

Index

Numbers and Symbols

.632 bootstrap, 371
 δ -bicluster algorithm, 517–518
 δ -pCluster, 518–519

A

absolute-error criterion, 455
absolute support, 246
abstraction levels, 281
accuracy
 attribute construction and, 105
 boosting, 382
 with bootstrap, 371
 classification, 377–385
 classifier, 330, 366
 with cross-validation, 370–371
 data, 84
 with holdout method, 370
 measures, 369
 random forests, 383
 with random subsampling, 370
 rule selection based on, 361
activation function, 402
active learning, 25, 430, 437
ad hoc data mining, 31
AdaBoost, 380–382
 algorithm illustration, 382
 TrAdaBoost, 436
adaptive probabilistic networks, 397
advanced data analysis, 3, 4
advanced database systems, 4
affinity matrix, 520, 521
agglomerative hierarchical method, 459
 AGNES, 459, 460
 divisive hierarchical clustering versus,
 459–460
Agglomerative Nesting (AGNES), 459, 460
aggregate cells, 189

aggregation, 112
 bootstrap, 379
 complex data types and, 166
 cube computation and, 193
 data cube, 110–111
 at multiple granularities, 230–231
 multiway array, 195–199
 simultaneous, 193, 195
AGNES. *See* Agglomerative Nesting
algebraic measures, 145
algorithms. *See specific algorithms*
all-confidence measure, 268, 272
all-versus-all (AVA), 430–431
analysis of variance (ANOVA), 600
analytical processing, 153
ancestor cells, 189
angle-based outlier detection (ABOD), 580
angle-based outlier factor (ABOF), 580
anomalies. *See* outliers
anomaly mining. *See* outlier analysis
anomaly-based detection, 614
antimonotonic constraints, 298, 301
antimonotonic measures, 194
antimonotonicity, 249
apex cuboids, 111, 138, 158
application domain-specific semantics, 282
applications, 33, 607–618
 business intelligence, 27
 computer science, 613
 domain-specific, 625
 engineering, 613, 624
 exploration, 623
 financial data analysis, 607–609
 intrusion detection/prevention, 614–615
 recommender systems, 615–618
 retail industry, 609–611
 science, 611–613
 social science and social studies, 613

- applications (*Continued*)
 - targeted, 27–28
 - telecommunications industry, 611
 - Web search engines, 28
- application-specific outlier detection, 548–549
- approximate patterns, 281
 - mining, 307–312
- Apriori algorithm, 248–253, 272
 - dynamic itemset counting, 256
 - efficiency, improving, 254–256
 - example, 250–252
 - hash-based technique, 255
 - join step, 249
 - partitioning, 255–256
 - prune step, 249–250
 - pseudocode, 253
 - sampling, 256
 - transaction reduction, 255
- Apriori property, 194, 201, 249
 - antimonotonicity, 249
 - in Apriori algorithm, 298
- Apriori pruning method, 194
- arrays
 - 3-D for dimensions, 196
 - sparse compression, 198–199
- association analysis, 17–18
- association rules, 245
 - approximate, 281
 - Boolean, 281
 - compressed, 281
 - confidence, 21, 245, 246, 416
 - constraint-based, 281
 - constraints, 296–297
 - correlation, 265, 272
 - discarded, 17
 - fittest, 426
 - frequent patterns and, 280
 - generation from frequent itemsets, 253, 254
 - hybrid-dimensional, 288
 - interdimensional, 288
 - intradimensional, 287
 - metarule-guided mining of, 295–296
 - minimum confidence threshold, 18, 245
 - minimum support threshold, 245
 - mining, 272
 - multidimensional, 17, 287–289, 320
 - multilevel, 281, 283–287, 320
 - near-match, 281
 - objective measures, 21
 - offspring, 426
 - quantitative, 281, 289, 320
 - redundancy-aware top-*k*, 281
 - single-dimensional, 17, 287
 - spatial, 595
 - strong, 264–265, 272
 - support, 21, 245, 246, 417
 - top-*k*, 281
 - types of values in, 281
- associative classification, 415, 416–419, 437
 - CBA, 417
 - CMAR, 417–418
 - CPAR, 418–419
 - rule confidence, 416
 - rule support, 417
 - steps, 417
- asymmetric binary dissimilarity, 71
- asymmetric binary similarity, 71
- attribute construction, 112
 - accuracy and, 105
 - multivariate splits, 344
- attribute selection measures, 331, 336–344
 - CHAID, 343
 - gain ratio, 340–341
 - Gini index, 341–343
 - information gain, 336–340
 - Minimum Description Length (MDL), 343–344
 - multivariate splits, 343–344
- attribute subset selection, 100, 103–105
 - decision tree induction, 105
 - forward selection/backward elimination combination, 105
 - greedy methods, 104–105
 - stepwise backward elimination, 105
 - stepwise forward selection, 105
- attribute vectors, 40, 328
- attribute-oriented induction (AOI), 166–178, 180
 - algorithm, 173
 - for class comparisons, 175–178
 - for data characterization, 167–172
 - data generalization by, 166–178
 - generalized relation, 172
 - implementation of, 172–174
- attributes, 9, 40
 - abstraction level differences, 99
 - behavioral, 546, 573
 - binary, 41–42, 79
 - Boolean, 41
 - categorical, 41
 - class label, 328
 - contextual, 546, 573
 - continuous, 44
 - correlated, 54–56
 - dimension correspondence, 10

- discrete, 44
- generalization, 169–170
- generalization control, 170
- generalization threshold control, 170
- grouping, 231
- interval-scaled, 43, 79
- of mixed type, 75–77
- nominal, 41, 79
- numeric, 43–44, 79
- ordered, 103
- ordinal, 41, 79
- qualitative, 41
- ratio-scaled, 43–44, 79
- reducts of, 427
- removal, 169
- repetition, 346
- set of, 118
- splitting, 333
- terminology for, 40
- type determination, 41
- types of, 39
- unordered, 103
- audio data mining, 604–607, 624
- automatic classification, 445
- AVA. *See* all-versus-all
- AVC-group, 347
- AVC-set, 347
- average()**, 215

B

- background knowledge, 30–31
- backpropagation, 393, 398–408, 437
 - activation function, 402
 - algorithm illustration, 401
 - biases, 402, 404
 - case updating, 404
 - efficiency, 404
 - epoch updating, 404
 - error, 403
 - functioning of, 400–403
 - hidden layers, 399
 - input layers, 399
 - input propagation, 401–402
 - interpretability and, 406–408
 - learning, 400
 - learning rate, 403–404
 - logistic function, 402
 - multilayer feed-forward neural network, 398–399
 - network pruning, 406–407
 - neural network topology definition, 400
 - output layers, 399

- sample learning calculations, 404–406
- sensitivity analysis, 408
- sigmoid function, 402
- squashing function, 403
- terminating conditions, 404
- unknown tuple classification, 406
- weights initialization, 401
- See also* classification
- bagging, 379–380
 - algorithm illustration, 380
 - boosting versus, 381–382
 - in building random forests, 383
- bar charts, 54
- base cells, 189
- base cuboids, 111, 137–138, 158
- Basic Local Alignment Search Tool (BLAST), 591
- Baum-Welch algorithm, 591
- Bayes' theorem, 350–351
- Bayesian belief networks, 393–397, 436
 - algorithms, 396
 - components of, 394
 - conditional probability table (CPT), 394, 395
 - directed acyclic graph, 394–395
 - gradient descent strategy, 396–397
 - illustrated, 394
 - mechanisms, 394–396
 - problem modeling, 395–396
 - topology, 396
 - training, 396–397
 - See also* classification
- Bayesian classification
 - basis, 350
 - Bayes' theorem, 350–351
 - class conditional independence, 350
 - naive, 351–355, 385
 - posterior probability, 351
 - prior probability, 351
- BCubed precision metric, 488, 489
- BCubed recall metric, 489
- behavioral attributes, 546, 573
- believability, data, 85
- BI (business intelligence), 27
- biases, 402, 404
- bicustering, 512–519, 538
 - application examples, 512–515
 - enumeration methods, 517, 518–519
 - gene expression example, 513–514
 - methods, 517–518
 - optimization-based methods, 517–518
 - recommender system example, 514–515
 - types of, 538

