - deviation detection, 651
 - distance based
 - definition, 666
 - exception mining, 651
 - issues, 656–658
 - outliers, 651
 - proximity-based, 666–668
 - distance-based, *see* anomaly detection, distance-based
 - strengths and weaknesses, 666
 - rare class, 654
 - relative density, 669–670
 - algorithm, 669
 - example, 669–670
 - semi-supervised, 656
 - statistical, 658–665
 - Gaussian, 659
 - Grubbs, 681
 - likelihood approach, 662
 - multivariate, 661
 - strengths and weaknesses, 665
 - strengths and weaknesses, 670
 - supervised, 655
 - techniques, 654–655
 - unsupervised, 656, 658–675
- Apriori
- algorithm, 335
 - principle, 333
- association
- analysis, 327
 - categorical attributes, 415
 - continuous attributes, 418
 - indirect, 467
 - pattern, 328
 - rule, *see* rule
- attribute, 22–29
- definition of, 23
 - number of values, 28
 - set, 146
 - type, 23–27
 - asymmetric, 28–29
 - binary, 28
 - continuous, 26, 28
 - discrete, 28
 - interval, 25, 26
 - nominal, 25, 26
 - ordinal, 25, 26

- qualitative, 26
- quantitative, 26
- ratio, 25
- axon, 246
- back-propagation, 254
- bagging, *see* classifier
- Bayes
 - classifier, *see* classifier
 - error, 238
 - naive, *see* classifier
 - network, *see* classifier
 - theorem, 228
- bias, 281
- bias variance decomposition, 281
- binarization, *see* discretization, binarization, 416, 419
- BIRCH, 633–635
- boosting, *see* classifier
- bootstrap, 188
- box plot, *see* visualization, box plot
- Bregman divergence, 79–80
- c4.5rules, 212
- candidate
 - elimination, 448
 - generation, 337, 338, 433, 448
 - itemset, 332
 - pruning, 338, 435, 453
 - rule, 350
 - sequence, 433
 - subgraph, 445
- capacity, 258
- case, *see* object
- chameleon, 616–622
 - algorithm, 620–621
 - graph partitioning, 621
 - merging strategy, 618
 - relative closeness, 618
 - relative interconnectivity, 619
 - self-similarity, 613, 618, 620, 621
 - strengths and limitations, 622
- characteristic, *see* attribute
- city block distance, *see* distance, city block
- class
 - imbalance, 294
 - label, 146
- classification
 - model, 146
- classifier
 - bagging, 283
 - base, 277
 - Bayes, 227
 - Bayesian belief, 240
 - boosting, 285
 - combination, 276
 - decision tree, 150
 - ensemble, 276
 - maximal margin, 259
 - naive-Bayes, 231
 - nearest-neighbor, 223
 - neural networks, 246
 - perceptron, 247
 - random forest, 290
 - Rote, 223
 - rule-based, 207
 - support vector machine, 256
 - unstable, 280
- climate indices, 630
- cluster analysis
 - algorithm characteristics, 575–576
 - mapping to another domain, 576
 - nondeterminism, 575
 - optimization, 576
 - order dependence, 575
 - parameter selection, *see* parameter selection
 - scalability, *see* scalability
 - applications, 487–489
 - as an optimization problem, 576
 - chameleon, *see* chameleon
 - choosing an algorithm, 639–642
 - cluster characteristics, 573–574
 - data distribution, 574
 - density, 574
 - poorly separated, 574
 - relationships, 574
 - shape, 574
 - size, 574
 - subspace, 574
 - cluster density, 574
 - cluster shape, 510, 574
 - cluster size, 574
 - data characteristics, 571–573
 - attribute types, 573
 - data types, 573
 - high-dimensionality, 572
 - mathematical properties, 573

- noise, 572
- outliers, 572
- scale, 573
- size, 572
- sparseness, 572
- DBSCAN, *see* DBSCAN
 - definition of, 487, 490
- DENCLUE, *see* DENCLUE
- density-based clustering, 600–612
- fuzzy clustering, *see* fuzzy clustering
- graph-based clustering, 612–630
 - sparsification, 613–614
- grid-based clustering, *see* grid-based clustering
- hierarchical, *see* hierarchical clustering
 - CURE, *see* CURE, *see* CURE
 - minimum spanning tree, 614–615
 - Jarvis-Patrick, *see* Jarvis-Patrick
 - K-means, *see* K-means
 - mixture modes, *see* mixture models
 - opossum, *see* opossum