- biclusters, 511
 - with coherent values, 516
 - with coherent values on rows, 516
 - with constant values, 515
 - with constant values on columns, 515
 - with constant values on rows, 515
 - as submatrix, 515
 - types of, 515–516
- bimodal, 47
- bin boundaries, 89
- binary attributes, 41, 79
 - asymmetric, 42, 70
 - as Boolean, 41
 - contingency table for, 70
 - dissimilarity between, 71–72
 - example, 41–42
 - proximity measures, 70–72
 - symmetric, 42, 70–71
 - See also* attributes
- binning
 - discretization by, 115
 - equal-frequency, 89
 - smoothing by bin boundaries, 89
 - smoothing by bin means, 89
 - smoothing by bin medians, 89
- biological sequences, 586, 624
 - alignment of, 590–591
 - analysis, 590
 - BLAST, 590
 - hidden Markov model, 591
 - as mining trend, 624
 - multiple sequence alignment, 590
 - pairwise alignment, 590
 - phylogenetic tree, 590
 - substitution matrices, 590
- bipartite graphs, 523
- BIRCH, 458, 462–466
 - CF-trees, 462–463, 464, 465–466
 - clustering feature, 462, 463, 464
 - effectiveness, 465
 - multiphase clustering technique, 464–465
 - See also* hierarchical methods
- bitmap indexing, 160–161, 179
- bitmapped join indexing, 163, 179
- bivariate distribution, 40
- BLAST. *See* Basic Local Alignment Search Tool
- BOAT. *See* Bootstrapped Optimistic Algorithm for Tree construction
- Boolean association rules, 281
- Boolean attributes, 41
- boosting, 380
 - accuracy, 382

- AdaBoost, 380–382
- bagging versus, 381–382
 - weight assignment, 381
- bootstrap method, 371, 386
- bottom-up design approach, 133, 151–152
- bottom-up subspace search, 510–511
- boxplots, 49
 - computation, 50
 - example, 50
 - five-number summary, 49
 - illustrated, 50
 - in outlier visualization, 555
- BUC, 200–204, 235
 - for 3-D data cube computation, 200
 - algorithm, 202
 - Apriori property, 201
 - bottom-up construction, 201
 - iceberg cube construction, 201
 - partitioning snapshot, 203
 - performance, 204
 - top-down processing order, 200, 201
- business intelligence (BI), 27
- business metadata, 135
- business query view, 151

C

- C4.5, 332, 385
 - class-based ordering, 358
 - gain ratio use, 340
 - greedy approach, 332
 - pessimistic pruning, 345
 - rule extraction, 358
 - See also* decision tree induction
- cannot-link constraints, 533
- CART, 332, 385
 - cost complexity pruning algorithm, 345
 - Gini index use, 341
 - greedy approach, 332
 - See also* decision tree induction
- case updating, 404
- case-based reasoning (CBR), 425–426
 - challenges, 426
- categorical attributes, 41
- CBA. *See* Classification Based on Associations
- CBLOF. *See* cluster-based local outlier factor
- CELL method, 562, 563
- cells, 10–11
 - aggregate, 189
 - ancestor, 189
 - base, 189
 - descendant, 189

- dimensional, 189
- exceptions, 231
- residual value, 234
- central tendency measures, 39, 44, 45–47
 - mean, 45–46
 - median, 46–47
 - midrange, 47
 - for missing values, 88
 - models, 47
- centroid distance, 108
- CF-trees, 462–463, 464
 - nodes, 465
 - parameters, 464
 - structure illustration, 464
- CHAID, 343
- Chameleon, 459, 466–467
 - clustering illustration, 466
 - relative closeness, 467
 - relative interconnectivity, 466–467
 - See also* hierarchical methods
- Chernoff faces, 60
 - asymmetrical, 61
 - illustrated, 62
- ChiMerge, 117
- chi-square test, 95
- chunking, 195
- chunks, 195
 - 2-D, 197
 - 3-D, 197
 - computation of, 198
 - scanning order, 197
- CLARA. *See* Clustering Large Applications
- CLARANS. *See* Clustering Large Applications based upon Randomized Search
- class comparisons, 166, 175, 180
 - attribute-oriented induction for, 175–178
 - mining, 176
 - presentation of, 175–176
 - procedure, 175–176
- class conditional independence, 350
- class imbalance problem, 384–385, 386
 - ensemble methods for, 385
 - on multiclass tasks, 385
 - oversampling, 384–385, 386
 - threshold-moving approach, 385
 - undersampling, 384–385, 386
- class label attributes, 328
- class-based ordering, 357
- class/concept descriptions, 15
- classes, 15, 166
 - contrasting, 15
 - equivalence, 427
 - target, 15
- classification, 18, 327–328, 385
 - accuracy, 330
 - accuracy improvement techniques, 377–385
 - active learning, 433–434
 - advanced methods, 393–442
 - applications, 327
 - associative, 415, 416–419, 437
 - automatic, 445
 - backpropagation, 393, 398–408, 437
 - bagging, 379–380
 - basic concepts, 327–330
 - Bayes methods, 350–355
 - Bayesian belief networks, 393–397, 436
 - boosting, 380–382
 - case-based reasoning, 425–426
 - of class-imbalanced data, 383–385
 - confusion matrix, 365–366, 386
 - costs and benefits, 373–374
 - decision tree induction, 330–350
 - discriminative frequent pattern-based, 437
 - document, 430
 - ensemble methods, 378–379
 - evaluation metrics, 364–370
 - example, 19
 - frequent pattern-based, 393, 415–422, 437
 - fuzzy set approaches, 428–429, 437
 - general approach to, 328
 - genetic algorithms, 426–427, 437
 - heterogeneous networks, 593
 - homogeneous networks, 593
 - IF-THEN rules for, 355–357
 - interpretability, 369
 - k*-nearest-neighbor, 423–425
 - lazy learners, 393, 422–426
 - learning step, 328
 - model representation, 18
 - model selection, 364, 370–377
 - multiclass, 430–432, 437
 - in multimedia data mining, 596
 - neural networks for, 19, 398–408
 - pattern-based, 282, 318
 - perception-based, 348–350
 - precision measure, 368–369
 - as prediction problem, 328
 - process, 328
 - process illustration, 329
 - random forests, 382–383
 - recall measure, 368–369
 - robustness, 369
 - rough set approach, 427–428, 437

- classification (*Continued*)
 - rule-based, 355–363, 386
 - scalability, 369
 - semi-supervised, 432–433, 437
 - sentiment, 434
 - spatial, 595
 - speed, 369
 - support vector machines (SVMs), 393, 408–415, 437
 - transfer learning, 434–436
 - tree pruning, 344–347, 385
 - web-document, 435
- Classification Based on Associations (CBA), 417
- Classification based on Multiple Association Rules (CMAR), 417–418
- Classification based on Predictive Association Rules (CPAR), 418–419
- classification-based outlier detection, 571–573, 582
 - one-class model, 571–572
 - semi-supervised learning, 572
 - See also* outlier detection
- classifiers, 328
 - accuracy, 330, 366
 - bagged, 379–380
 - Bayesian, 350, 353
 - case-based reasoning, 425–426
 - comparing with ROC curves, 373–377
 - comparison aspects, 369
 - decision tree, 331
 - error rate, 367
 - k*-nearest-neighbor, 423–425
 - Naive Bayesian, 351–352
 - overfitting data, 330
 - performance evaluation metrics, 364–370
 - recognition rate, 366–367
 - rule-based, 355
- Clementine, 603, 606
- CLIQUE, 481–483
 - clustering steps, 481–482
 - effectiveness, 483
 - strategy, 481
 - See also* cluster analysis; grid-based methods
- closed data cubes, 192
- closed frequent itemsets, 247, 308
 - example, 248
 - mining, 262–264
 - shortcomings for compression, 308–309
- closed graphs, 591
- closed patterns, 280
 - top-*k* most frequent, 307
- closure checking, 263–264
- cloud computing, 31
- cluster analysis, 19–20, 443–495
 - advanced, 497–541
 - agglomerative hierarchical clustering, 459–461
 - applications, 444, 490
 - attribute types and, 446
 - as automatic classification, 445
 - bicustering, 511, 512–519
 - BIRCH, 458, 462–466
 - Chameleon, 458, 466–467
 - CLIQUE, 481–483
 - clustering quality measurement, 484, 487–490
 - clustering tendency assessment, 484–486
 - constraint-based, 447, 497, 532–538
 - correlation-based, 511
 - as data redundancy technique, 108
 - as data segmentation, 445
 - DBSCAN, 471–473
 - DENCLUE, 476–479
 - density-based methods, 449, 471–479, 491
 - in derived space, 519–520
 - dimensionality reduction methods, 519–522
 - discretization by, 116
 - distance measures, 461–462
 - distance-based, 445
 - divisive hierarchical clustering, 459–461
 - evaluation, 483–490, 491
 - example, 20
 - expectation-maximization (EM) algorithm, 505–508
 - graph and network data, 497, 522–532
 - grid-based methods, 450, 479–483, 491
 - heterogeneous networks, 593
 - hierarchical methods, 449, 457–470, 491
 - high-dimensional data, 447, 497, 508–522
 - homogeneous networks, 593
 - in image recognition, 444
 - incremental, 446
 - interpretability, 447
 - k*-means, 451–454
 - k*-medoids, 454–457
 - k*-modes, 454
 - in large databases, 445
 - as learning by observation, 445
 - low-dimensional, 509
 - methods, 448–451
 - multiple-phase, 458–459
 - number of clusters determination, 484, 486–487
 - OPTICS, 473–476
 - orthogonal aspects, 491
 - for outlier detection, 445
 - outlier detection and, 543