 - parameter selection, 529, 546, 575
 - prototype-based clustering, 577–600
 - seeshared nearest neighbor, density-based clustering, 629
 - self-organizing maps, *see* self-organizing maps
 - subspace clustering, *see* subspace clustering
 - subspace clusters, 574
 - types of clusterings, 491–493
 - complete, 493
 - exclusive, 492
 - fuzzy, 492
 - hierarchical, 491
 - overlapping, 492
 - partial, 493
 - partitional, 491
 - types of clusters, 493–495
 - conceptual, 495
 - density-based, 494
 - graph-based, 494
 - prototype-based, 494
 - well-separated, 493
 - validation, *see* cluster validation
- cluster validation, 532–555
 - assessment of measures, 553–555
 - clustering tendency, 533, 547
- cohesion, 536–540
- cophenetic correlation, 545
- for individual clusters, 541
- for individual objects, 541
- hierarchical, 544, 552
- number of clusters, 546
- relative measures, 535
- separation, 536–540
- silhouette coefficient, 541–542
- supervised, 548–553
 - classification measures, 549–550
 - similarity measures, 550–552
- supervised measures, 535
- unsupervised, 536–548
- unsupervised measures, 535
 - with proximity matrix, 542–544
- codeword, 308
- compaction factor, 370
- concept hierarchy, 426
- conditional independence, 241
- confidence
 - factor, 208
 - interval, 186
 - level, 725
 - measure, *see* measure
- confusion matrix, 149
- constraint
 - maxgap, 437
 - maxspan, 437
 - mingap, 437
 - timing, 436
 - window size, 439
- constructive induction, 172
- contingency table, 372
- convex optimization, 262
- correlation
 - ϕ -coefficient, 375
 - limitation, 375
- cost
 - learning, 302
- cost-sensitive learning, *see* cost
- coverage, 208
- cross-validation, 187
- CURE, 635–639
 - algorithm, 636, 637
 - cluster feature, 634
 - clustering feature
 - tree, 634
 - use of partitioning, 638–639

- use of sampling, 638
- curse of dimensionality, 271
- dag*, *see* graph
- data
 - attribute, *see* attribute
 - attribute types, 573
 - cleaning, *see* data quality, data cleaning
 - distribution, 574
 - exploration, *see* data exploration
 - fragmentation, 170
 - high dimensional
 - problems with similarity, 622
 - high-dimensional, 572
 - market basket, 327
 - mathematical properties, 573
 - noise, 572
 - object, *see* object
 - outliers, 572
 - preprocessing, *see* preprocessing
 - quality, *see* data quality
 - scale, 573
 - set, *see* data set
 - similarity, *see* similarity
 - size, 572
 - sparse, 572
 - transformations, *see* transformations
 - types, 573
 - types of, 19, 22–36
- data exploration, 97–98
 - OLAP, *see* OLAP
 - statistics, *see* statistics
 - visualization, *see* visualization
- data quality, 19, 36–44
 - application issues, 43–44
 - data documentation, 44
 - relevance, 44
 - timeliness, 44
 - data cleaning, 36
 - errors, 37–43
 - accuracy, 39
 - artifacts, 39
 - bias, 39
 - collection, 37
 - duplicate data, 42–43
 - inconsistent values, 41–42
 - measurement, 37
 - missing values, 40–41
- noise, 37–39
- outliers, 40
- precision, 39
- significant digits, 39–40
- data set, 22
 - characteristics, 29–30
 - dimensionality, 29
 - resolution, 29
 - sparsity, 29
 - types of, 29–36
 - graph-based, 32
 - matrix, *see* matrix
 - ordered, 33–35
 - record, 30–32
 - sequence, 34
 - sequential, 33
 - spatial, 35
 - temporal, 33
 - time series, 35
 - transaction, 30
- DBSCAN, 526–532
 - algorithm, 528
 - comparison to K-means, 570–571
 - complexity, 528
 - definition of density, 527
 - parameter selection, 529
 - types of points, 527
 - border, 527
 - core, 527
 - noise, 527
- decision
 - boundary, 170
 - list, 211
 - stump, 284
 - tree, *see* classifier
- decision boundary
 - linear, 259
- deduplication, 43
- DENCLUE, 608–612
 - algorithm, 610
 - implementation issues, 610
 - kernel density estimation, 609–610
 - strengths and limitations, 611