- partitioning methods, 448, 451–457, 491
- pattern, 282, 308–310
- probabilistic hierarchical clustering, 467–470
- probability model-based, 497–508
- PROCLUS, 511
- requirements, 445–448, 490–491
- scalability, 446
- in search results organization, 444
- spatial, 595
- spectral, 519–522
- as standalone tool, 445
- STING, 479–481
- subspace, 318–319, 448
- subspace search methods, 510–511
- taxonomy formation, 20
- techniques, 443, 444
- as unsupervised learning, 445
- usability, 447
- use of, 444
- cluster computing, 31
- cluster samples, 108–109
- cluster-based local outlier factor (CBLOF), 569–570
- clustering. *See* cluster analysis
- clustering features, 462, 463, 464
- Clustering Large Applications based upon Randomized Search (CLARANS), 457
- Clustering Large Applications (CLARA), 456–457
- clustering quality measurement, 484t, 487–490
 - cluster completeness, 488
 - cluster homogeneity, 487–488
 - extrinsic methods, 487–489
 - intrinsic methods, 487, 489–490
 - rag bag, 488
 - silhouette coefficient, 489–490
 - small cluster preservation, 488
- clustering space, 448
- clustering tendency assessment, 484–486
 - homogeneous hypothesis, 486
 - Hopkins statistic, 484–485
 - nonhomogeneous hypothesis, 486
 - nonuniform distribution of data, 484*See also* cluster analysis
- clustering with obstacles problem, 537
- clustering-based methods, 552, 567–571
 - example, 553
 - See also* outlier detection
- clustering-based outlier detection, 567–571, 582
 - approaches, 567
 - distance to closest cluster, 568–569
 - fixed-width clustering, 570
 - intrusion detection by, 569–570
 - objects not belonging to a cluster, 568
 - in small clusters, 570–571
 - weakness of, 571
- clustering-based quantitative associations, 290–291
- clusters, 66, 443, 444, 490
 - arbitrary shape, discovery of, 446
 - assignment rule, 497–498
 - completeness, 488
 - constraints on, 533
 - cuts and, 529–530
 - density-based, 472
 - determining number of, 484, 486–487
 - discovery of, 318
 - fuzzy, 499–501
 - graph clusters, finding, 528–529
 - on high-dimensional data, 509
 - homogeneity, 487–488
 - merging, 469, 470
 - ordering, 474–475, 477
 - pattern-based, 516
 - probabilistic, 502–503
 - separation of, 447
 - shapes, 471
 - small, preservation, 488
- CMAR. *See* Classification based on Multiple Association Rules
- CN2, 359, 363
- collaborative recommender systems, 610, 617, 618
- collective outlier detection, 548, 582
 - categories of, 576
 - contextual outlier detection versus, 575
 - on graph data, 576
 - structure discovery, 575
- collective outliers, 575, 581
 - mining, 575–576
- co-location patterns, 319, 595
- colossal patterns, 302, 320
 - core descendants, 305, 306
 - core patterns, 304–305
 - illustrated, 303
 - mining challenge, 302–303
 - Pattern-Fusion mining, 302–307
- combined significance, 312
- complete-linkage algorithm, 462
- completeness
 - data, 84–85
 - data mining algorithm, 22
- complex data types, 166
 - biological sequence data, 586, 590–591
 - graph patterns, 591–592
 - mining, 585–598, 625
 - networks, 591–592
 - in science applications, 612

- complex data types (*Continued*)
 - summary, 586
 - symbolic sequence data, 586, 588–590
 - time-series data, 586, 587–588
- composite join indices, 162
- compressed patterns, 281
 - mining, 307–312
 - mining by pattern clustering, 308–310
- compression, 100, 120
 - lossless, 100
 - lossy, 100
 - theory, 601
- computer science applications, 613
- concept characterization, 180
- concept comparison, 180
- concept description, 166, 180
- concept hierarchies, 142, 179
 - for generalizing data, 150
 - illustrated, 143, 144
 - implicit, 143
 - manual provision, 144
 - multilevel association rule mining with, 285
 - multiple, 144
 - for nominal attributes, 284
 - for specializing data, 150
- concept hierarchy generation, 112, 113, 120
 - based on number of distinct values, 118
 - illustrated, 112
 - methods, 117–119
 - for nominal data, 117–119
 - with prespecified semantic connections, 119
 - schema, 119
- conditional probability table (CPT), 394, 395–396
- confidence, 21
 - association rule, 21
 - interval, 219–220
 - limits, 373
 - rule, 245, 246
- conflict resolution strategy, 356
- confusion matrix, 365–366, 386
 - illustrated, 366
- connectionist learning, 398
- consecutive rules, 92
- Constrained Vector Quantization Error (CVQE)
 - algorithm, 536
- constraint-based clustering, 447, 497, 532–538, 539
 - categorization of constraints and, 533–535
 - hard constraints, 535–536
 - methods, 535–538
 - soft constraints, 536–537
 - speeding up, 537–538
 - See also* cluster analysis
- constraint-based mining, 294–301, 320
 - interactive exploratory mining/analysis, 295
 - as mining trend, 623
- constraint-based patterns/rules, 281
- constraint-based sequential pattern mining, 589
- constraint-guided mining, 30
- constraints
 - antimonotonic, 298, 301
 - association rule, 296–297
 - cannot-link, 533
 - on clusters, 533
 - coherence, 535
 - conflicting, 535
 - convertible, 299–300
 - data, 294
 - data-antimonotonic, 300
 - data-pruning, 300–301, 320
 - data-succinct, 300
 - dimension/level, 294, 297
 - hard, 534, 535–536, 539
 - inconvertible, 300
 - on instances, 533, 539
 - interestingness, 294, 297
 - knowledge type, 294
 - monotonic, 298
 - must-link, 533, 536
 - pattern-pruning, 297–300, 320
 - rules for, 294
 - on similarity measures, 533–534
 - soft, 534, 536–537, 539
 - succinct, 298–299
- content-based retrieval, 596
- context indicators, 314
- context modeling, 316
- context units, 314
- contextual attributes, 546, 573
- contextual outlier detection, 546–547, 582
 - with identified context, 574
 - normal behavior modeling, 574–575
 - structures as contexts, 575
 - summary, 575
 - transformation to conventional outlier detection, 573–574
- contextual outliers, 545–547, 573, 581
 - example, 546, 573
 - mining, 573–575
- contingency tables, 95
- continuous attributes, 44
- contrasting classes, 15, 180
 - initial working relations, 177
 - prime relation, 175, 177
- convertible constraints, 299–300

COP k -means algorithm, 536
 core descendants, 305
 colossal patterns, 306
 merging of core patterns, 306
 core patterns, 304–305
 core ratio, 305
 correlation analysis, 94
 discretization by, 117
 interestingness measures, 264
 with lift, 266–267
 nominal data, 95–96
 numeric data, 96–97
 redundancy and, 94–98
 correlation coefficient, 94, 96
 numeric data, 96–97
 correlation rules, 265, 272
 correlation-based clustering methods, 511
 correlations, 18
 cosine measure, 268
 cosine similarity, 77
 between two term-frequency vectors, 78
 cost complexity pruning algorithm, 345
 cotraining, 432–433
 covariance, 94, 97
 numeric data, 97–98
 CPAR. *See* Classification based on Predictive Association Rules
 credit policy analysis, 608–609
 CRM. *See* customer relationship management
 crossover operation, 426
 cross-validation, 370–371, 386
 k -fold, 370
 leave-one-out, 371
 in number of clusters determination, 487
 stratified, 371
 cube gradient analysis, 321
 cube shells, 192, 211
 computing, 211
 cube space
 discovery-driven exploration, 231–234
 multidimensional data analysis in, 227–234
 prediction mining in, 227
 subspaces, 228–229
 cuboid trees, 205
 cuboids, 137
 apex, 111, 138, 158
 base, 111, 137–138, 158
 child, 193
 individual, 190
 lattice of, 139, 156, 179, 188–189,
 234, 290
 sparse, 190

 subset selection, 160
 See also data cubes
 curse of dimensionality, 158, 179
 customer relationship management (CRM),
 619
 customer retention analysis, 610
 CVQE. *See* Constrained Vector Quantization Error
 algorithm
 cyber-physical systems (CPS), 596, 623–624

D

data
 antimonotonicity, 300
 archeology, 6
 biological sequence, 586, 590–591
 complexity, 32
 conversion to knowledge, 2
 cyber-physical system, 596
 for data mining, 8
 data warehouse, 13–15
 database, 9–10
 discrimination, 16
 dredging, 6
 generalizing, 150
 graph, 14
 growth, 2
 linearly inseparable, 413–415
 linearly separated, 409
 multimedia, 14, 596
 multiple sources, 15, 32
 multivariate, 556
 networked, 14
 overfitting, 330
 relational, 10
 sample, 219
 similarity and dissimilarity measures, 65–78
 skewed, 47, 271
 spatial, 14, 595
 spatiotemporal, 595–596
 specializing, 150
 statistical descriptions, 44–56
 streams, 598
 symbolic sequence, 586, 588–589
 temporal, 14
 text, 14, 596–597
 time-series, 586, 587
 “tombs,” 5
 training, 18
 transactional, 13–14
 types of, 33
 web, 597–598
 data auditing tools, 92

- data characterization, 15, 166
 - attribute-oriented induction, 167–172
 - data mining query, 167–168
 - example, 16
 - methods, 16
 - output, 16
- data classification. *See* classification
- data cleaning, 6, 85, 88–93, 120
 - in back-end tools/utilities, 134
 - binning, 89–90
 - discrepancy detection, 91–93
 - by information network analysis, 592–593
 - missing values, 88–89
 - noisy data, 89
 - outlier analysis, 90
 - pattern mining for, 318
 - as process, 91–93
 - regression, 90
 - See also* data preprocessing
- data constraints, 294
 - antimonotonic, 300
 - pruning data space with, 300–301
 - succinct, 300
 - See also* constraints
- data cube aggregation, 110–111
- data cube computation, 156–160, 214–215
 - aggregation and, 193
 - `average()`, 215
 - BUC, 200–204, 235
 - cube operator, 157–159
 - cube shells, 211
 - full, 189–190, 195–199
 - general strategies for, 192–194
 - iceberg, 160, 193–194
 - memory allocation, 199
 - methods, 194–218, 235
 - multiway array aggregation, 195–199
 - one-pass, 198
 - preliminary concepts, 188–194
 - shell fragments, 210–218, 235
 - Star-Cubing, 204–210, 235
- data cubes, 10, 136, 178, 188
 - 3-D, 138
 - 4-D, 138, 139
 - apex cuboid, 111, 138, 158
 - base cuboid, 111, 137–138, 158
 - closed, 192
 - cube shell, 192
 - cuboids, 137
 - curse of dimensionality, 158
 - discovery-driven exploration, 231–234
 - example, 11–13
 - full, 189–190, 196–197
 - gradient analysis, 321
 - iceberg, 160, 190–191, 201, 235
 - lattice of cuboids, 157, 234, 290
 - materialization, 159–160, 179, 234
 - measures, 145
 - multidimensional, 12, 136–139
 - multidimensional data mining and, 26
 - multifeature, 227, 230–231, 235
 - multimedia, 596
 - prediction, 227–230, 235
 - qualitative association mining, 289–290
 - queries, 230
 - query processing, 218–227
 - ranking, 225–227, 235
 - sampling, 218–220, 235
 - shell, 160, 211
 - shell fragments, 192, 210–218, 235
 - sparse, 190
 - spatial, 595
 - technology, 187–242
- data discretization. *See* discretization
- data dispersion, 44, 48–51
 - boxplots, 49–50
 - five-number summary, 49
 - quartiles, 48–49
 - standard deviation, 50–51
 - variance, 50–51
- data extraction, in back-end tools/utilities, 134
- data focusing, 168
- data generalization, 179–180
 - by attribute-oriented induction, 166–178
- data integration, 6, 85–86, 93–99, 120
 - correlation analysis, 94–98
 - detection/resolution of data value conflicts, 99
 - entity identification problem, 94
 - by information network analysis, 592–593
 - object matching, 94
 - redundancy and, 94–98
 - schema, 94
 - tuple duplication, 98–99
 - See also* data preprocessing
- data marts, 132, 142
 - data warehouses versus, 142
 - dependent, 132
 - distributed, 134
 - implementation, 132
 - independent, 132
- data matrix, 67–68
 - dissimilarity matrix versus, 67–68
 - relational table, 67–68

- rows and columns, 68
 - as two-mode matrix, 68
- data migration tools, 93
- data mining, 5–8, 33, 598, 623
 - ad hoc, 31
 - applications, 607–618
 - biological data, 624
 - complex data types, 585–598, 625
 - cyber-physical system data, 596
 - data streams, 598
 - data types for, 8
 - data warehouses for, 154
 - database types and, 32
 - descriptive, 15
 - distributed, 615, 624
 - efficiency, 31
 - foundations, views on, 600–601
 - functionalities, 15–23, 34
 - graphs and networks, 591–594
 - incremental, 31
 - as information technology evolution, 2–5
 - integration, 623
 - interactive, 30
 - as interdisciplinary effort, 29–30
 - invisible, 33, 618–620, 625
 - issues in, 29–33, 34
 - in knowledge discovery, 7
 - as knowledge search through data, 6
 - machine learning similarities, 26
 - methodologies, 29–30, 585–607
 - motivation for, 1–5
 - multidimensional, 11–13, 26, 33–34, 155–156, 179, 227–230
 - multimedia data, 596
 - OLAP and, 154
 - as pattern/knowledge discovery process, 8
 - predictive, 15
 - presentation/visualization of results, 31
 - privacy-preserving, 32, 621–622, 624–625, 626
 - query languages, 31
 - relational databases, 10
 - scalability, 31
 - sequence data, 586
 - social impacts, 32
 - society and, 618–622
 - spatial data, 595
 - spatiotemporal data and moving objects, 595–596, 623–624
 - statistical, 598
 - text data, 596–597, 624
 - trends, 622–625, 626
 - ubiquitous, 618–620, 625
 - user interaction and, 30–31
 - visual and audio, 602–607, 624, 625
 - Web data, 597–598, 624
- data mining systems, 10
- data models
 - entity-relationship (ER), 9, 139
 - multidimensional, 135–146
- data objects, 40, 79
 - similarity, 40
 - terminology for, 40
- data preprocessing, 83–124
 - cleaning, 88–93
 - forms illustration, 87
 - integration, 93–99
 - overview, 84–87
 - quality, 84–85
 - reduction, 99–111
 - in science applications, 612
 - summary, 87
 - tasks in, 85–87
 - transformation, 111–119
- data quality, 84, 120
 - accuracy, 84
 - believability, 85
 - completeness, 84–85
 - consistency, 85
 - interpretability, 85
 - timeliness, 85
- data reduction, 86, 99–111, 120
 - attribute subset selection, 103–105
 - clustering, 108
 - compression, 100, 120
 - data cube aggregation, 110–111
 - dimensionality, 86, 99–100, 120
 - histograms, 106–108
 - numerosity, 86, 100, 120
 - parametric, 105–106
 - principle components analysis, 102–103
 - sampling, 108
 - strategies, 99–100
 - theory, 601
 - wavelet transforms, 100–102
 - See also* data preprocessing
- data rich but information poor, 5
- data scrubbing tools, 92
- data security-enhancing techniques, 621
- data segmentation, 445
- data selection, 8
- data source view, 151
- data streams, 14, 598, 624
- data transformation, 8, 87, 111–119, 120
 - aggregation, 112