- dendrite, 246
- density
 - center based, 527
- dimension, *see* attribute
- dimensionality
 - curse, 51–52

- dimensionality reduction, 50–52, 701–716
 factor analysis, 708–710
 FastMap, 713
 ISOMAP, 714–715
 issues, 715–716
 Locally Linear Embedding, 710–712
 multidimensional scaling, 712–713
 PCA, 52
 SVD, 52
 discretization, 57–63, 233
 association, *see* association
 binarization, 58–59
 clustering, 420
 equal frequency, 420
 equal width, 420
 of binary attributes, *see* discretization, binarization
 of categorical variables, 62–63
 of continuous attributes, 59–62
 supervised, 60–62
 unsupervised, 59–60
 disjunct, 207
 dissimilarity, 69–72, 79–80
 choosing, 83–84
 definition of, 66
 distance, *see* distance
 non-metric, 72
 transformations, 66–69
 distance, 69–71
 city block, 70
 Euclidean, 70, 690
 Hamming, 308
 L_1 norm, 70
 L_2 norm, 70
 L_∞ , 70
 L_{\max} , 70
 Mahalanobis, 81
 Manhattan, 70
 metric, 70–71
 positivity, 70
 symmetry, 71
 triangle inequality, 71
 Minkowski, 69–70
 supremum, 70
 distribution
 binomial, 183
 Gaussian, 183, 233
 eager learner, *see* learner
 edge, 443
 growing, 449
 element, 430
 EM algorithm, 587–591
 ensemble method, *see* classifier
 entity, *see* object
 entropy, 60, 158
 use in discretization, *see* discretization, supervised
 equivalent sample size, 236
 error
 apparent, 172
 error-correcting output coding, 307
 function, 253
 generalization, 172, 179, 258
 optimistic, 180
 pessimistic, 181
 resubstitution, 172
 training, 172
 error rate, 149
 Euclidean distance, *see* distance, Euclidean
 evaluation
 association, 370
 example, 146
 exhaustive, 210
 exploratory data analysis, *see* data exploration
 factor analysis, *see* dimensionality reduction, factor analysis
 false negative, 297
 false positive, 296
 FastMap, *see* dimensionality reduction, FastMap
 feature
 redundant, 169, 256
 feature construction, *see* feature creation, feature construction
 feature creation, 55–57
 feature construction, 57
 feature extraction, 55
 mapping data to a new space, 56–57
 feature extraction, *see* feature creation, feature extraction
 feature selection, 52–55
 architecture for, 53–54
 feature weighting, 55
 irrelevant features, 52
 redundant features, 52
 types of, 52–53

- embedded, 53
- filter, 53
- wrapper, 53
- field, *see* attribute
- Fourier transform, 56
- FP-growth, 363
- FP-tree, *see* tree
- frequent subgraph, 442
 - Apriori-like algorithm, 447
- fuzzy clustering, 577–582
 - fuzzy c-means, 579–582
 - algorithm, 579
 - centroids, 580
 - example, 582
 - initialization, 580
 - SSE, 580
 - strengths and limitations, 582
 - weight update, 581
 - fuzzy sets, 578
 - fuzzy pseudo-partition, 579
- gain ratio, 164
- general-to-specific, 216
- generalization, 148, *see* rule
- gini index, 158
- gradient descent, 254
- graph, 443
 - connected, 444
 - directed acyclic, 241, 426
 - isomorphism, 454
 - undirected, 444
- greedy search, 216
- grid-based clustering, 601–604
 - algorithm, 601
 - clusters, 603
 - density, 602
 - grid cells, 601
- hidden
 - layer, 251
 - node, 251
- hierarchical clustering, 515–526
 - agglomerative algorithm, 516
 - centroid methods, 523
 - cluster proximity, 517
 - Lance-Williams formula, 524
 - complete link, 517, 520–521
 - complexity, 518
 - group average, 517, 521–522
- inversions, 523
- MAX, *see* complete link
- merging decisions, 525
- MIN, *see* single link
- single link, 517, 519–520
- Ward's method, 523
- high-dimensionality
 - seed data, high-dimensional, 572
- histogram, *see* visualization, histogram
- holdout, 186
- hypothesis
 - alternative, 423, 727
 - null, 423, 727
- independence
 - conditional, 231
- information gain
 - entropy-based, 160
 - FOIL's, 218
- instance, 146
- interest, *see* measure
- ISOMAP, *see* dimensionality reduction, ISOMAP
- item, *see* attribute, 328