- data transformation (*Continued*)
 - attribute construction, 112
 - in back-end tools/utilities, 134
 - concept hierarchy generation, 112, 120
 - discretization, 111, 112, 120
 - normalization, 112, 113–115, 120
 - smoothing, 112
 - strategies, 112–113
 - See also* data preprocessing
- data types
 - complex, 166
 - complex, mining, 585–598
 - for data mining, 8
- data validation, 592–593
- data visualization, 56–65, 79, 602–603
 - complex data and relations, 64–65
 - geometric projection techniques, 58–60
 - hierarchical techniques, 63–64
 - icon-based techniques, 60–63
 - mining process, 603
 - mining result, 603, 605
 - pixel-oriented techniques, 57–58
 - in science applications, 613
 - summary, 65
 - tag clouds, 64, 66
 - techniques, 39–40
- data warehouses, 10–13, 26, 33, 125–185
 - analytical processing, 153
 - back-end tools/utilities, 134, 178
 - basic concepts, 125–135
 - bottom-up design approach, 133, 151–152
 - business analysis framework for, 150
 - business query view, 151
 - combined design approach, 152
 - data mart, 132, 142
 - data mining, 154
 - data source view, 151
 - design process, 151
 - development approach, 133
 - development tools, 153
 - dimensions, 10
 - enterprise, 132
 - extractors, 151
 - fact constellation, 141–142
 - for financial data, 608
 - framework illustration, 11
 - front-end client layer, 132
 - gateways, 131
 - geographic, 595
 - implementation, 156–165
 - information processing, 153
 - integrated, 126
 - metadata, 134–135
 - modeling, 10, 135–150
 - models, 132–134
 - multitier, 134
 - multitiered architecture, 130–132
 - nonvolatile, 127
 - OLAP server, 132
 - operational database systems versus, 128–129
 - planning and analysis tools, 153
 - retail industry, 609–610
 - in science applications, 612
 - snowflake schema, 140–141
 - star schema, 139–140
 - subject-oriented, 126
 - three-tier architecture, 131, 178
 - time-variant, 127
 - tools, 11
 - top-down design approach, 133, 151
 - top-down view, 151
 - update-driven approach, 128
 - usage for information processing, 153
 - view, 151
 - virtual, 133
 - warehouse database server, 131
- database management systems (DBMSs), 9
- database queries. *See* queries
- databases, 9
 - inductive, 601
 - relational. *See* relational databases
 - research, 26
 - statistical, 148–149
 - technology evolution, 3
 - transactional, 13–15
 - types of, 32
 - web-based, 4
- data/pattern analysis. *See* data mining
- DBSCAN, 471–473
 - algorithm illustration, 474
 - core objects, 472
 - density estimation, 477
 - density-based cluster, 472
 - density-connected, 472, 473
 - density-reachable, 472, 473
 - directly density-reachable, 472
 - neighborhood density, 471
 - See also* cluster analysis; density-based methods
- DDPMine, 422
- decimal scaling, normalization by, 115
- decision tree analysis, discretization by, 116
- decision tree induction, 330–350, 385
 - algorithm differences, 336
 - algorithm illustration, 333

- attribute selection measures, 336–344
- attribute subset selection, 105
- C4.5, 332
- CART, 332
- CHAID, 343
- gain ratio, 340–341
- Gini index, 332, 341–343
- ID3, 332
- incremental versions, 336
- information gain, 336–340
- multivariate splits, 344
- parameters, 332
- scalability and, 347–348
- splitting criterion, 333
- from training tuples, 332–333
- tree pruning, 344–347, 385
- visual mining for, 348–350
- decision trees, 18, 330
 - branches, 330
 - illustrated, 331
 - internal nodes, 330
 - leaf nodes, 330
 - pruning, 331, 344–347
 - root node, 330
 - rule extraction from, 357–359
- deep web, 597
- default rules, 357
- DENCLUE, 476–479
 - advantages, 479
 - clusters, 478
 - density attractor, 478
 - density estimation, 476
 - kernel density estimation, 477–478
 - kernels, 478
 - See also* cluster analysis; density-based methods
- dendrograms, 460
- densification power law, 592
- density estimation, 476
 - DENCLUE, 477–478
 - kernel function, 477–478
- density-based methods, 449, 471–479, 491
 - DBSCAN, 471–473
 - DENCLUE, 476–479
 - object division, 449
 - OPTICS, 473–476
 - STING as, 480
 - See also* cluster analysis
- density-based outlier detection, 564–567
 - local outlier factor, 566–567
 - local proximity, 564
 - local reachability density, 566
 - relative density, 565
- descendant cells, 189
- descriptive mining tasks, 15
- DIANA (Divisive Analysis), 459, 460
- dice operation, 148
- differential privacy, 622
- dimension tables, 136
- dimensional cells, 189
- dimensionality reduction, 86, 99–100, 120
- dimensionality reduction methods, 510, 519–522, 538
 - list of, 587
 - spectral clustering, 520–522
- dimension/level
 - application of, 297
 - constraints, 294
- dimensions, 10, 136
 - association rule, 281
 - cardinality of, 159
 - concept hierarchies and, 142–144
 - in multidimensional view, 33
 - ordering of, 210
 - pattern, 281
 - ranking, 225
 - relevance analysis, 175
 - selection, 225
 - shared, 204
 - See also* data warehouses
- direct discriminative pattern mining, 422
- directed acyclic graphs, 394–395
- discernibility matrix, 427
- discovery-driven exploration, 231–234, 235
- discrepancy detection, 91–93
- discrete attributes, 44
- discrete Fourier transform (DFT), 101, 587
- discrete wavelet transform (DWT), 100–102, 587
- discretization, 112, 120
 - by binning, 115
 - by clustering, 116
 - by correlation analysis, 117
 - by decision tree analysis, 116
 - by histogram analysis, 115–116
 - techniques, 113
- discriminant analysis, 600
- discriminant rules, 16
- discriminative frequent pattern-based classification, 416, 419–422, 437
 - basis for, 419
 - feature generation, 420
 - feature selection, 420–421
 - framework, 420–421
 - learning of classification model, 421

- dispersion of data, 44, 48–51
 - dissimilarity
 - asymmetric binary, 71
 - between attributes of mixed type, 76–77
 - between binary attributes, 71–72
 - measuring, 65–78, 79
 - between nominal attributes, 69
 - on numeric data, 72–74
 - between ordinal attributes, 75
 - symmetric binary, 70–71
 - dissimilarity matrix, 67, 68
 - data matrix versus, 67–68
 - n*-by-*n* table representation, 68
 - as one-mode matrix, 68
 - distance measures, 461–462
 - Euclidean, 72–73
 - Manhattan, 72–73
 - Minkowski, 73
 - supremum, 73–74
 - types of, 72
 - distance-based cluster analysis, 445
 - distance-based outlier detection, 561–562
 - nested loop algorithm, 561, 562
 - See also* outlier detection
 - distributed data mining, 615, 624
 - distributed privacy preservation, 622
 - distributions
 - boxplots for visualizing, 49–50
 - five-number summary, 49
 - distributive measures, 145
 - Divisive Analysis (DIANA), 459, 460
 - divisive hierarchical method, 459
 - agglomerative hierarchical clustering versus, 459–460
 - DIANA, 459, 460
 - DNA chips, 512
 - document classification, 430
 - documents
 - language model, 26
 - topic model, 26–27
 - drill-across operation, 148
 - drill-down operation, 11, 146–147
 - drill-through operation, 148
 - dynamic itemset counting, 256
- E**
- eager learners, 423, 437
 - Eclat (Equivalence Class Transformation) algorithm, 260, 272
 - e-commerce, 609
 - editing method, 425
 - efficiency
 - Apriori algorithm, 255–256
 - backpropagation, 404
 - data mining algorithms, 31
 - elbow method, 486
 - email spam filtering, 435
 - engineering applications, 613
 - ensemble methods, 378–379, 386
 - bagging, 379–380
 - boosting, 380–382
 - for class imbalance problem, 385
 - random forests, 382–383
 - types of, 378, 386
 - enterprise warehouses, 132
 - entity identification problem, 94
 - entity-relationship (ER) data model, 9, 139
 - epoch updating, 404
 - equal-frequency histograms, 107, 116
 - equal-width histograms, 107, 116
 - equivalence classes, 427
 - error rates, 367
 - error-correcting codes, 431–432
 - Euclidean distance, 72
 - mathematical properties, 72–73
 - weighted, 74
 - See also* distance measures
 - evaluation metrics, 364–370
 - evolution, of database system technology, 3–5
 - evolutionary searches, 579
 - exception-based, discovery-driven exploration, 231–234, 235
 - exceptions, 231
 - exhaustive rules, 358
 - expectation-maximization (EM) algorithm, 505–508, 538
 - expectation step (E-step), 505
 - fuzzy clustering with, 505–507
 - maximization step (M-step), 505
 - for mixture models, 507–508
 - for probabilistic model-based clustering, 507–508
 - steps, 505
 - See also* probabilistic model-based clustering
 - expected values, 97
 - cell, 234
 - exploratory data mining. *See* multidimensional data mining
 - extraction
 - data, 134
 - rule, from decision tree, 357–359
 - extraction/transformation/loading (ETL) tools, 93
 - extractors, 151

F

- fact constellation, 141
 - example, 141–142
 - illustrated, 142
- fact tables, 136
 - summary, 165
- factor analysis, 600
- facts, 136
- false negatives, 365
- false positives, 365
- farthest-neighbor clustering algorithm, 462
- field overloading, 92
- financial data analysis, 607–609
 - credit policy analysis, 608–609
 - crimes detection, 609
 - data warehouses, 608
 - loan payment prediction, 608–609
 - targeted marketing, 609
- FindCBLOF algorithm, 569–570
- five-number summary, 49
- fixed-width clustering, 570
- FOIL, 359, 363, 418
- Forest-RC, 383
- forward algorithm, 591
- FP-growth, 257–259, 272
 - algorithm illustration, 260
 - example, 257–258
 - performance, 259
- FP-trees, 257
 - condition pattern base, 258
 - construction, 257–258
 - main memory-based, 259
 - mining, 258, 259
- Frag-Shells, 212, 213
- fraudulent analysis, 610–611
- frequency patterns
 - approximate, 281, 307–312
 - compressed, 281, 307–312
 - constraint-based, 281
 - near-match, 281
 - redundancy-aware top-*k*, 281
 - top-*k*, 281
- frequent itemset mining, 18, 272, 282
 - Apriori algorithm, 248–253
 - closed patterns, 262–264
 - market basket analysis, 244–246
 - max patterns, 262–264
 - methods, 248–264
 - pattern-growth approach, 257–259
 - with vertical data format, 259–262, 272
- frequent itemsets, 243, 246, 272
 - association rule generation from, 253, 254
 - closed, 247, 248, 262–264, 308
 - finding, 247
 - finding by confined candidate generation, 248–253
 - maximal, 247, 248, 262–264, 308
 - subsets, 309
- frequent pattern mining, 279
 - advanced forms of patterns, 320
 - application domain-specific semantics, 282
 - applications, 317–319, 321
 - approximate patterns, 307–312
 - classification criteria, 280–283
 - colossal patterns, 301–307
 - compressed patterns, 307–312
 - constraint-based, 294–301, 320
 - data analysis usages, 282
 - for data cleaning, 318
 - direct discriminative, 422
 - high-dimensional data, 301–307
 - in high-dimensional space, 320
 - in image data analysis, 319
 - for indexing structures, 319
 - kinds of data and features, 282
 - multidimensional associations, 287–289
 - in multilevel, multidimensional space, 283–294
 - multilevel associations, 283–294
 - in multimedia data analysis, 319
 - negative patterns, 291–294
 - for noise filtering, 318
 - Pattern-Fusion, 302–307
 - quantitative association rules, 289–291
 - rare patterns, 291–294
 - in recommender systems, 319
 - road map, 279–283
 - scalable computation and, 319
 - scope of, 319–320
 - in sequence or structural data analysis, 319
 - in spatiotemporal data analysis, 319
 - for structure and cluster discovery, 318
 - for subspace clustering, 318–319
 - in time-series data analysis, 319
 - top-*k*, 310
 - in video data analysis, 319
 - See also* frequent patterns
- frequent pattern-based classification, 415–422, 437
 - associative, 415, 416–419
 - discriminative, 416, 419–422
 - framework, 422
- frequent patterns, 17, 243
 - abstraction levels, 281
 - association rule mapping, 280
 - basic, 280

- frequent patterns (*Continued*)
 - closed, 262–264, 280
 - concepts, 243–244
 - constraint-based, 281
 - dimensions, 281
 - diversity, 280
 - exploration, 313–319
 - growth, 257–259, 272
 - max, 262–264, 280
 - mining, 243–244, 279–325
 - mining constraints or criteria, 281
 - number of dimensions involved in, 281
 - semantic annotation of, 313–317
 - sequential, 243
 - strong associations, 437
 - structured, 243
 - trees, 257–259
 - types of values in, 281
- frequent subgraphs, 591
- front-end client layer, 132
- full materialization, 159, 179, 234
- fuzzy clustering, 499–501, 538
 - data set for, 506
 - with EM algorithm, 505–507
 - example, 500
 - expectation step (E-step), 505
 - flexibility, 501
 - maximization step (M-step), 506–507
 - partition matrix, 499
 - as soft clusters, 501
- fuzzy logic, 428
- fuzzy sets, 428–429, 437, 499
 - evaluation, 500–501
 - example, 499

G

- gain ratio, 340
 - C4.5 use of, 340
 - formula, 341
 - maximum, 341
- gateways, 131
- gene expression, 513–514
- generalization
 - attribute, 169–170
 - attribute, control, 170
 - attribute, threshold control, 170
 - in multimedia data mining, 596
 - process, 172
 - results presentation, 174
 - synchronous, 175
- generalized linear models, 599–600
- generalized relations
 - attribute-oriented induction, 172
 - presentation of, 174
 - threshold control, 170
- generative model, 467–469
- genetic algorithms, 426–427, 437
- genomes, 15
- geodesic distance, 525–526, 539
 - diameter, 525
 - eccentricity, 525
 - measurements based on, 526
 - peripheral vertex, 525
 - radius, 525
- geographic data warehouses, 595
- geometric projection visualization, 58–60
- Gini index, 341
 - binary enforcement, 332
 - binary indexes, 341
 - CART use of, 341
 - decision tree induction using,
 - 342–343
 - minimum, 342
 - partitioning and, 342
- global constants, for missing values, 88
- global outliers, 545, 581
 - detection, 545
 - example, 545
- Google
 - Flu Trends*, 2
 - popularity of, 619–620
- gradient descent strategy, 396–397
 - algorithms, 397
 - greedy hill-climbing, 397
 - as iterative, 396–397
- graph and network data clustering, 497,
 - 522–532, 539
 - applications, 523–525
 - bipartite graph, 523
 - challenges, 523–525, 530
 - cuts and clusters, 529–530
 - generic method, 530–531
 - geodesic distance, 525–526
 - methods, 528–532
 - similarity measures, 525–528
 - SimRank, 526–528
 - social network, 524–525
 - web search engines, 523–524
 - See also* cluster analysis
- graph cuts, 539
- graph data, 14
- graph index structures, 591
- graph pattern mining, 591–592, 612–613
- graphic displays
 - data presentation software, 44–45
 - histogram, 54, 55

- quantile plot, 51–52
- quantile-quantile plot, 52–54
- scatter plot, 54–56
- greedy hill-climbing, 397
- greedy methods, attribute subset selection, 104–105
- grid-based methods, 450, 479–483, 491
 - CLIQUE, 481–483
 - STING, 479–481
 - See also* cluster analysis
- grid-based outlier detection, 562–564
 - CELL method, 562, 563
 - cell properties, 562
 - cell pruning rules, 563
 - See also* outlier detection
- group-based support, 286
- group-by** clause, 231
- grouping attributes, 231
- grouping variables, 231
- Grubb's test, 555

H

- hamming distance, 431
- hard constraints, 534, 539
 - example, 534
 - handling, 535–536
- harmonic mean, 369
- hash-based technique, 255
- heterogeneous networks, 592
 - classification of, 593
 - clustering of, 593
 - ranking of, 593
- heterogeneous transfer learning, 436
- hidden Markov model (HMM), 590, 591
- hierarchical methods, 449, 457–470, 491
 - agglomerative, 459–461
 - algorithmic, 459, 461–462
 - Bayesian, 459
 - BIRCH, 458, 462–466
 - Chameleon, 458, 466–467
 - complete linkages, 462, 463
 - distance measures, 461–462
 - divisive, 459–461
 - drawbacks, 449
 - merge or split points and, 458
 - probabilistic, 459, 467–470
 - single linkages, 462, 463
 - See also* cluster analysis
- hierarchical visualization, 63
 - treemaps, 63, 65
 - Worlds-with-Worlds, 63, 64
- high-dimensional data, 301
 - clustering, 447

- data distribution of, 560
- frequent pattern mining, 301–307
- outlier detection in, 576–580, 582
- row enumeration, 302
- high-dimensional data clustering, 497, 508–522, 538, 553
 - biclustering, 512–519
 - dimensionality reduction methods, 510, 519–522
 - example, 508–509
 - problems, challenges, and methodologies, 508–510
 - subspace clustering methods, 509, 510–511
 - See also* cluster analysis
- HilOut algorithm, 577–578
- histograms, 54, 106–108, 116
 - analysis by discretization, 115–116
 - attributes, 106
 - binning, 106
 - construction, 559
 - equal-frequency, 107
 - equal-width, 107
 - example, 54
 - illustrated, 55, 107
 - multidimensional, 108
 - as nonparametric model, 559
 - outlier detection using, 558–560
- holdout method, 370, 386
- holistic measures, 145
- homogeneous networks, 592
 - classification of, 593
 - clustering of, 593
- Hopkins statistic, 484–485
- horizontal data format, 259
- hybrid OLAP (HOLAP), 164–165, 179
- hybrid-dimensional association rules, 288

I

- IBM Intelligent Miner, 603, 606
- iceberg condition, 191
- iceberg cubes, 160, 179, 190, 235
 - BUC construction, 201
 - computation, 160, 193–194, 319
 - computation and storage, 210–211
 - computation with Star-Cubing algorithm, 204–210
 - materialization, 319
 - specification of, 190–191
 - See also* data cubes
- icon-based visualization, 60
 - Chernoff faces, 60–61