 - competing, 457
 - negative, 458
- itemset, 329
 - candidate, *see* candidate
 - closed, 355
 - frequent, 331
 - maximal, 354
- Jarvis-Patrick, 625–627
 - algorithm, 625
 - example, 626
 - strengths and limitations, 626
- K-means, 496–515
 - algorithm, 497–498
 - bisecting, 508–509
 - centroids, 499, 501
 - choosing initial, 501–505
 - comparison to DBSCAN, 570–571
 - complexity, 505
 - derivation, 513–515
 - empty clusters, 506
 - incremental, 508
 - limitations, 510–513
 - objective functions, 499, 501
 - outliers, 506

- reducing SEE, 507–508
- k-nearest neighbor graph, 613, 619, 620
- Karush-Kuhn-Tucker, *see* KKT
- kernel density estimation, 609–610
- kernel function, 273
- kernel trick, 273
- KKT, 263
- L₁ norm, *see* distance, L₁ norm
- L₂ norm, *see* distance, L₂ norm
- Lagrange multiplier, 262
- Lagrangian, 262
 - dual, 263
- Laplace, 217
- lazy learner, *see* learner
- learner
 - eager, 223, 226
 - lazy, 223, 226
- learning algorithm, 148
- learning rate, 254
- least squares, 699
- leave-one-out, 187
- lexicographic order, 341
- likelihood ratio, 217
- linear algebra, 685–700
 - matrix, *see* matrix
 - vector, *see* vector
- linear regression, 699
- linear separable, 256
- linear systems of equations, 699
- linear transformation, *see* matrix, linear transformation
- Locally Linear Embedding, *see* dimensionality reduction, Locally Linear Embedding
- m-estimate, 217, 236
- majority voting, *see* voting
- Manhattan distance, *see* distance, Manhattan
- margin
 - maximal, 256
 - soft, 266
- market basket data, *see* data
- matrix, 30, 691–697
 - addition, 692
 - column vector, 692
 - confusion, *see* confusion matrix
 - definition, 691–692
- document-term, 32
- eigenvalue, 697
- eigenvalue decomposition, 697–698
- eigenvector, 697
- in data analysis, 699–700
- inverse, 696–697
- linear combination, 703
- linear transformations, 695–697
 - column space, 696
 - left nullspace, 696
 - nullspace, 696
 - projection, 695
 - reflection, 695
 - rotation, 695
 - row space, 696
 - scaling, 695
- multiplication, 693–695
- permutation, 455
- positive semidefinite, 703
- rank, 696
- row vector, 692
- scalar multiplication, 692–693
- singular value, 698
- singular value decomposition, 698
- singular vector, 698
- sparse, 31
- maxgap, *see* constraint
- maximum likelihood estimation, 585–587
- maxspan, *see* constraint
- MDL, 182, 258
- mean, 233
- measure
 - asymmetric, 377
 - confidence, 329
 - consistency, 377
 - interest, 373
 - IS, 375
 - objective, 371
 - properties, 379
 - subjective, 371
 - support, 329
 - symmetric, 377
- measurement, 23–27
 - definition of, 23
 - scale, 23
 - permissible transformations, 26–27
 - types, 23–27
- Mercer's theorem, 274
- metric

- accuracy, 149
- performance, 149
- min-Apriori, 425
- mingap, *see* constraint
- minimum description length, *see* MDL
- missing values, *see* data quality, errors, missing values
- mixture models, 583–591
 - advantages and limitations, 591
 - definition of, 583–585
 - EM algorithm, 587–591
 - maximum likelihood estimation, 585–587
- model, 146
 - comparison, 188
 - descriptive, 146
 - evaluation, 186
 - overfitting, 165, 172
 - predictive, 147
 - selection, 186
- monotone, 258
- monotonicity, 334
- multiclass, 306
- multidimensional data analysis, *see* OLAP, multidimensional data analysis
- multidimensional scaling, *see* dimensionality reduction, multidimensional scaling
- multiple comparison, 178
- mutual exclusive, 210
- nearest-neighbor classifier, *see* classifier network
 - Bayesian, *see* classifier
 - feed-forward, 251
 - multilayer, *see* classifier
 - neural, *see* classifier
 - recurrent, 251
 - topology, 243
- neuron, 246
- node
 - hidden, *see* hidden
 - internal, 150
 - leaf, 150
 - non-terminal, 150
 - root, 150
 - terminal, 150
- noise, 175, 227, 281
- object, 22