- icon-based visualization (*Continued*)
 - stick figure technique, 61–63
 - See also* data visualization
- ID3, 332, 385
 - greedy approach, 332
 - information gain, 336
 - See also* decision tree induction
- IF-THEN rules, 355–357
 - accuracy, 356
 - conflict resolution strategy, 356
 - coverage, 356
 - default rule, 357
 - extracting from decision tree, 357
 - form, 355
 - rule antecedent, 355
 - rule consequent, 355
 - rule ordering, 357
 - satisfied, 356
 - triggered, 356
- illustrated, 149
- image data analysis, 319
- imbalance problem, 367
- imbalance ratio (IR), 270
 - skewness, 271
- inconvertible constraints, 300
- incremental data mining, 31
- indexes
 - bitmapped join, 163
 - composite join, 162
 - Gini, 332, 341–343
 - inverted, 212, 213
- indexing
 - bitmap, 160–161, 179
 - bitmapped join, 179
 - frequent pattern mining for, 319
 - join, 161–163, 179
 - OLAP, 160–163
- inductive databases, 601
- inferential statistics, 24
- information age, moving toward, 1–2
- information extraction systems, 430
- information gain, 336–340
 - decision tree induction using, 338–339
 - ID3 use of, 336
 - pattern frequency support versus, 421
 - single feature plot, 420
 - split-point, 340
- information networks
 - analysis, 592–593
 - evolution of, 594
 - link prediction in, 593–594
 - mining, 623
 - OLAP in, 594
 - role discovery in, 593–594
 - similarity search in, 594
- information processing, 153
- information retrieval (IR), 26–27
 - challenges, 27
 - language model, 26
 - topic model, 26–27
- informativeness model, 535
- initial working relations, 168, 169, 177
- instance-based learners. *See* lazy learners
- instances, constraints on, 533, 539
- integrated data warehouses, 126
- integrators, 127
- intelligent query answering, 618
- interactive data mining, 604, 607
- interactive mining, 30
- intercuboid query expansion, 221
 - example, 224–225
 - method, 223–224
- interdimensional association rules, 288
- interestingness, 21–23
 - assessment methods, 23
 - components of, 21
 - expected, 22
 - objective measures, 21–22
 - strong association rules, 264–265
 - subjective measures, 22
 - threshold, 21–22
 - unexpected, 22
- interestingness constraints, 294
 - application of, 297
- interpretability
 - backpropagation and, 406–408
 - classification, 369
 - cluster analysis, 447
 - data, 85
 - data quality and, 85
 - probabilistic hierarchical clustering, 469
- interquartile range (IQR), 49, 555
- interval-scaled attributes, 43, 79
- intracuboid query expansion, 221
 - example, 223
 - method, 221–223
 - value usage, 222
- intradimensional association rules, 287
- intrusion detection, 569–570
 - anomaly-based, 614
 - data mining algorithms, 614–615
 - discriminative classifiers, 615
 - distributed data mining, 615

- signature-based, 614
- stream data analysis, 615
- visualization and query tools, 615
- inverted indexes, 212, 213
- invisible data mining, 33, 618–620, 625
- IQR. *See* Interquartile range
- IR. *See* information retrieval
- item merging, 263
- item skipping, 263
- items, 13
- itemsets, 246
 - candidate, 251, 252
 - dependent, 266
 - dynamic counting, 256
 - imbalance ratio (IR), 270, 271
 - negatively correlated, 292
 - occurrence independence, 266
 - strongly negatively correlated, 292
 - See also* frequent itemsets
- iterative Pattern-Fusion, 306
- iterative relocation techniques, 448

J

- Jaccard coefficient, 71
- join indexing, 161–163, 179

K

- k*-anonymity method, 621–622
- Karush-Kuhn-Tucker (KKT) conditions, 412
- k*-distance neighborhoods, 565
- kernel density estimation, 477–478
- kernel function, 415
- k*-fold cross-validation, 370–371
- k*-means, 451–454
 - algorithm, 452
 - application of, 454
 - CLARANS, 457
 - within-cluster variation, 451, 452
 - clustering by, 453
 - drawback of, 454–455
 - functioning of, 452
 - scalability, 454
 - time complexity, 453
 - variants, 453–454
- k*-means clustering, 536
- k*-medoids, 454–457
 - absolute-error criterion, 455
 - cost function for, 456
 - PAM, 455–457
- k*-nearest-neighbor classification, 423
 - closeness, 423
 - distance-based comparisons, 425

- editing method, 425
- missing values and, 424
- number of neighbors, 424–425
- partial distance method, 425
- speed, 425
- knowledge
 - background, 30–31
 - mining, 29
 - presentation, 8
 - representation, 33
 - transfer, 434
- knowledge bases, 5, 8
- knowledge discovery
 - data mining in, 7
 - process, 8
- knowledge discovery from data (KDD), 6
- knowledge extraction. *See* data mining
- knowledge mining. *See* data mining
- knowledge type constraints, 294
- k*-predicate sets, 289
- Kulczynski measure, 268, 272
 - negatively correlated pattern based on, 293–294

L

- language model, 26
- Laplacian correction, 355
- lattice of cuboids, 139, 156, 179, 188–189, 234
- lazy learners, 393, 422–426, 437
 - case-based reasoning classifiers, 425–426
 - k*-nearest-neighbor classifiers, 423–425
- l*-diversity method, 622
- learning
 - active, 430, 433–434, 437
 - backpropagation, 400
 - as classification step, 328
 - connectionist, 398
 - by examples, 445
 - by observation, 445
 - rate, 397
 - semi-supervised, 572
 - supervised, 330
 - transfer, 430, 434–436, 438
 - unsupervised, 330, 445, 490
- learning rates, 403–404
- leave-one-out, 371
- lift, 266, 272
 - correlation analysis with, 266–267
- likelihood ratio statistic, 363
- linear regression, 90, 105
 - multiple, 106
- linearly, 412–413
- linearly inseparable data, 413–415

- link mining, 594
- link prediction, 594
- load, in back-end tools/utilities, 134
- loan payment prediction, 608–609
- local outlier factor, 566–567
- local proximity-based outliers, 564–565
- logistic function, 402
- log-linear models, 106
- lossless compression, 100
- lossy compression, 100
- lower approximation, 427

M

- machine learning, 24–26
 - active, 25
 - data mining similarities, 26
 - semi-supervised, 25
 - supervised, 24
 - unsupervised, 25
- Mahalanobis distance, 556
- majority voting, 335
- Manhattan distance, 72–73
- MaPle, 519
- margin, 410
- market basket analysis, 244–246, 271–272
 - example, 244
 - illustrated, 244
- Markov chains, 591
- materialization
 - full, 159, 179, 234
 - iceberg cubes, 319
 - no, 159
 - partial, 159–160, 192, 234
 - semi-offline, 226
- max patterns, 280
- max_confidence** measure, 268, 272
- maximal frequent itemsets, 247, 308
 - example, 248
 - mining, 262–264
 - shortcomings for compression, 308–309
- maximum marginal hyperplane (MMH), 409
 - SVM finding, 412
- maximum normed residual test, 555
- mean, 39, 45
 - bin, smoothing by, 89
 - example, 45
 - for missing values, 88
 - trimmed, 46
 - weighted arithmetic, 45
- measures, 145
 - accuracy-based, 369
 - algebraic, 145
 - all_confidence**, 272
 - antimonotonic, 194
 - attribute selection, 331
 - categories of, 145
 - of central tendency, 39, 44, 45–47
 - correlation, 266
 - data cube, 145
 - dispersion, 48–51
 - distance, 72–74, 461–462
 - distributive, 145
 - holistic, 145
 - Kulczynski, 272
 - max_confidence**, 272
 - of multidimensional databases, 146
 - null-invariant, 272
 - pattern evaluation, 267–271
 - precision, 368–369
 - proximity, 67, 68–72
 - recall, 368–369
 - sensitivity, 367
 - significance, 312
 - similarity/dissimilarity, 65–78
 - specificity, 367
- median, 39, 46
 - bin, smoothing by, 89
 - example, 46
 - formula, 46–47
 - for missing values, 88
- metadata, 92, 134, 178
 - business, 135
 - importance, 135
 - operational, 135
 - repositories, 134–135
- metarule-guided mining
 - of association rules, 295–296
 - example, 295–296
- metrics, 73
 - classification evaluation, 364–370
- microeconomic view, 601
- midrange, 47
- MineSet, 603, 605
- minimal interval size, 116
- minimal spanning tree algorithm, 462
- minimum confidence threshold, 18, 245
- Minimum Description Length (MDL), 343–344
- minimum support threshold, 18, 190
 - association rules, 245
 - count, 246
- Minkowski distance, 73
- min-max normalization, 114
- missing values, 88–89
- mixed-effect models, 600