- objective function, 261
- observation, *see* object
- Occam's razor, 181
- OLAP, 46, 131–139
 - computing aggregates, 135–137
 - cross tabulation, 137
 - data cube, 135–139
 - dicing, 135, 138
 - dimensionality reduction, 135, 138
 - drill-down, 135, 138–139
 - fact table, 134
 - MOLAP, 139
 - multidimensional data, 131–135
 - multidimensional data analysis, 135–139
 - pivoting, 138
 - ROLAP, 139
 - roll-up, 135, 138–139
 - slicing, 135, 138
- online analytical processing, *see* OLAP
- opossum, 616
 - algorithm, 616
 - strengths and weaknesses, 616
- outliers, *see* data quality
- overfitting, *see* model, 174
- oversampling, 305
- pattern
 - cross-support, 387
 - hyperclique, 390
 - infrequent, 457
 - negative, 458
 - negatively correlated, 458, 460
 - sequential, *see* sequential
 - subgraph, *see* subgraph
- PCA, 701–704
 - examples, 704
 - mathematics, 702–703
- perceptron, *see* classifier
 - learning, 248
- pie chart, *see* visualization, pie chart
- point, *see* object
- precision, 297
- precondition, 208
- preprocessing, 19, 44–65
 - aggregation, *see* aggregation
 - binarization, *see* discretization, binarization

- dimensionality reduction, 50
- discretization, *see* discretization
- feature creation, *see* feature creation
- feature selection, *see* feature selection
- sampling, *see* sampling
- transformations, *see* transformations
- prevalence, 213
- probability
 - class-conditional, 230
 - density, 234
 - posterior, 229
 - prior, 229
 - table, 241
- proximity, 65–84
 - choosing, 83–84
 - cluster, 517
 - definition of, 65
 - dissimilarity, *see* dissimilarity
 - distance, *see* distance
 - for simple attributes, 67–69
 - issues, 80–83
 - attribute weights, 82–83
 - combining proximities, 81–82
 - correlation, 81
 - standardization, 81
 - similarity, *see* similarity
 - transformations, 66–67
- pruning
 - post-pruning, 185
 - prepruning, 184
- quadratic programming, 264
- random forest
 - see also* classifier, 290
- random subsampling, 187
- recall, 297
- Receiver Operating Characteristic curve, *see* ROC
- record, *see* object, 146
- reduced error pruning, 201, 316
- Reproducing kernel Hilbert space, *see* RKHS
- RIPPER, 212, 220
- RKHS, 274
- ROC, 298
- Rote classifier, *see* classifier
- rule
 - antecedent, 208
- association, 329
- candidate, *see* candidate
- classification, 207
- consequent, 208
- evaluation, 216
- generalization, 422
- generation, 222, 331, 349, 422
- growing, 215, 220
- ordered, 211
- ordering, 222
 - class-based, 212
 - rule-based, 211
- pruning, 218
- quantitative, 418
 - discretization-based, 418
 - non-discretization, 424
 - statistics-based, 422
- redundant, 422
- specialization, 422
- unordered, 211
- validation, 423
- rule set, 207
- sample, *see* object
- sampling, 47–50, 305
 - approaches, 48
 - progressive, 50
 - random, 48
 - stratified, 48
 - with replacement, 48
 - without replacement, 48
 - sample size, 48–49
- scalability
 - clustering algorithms, 630–639
 - BIRCH, 633–635
 - CURE, 635–639
 - general issues, 630–633
 - scatter plot, *see* visualization, scatter plot
- segmentation, 491
- self-organizing maps, 594–600
 - algorithm, 594–597
 - applications, 598
 - strengths and limitations, 599
- sensitivity, 296
- sequence
 - data sequence, 431
 - definition, 430
- sequential
 - pattern, 429

- pattern discovery, 431
- timing constraints, *see* constraint
- sequential covering, 213
- shared nearest neighbor, 613
 - density, 627–628
 - density-based clustering, 629–630
 - algorithm, 629
 - example, 629
 - strengths and limitations, 630
 - principle, 613
 - similarity, 622–625
 - computation, 624–625
 - differences in density, 623–624
 - versus direct similarity, 625
- significance
 - level, 727
- similarity, 20, 72–79
 - choosing, 83–84
 - correlation, 76–79
 - cosine, 74–76, 690
 - definition of, 66
 - extended Jaccard, 76
 - Jaccard, 74
 - shared nearest neighbor, *see* shared nearest neighbor, similarity
- simple matching coefficient, 73–74
- Tanimoto, 76
 - transformations, 66–69
- Simpson's paradox, 384
- soft splitting, 194
- SOM, 574, *see* self-organizing maps
- specialization, *see* rule
- specific-to-general, 216
- specificity, 296
- split information, 164
- statistical databases, 139
- statistics
 - absolute average deviation, 103
 - covariance matrix, 702
 - frequencies, 99–100
 - interquartile range, 103
 - mean, 101–102
 - trimmed, 102
 - median, 101–102
 - median average deviation, 103
 - mode, 99–100
 - percentiles, 100–101
 - range, 102–103
 - skewness, 105
- summary, 98–105
- variance, 102–103
- stem and leaf plot, *see* visualization, stem and leaf plot
- stopping criterion, 216
- structural risk minimization, 258
- subclass, 147
- subgraph
 - core, 448
 - definition, 443
 - pattern, 442
 - support, *see* support
- subsequence, 430
 - contiguous, 438
- subspace clustering, 604–608
 - CLIQUE, 607
 - algorithm, 608
 - monotonicity property, 607
 - strengths and limitations, 607
 - example, 604
 - subtree
 - raising, 186
 - replacement, 186
- superclass, 147
- support
 - count, 329
 - counting, 342, 435, 439, 457
 - limitation, 372
 - measure, *see* measure
 - pruning, 334
 - sequence, 431
 - subgraph, 443
- support vector, 256
- support vector machine, *see* classifier
- SVD, 706–708
 - example, 706–708
 - mathematics, 706
- SVM, *see* classifier
 - linear, 259
 - linear separable, 262
 - nonlinear, 270
- svm
 - nonseparable, 266
- synapse, 246
- target
 - attribute, 146
 - function, 146
- taxonomy, *see* concept hierarchy

- test set, 149
- topological equivalence, 450
- training set, 148
- transaction, 328
 - extended, 428
 - width, 346
- transformations, 63–65
 - between similarity and dissimilarity, 66–67
 - normalization, 64–65
 - simple functions, 63–64
 - standardization, 64–65
- tree
 - conditional FP-tree, 369
 - decision, *see* classifier
 - FP-tree, 363
 - hash, 344
 - oblique, 172
 - pruning, 165
- triangle inequality, 71
- true negative, 296
- true positive, 296
- underfitting, 174
- undersampling, 305
- universal approximator, 255
- validation set, 184, 218
- variable, *see* attribute
- variance, 233, 281
- vector, 685–691
 - addition, 685–686
 - column, *see* matrix, column vector
 - definition, 685
 - dot product, 688–690
 - in data analysis, 690–691
 - linear independence, 689–690
 - mean, 691
 - multiplication by a scalar, 686–687
 - norm, 688
 - orthogonal, 687–689
 - orthogonal projection, 689
 - row, *see* matrix, row vector
 - space, 687–688
 - basis, 687
 - dimension, 687
 - independent components, 687
 - linear combination, 687
 - span, 687
- vector quantization, 489
- vertex, 443
 - growing, 448
- visualization, 105–131
 - animation, 123
 - basic concepts
 - arrangement, 108–109
 - representation, 107–108
 - selection, 109–110
 - box plot, 114
 - Chernoff faces, 126–130
 - contour plot, 121
 - correlation matrix, 125–126
 - cumulative distribution functions, 115–116
 - data matrix, 124–125
 - do's and don'ts, 130–131
 - histogram
 - basic, 111–112
 - two-dimensional, 113
 - motivations for, 105–106
 - parallel coordinates, 126
 - percentile plot, 116
 - pie chart, 114–115
 - scatter plot, 116–119
 - extended, 119
 - matrix, 116
 - three-dimensional, 119
 - two-dimensional, 116–119
 - slices, 122–123
 - star coordinates, 126–130
 - stem and leaf plot, 111
 - surface plot, 121–122
 - techniques, 110–131
 - for few attributes, 111–119
 - for many attributes, 124–130
 - for spatio-temporal data, 119–123
 - vector field plot, 122
- vote
 - majority, 304
- voting
 - distance-weighted, 226
 - majority, 226
- wavelet transform, 57
- Web crawler, 166
- window size, *see* constraint