- mixture models, 503, 538
 - EM algorithm for, 507–508
 - univariate Gaussian, 504
 - mode, 39, 47
 - example, 47
 - model selection, 364
 - with statistical tests of significance, 372–373
 - models, 18
 - modularity
 - of clustering, 530
 - use of, 539
 - MOLAP. *See* multidimensional OLAP
 - monotonic constraints, 298
 - motifs, 587
 - moving-object data mining, 595–596, 623–624
 - multiclass classification, 430–432, 437
 - all-versus-all (AVA), 430–431
 - error-correcting codes, 431–432
 - one-versus-all (OVA), 430
 - multidimensional association rules, 17, 283, 288, 320
 - hybrid-dimensional, 288
 - interdimensional, 288
 - mining, 287–289
 - mining with static discretization of quantitative attributes, 288
 - with no repeated predicates, 288
 - See also* association rules
 - multidimensional data analysis
 - in cube space, 227–234
 - in multimedia data mining, 596
 - spatial, 595
 - of top-*k* results, 226
 - multidimensional data mining, 11–13, 34 155–156, 179, 187, 227, 235
 - data cube promotion of, 26
 - dimensions, 33
 - example, 228–229
 - retail industry, 610
 - multidimensional data model, 135–146, 178
 - data cube as, 136–139
 - dimension table, 136
 - dimensions, 142–144
 - fact constellation, 141–142
 - fact table, 136
 - snowflake schema, 140–141
 - star schema, 139–140
 - multidimensional databases
 - measures of, 146
 - querying with starnet model, 149–150
 - multidimensional histograms, 108
 - multidimensional OLAP (MOLAP), 132, 164, 179
 - multifeature cubes, 227, 230, 235
 - complex query support, 231
 - examples, 230–231
 - multilayer feed-forward neural networks, 398–399
 - example, 405
 - illustrated, 399
 - layers, 399
 - units, 399
 - multilevel association rules, 281, 283, 284, 320
 - ancestors, 287
 - concept hierarchies, 285
 - dimensions, 281
 - group-based support, 286
 - mining, 283–287
 - reduced support, 285, 286
 - redundancy, checking, 287
 - uniform support, 285–286
 - multimedia data, 14
 - multimedia data analysis, 319
 - multimedia data mining, 596
 - multimodal, 47
 - multiple linear regression, 90, 106
 - multiple sequence alignment, 590
 - multiple-phase clustering, 458–459
 - multitier data warehouses, 134
 - multivariate outlier detection, 556
 - with Mahalanobis distance, 556
 - with multiple clusters, 557
 - with multiple parametric distributions, 557
 - with χ^2 -static, 556
 - multiway array aggregation, 195, 235
 - for full cube computation, 195–199
 - minimum memory requirements, 198
 - must-link constraints, 533, 536
 - mutation operator, 426
 - mutual information, 315–316
 - mutually exclusive rules, 358
- ## N
- naive Bayesian classification, 385
 - class label prediction with, 353–355
 - functioning of, 351–352
 - nearest-neighbor clustering algorithm, 461
 - near-match patterns/rules, 281
 - negative correlation, 55, 56
 - negative patterns, 280, 283, 320
 - example, 291–292
 - mining, 291–294
 - negative transfer, 436
 - negative tuples, 364
 - negatively skewed data, 47

- neighborhoods
 - density, 471
 - distance-based outlier detection, 560
 - k*-distance, 565
 - nested loop algorithm, 561, 562
 - networked data, 14
 - networks, 592
 - heterogeneous, 592, 593
 - homogeneous, 592, 593
 - information, 592–594
 - mining in science applications, 612–613
 - social, 592
 - statistical modeling of, 592–594
 - neural networks, 19, 398
 - backpropagation, 398–408
 - as black boxes, 406
 - for classification, 19, 398
 - disadvantages, 406
 - fully connected, 399, 406–407
 - learning, 398
 - multilayer feed-forward, 398–399
 - pruning, 406–407
 - rule extraction algorithms, 406, 407
 - sensitivity analysis, 408
 - three-layer, 399
 - topology definition, 400
 - two-layer, 399
 - neurodes, 399
 - Ng-Jordan-Weiss algorithm, 521, 522
 - no materialization, 159
 - noise filtering, 318
 - noisy data, 89–91
 - nominal attributes, 41
 - concept hierarchies for, 284
 - correlation analysis, 95–96
 - dissimilarity between, 69
 - example, 41
 - proximity measures, 68–70
 - similarity computation, 70
 - values of, 79, 288
 - See also* attributes
 - nonlinear SVMs, 413–415
 - nonparametric statistical methods, 553–558
 - nonvolatile data warehouses, 127
 - normalization, 112, 120
 - data transformation by, 113–115
 - by decimal scaling, 115
 - min-max, 114
 - z*-score, 114–115
 - null rules, 92
 - null-invariant measures, 270–271, 272
 - null-transactions, 270, 272
 - number of, 270
 - problem, 292–293
 - numeric attributes, 43–44, 79
 - covariance analysis, 98
 - interval-scaled, 43, 79
 - ratio-scaled, 43–44, 79
 - numeric data, dissimilarity on, 72–74
 - numeric prediction, 328, 385
 - classification, 328
 - support vector machines (SVMs) for, 408
 - numerosity reduction, 86, 100, 120
 - techniques, 100
- O**
- object matching, 94
 - objective interestingness measures, 21–22
 - one-class model, 571–572
 - one-pass cube computation, 198
 - one-versus-all (OVA), 430
 - online analytical mining (OLAM), 155, 227
 - online analytical processing (OLAP), 4, 33, 128, 179
 - access patterns, 129
 - data contents, 128
 - database design, 129
 - dice operation, 148
 - drill-across operation, 148
 - drill-down operation, 11, 135–136, 146
 - drill-through operation, 148
 - example operations, 147
 - functionalities of, 154
 - hybrid OLAP, 164–165, 179
 - indexing, 125, 160–163
 - in information networks, 594
 - in knowledge discovery process, 125
 - market orientation, 128
 - multidimensional (MOLAP), 132, 164, 179
 - OLTP versus, 128–129, 130
 - operation integration, 125
 - operations, 146–148
 - pivot (rotate) operation, 148
 - queries, 129, 130, 163–164
 - query processing, 125, 163–164
 - relational OLAP, 132, 164, 165, 179
 - roll-up operation, 11, 135–136, 146
 - sample data effectiveness, 219
 - server architectures, 164–165
 - servers, 132
 - slice operation, 148
 - spatial, 595
 - statistical databases versus, 148–149

- user-control versus automation, 167
 - view, 129
 - online transaction processing (OLTP), 128
 - access patterns, 129
 - customer orientation, 128
 - data contents, 128
 - database design, 129
 - OLAP versus, 128–129, 130
 - view, 129
 - operational metadata, 135
 - OPTICS, 473–476
 - cluster ordering, 474–475, 477
 - core-distance, 475
 - density estimation, 477
 - reachability-distance, 475
 - structure, 476
 - terminology, 476
 - See also* cluster analysis; density-based methods
 - ordered attributes, 103
 - ordering
 - class-based, 358
 - dimensions, 210
 - rule, 357
 - ordinal attributes, 42, 79
 - dissimilarity between, 75
 - example, 42
 - proximity measures, 74–75
 - outlier analysis, 20–21
 - clustering-based techniques, 66
 - example, 21
 - in noisy data, 90
 - spatial, 595
 - outlier detection, 543–584
 - angle-based (ABOD), 580
 - application-specific, 548–549
 - categories of, 581
 - CELL method, 562–563
 - challenges, 548–549
 - clustering analysis and, 543
 - clustering for, 445
 - clustering-based methods, 552–553, 560–567
 - collective, 548, 575–576
 - contextual, 546–547, 573–575
 - distance-based, 561–562
 - extending, 577–578
 - global, 545
 - handling noise in, 549
 - in high-dimensional data, 576–580, 582
 - with histograms, 558–560
 - intrusion detection, 569–570
 - methods, 549–553
 - mixture of parametric distributions, 556–558
 - multivariate, 556
 - novelty detection relationship, 545
 - proximity-based methods, 552, 560–567, 581
 - semi-supervised methods, 551
 - statistical methods, 552, 553–560, 581
 - supervised methods, 549–550
 - understandability, 549
 - univariate, 554
 - unsupervised methods, 550
 - outlier subgraphs, 576
 - outliers
 - angle-based, 20, 543, 544, 580
 - collective, 547–548, 581
 - contextual, 545–547, 573, 581
 - density-based, 564
 - distance-based, 561
 - example, 544
 - global, 545, 581
 - high-dimensional, modeling, 579–580
 - identifying, 49
 - interpretation of, 577
 - local proximity-based, 564–565
 - modeling, 548
 - in small clusters, 571
 - types of, 545–548, 581
 - visualization with boxplot, 555
 - oversampling, 384, 386
 - example, 384–385
- ## P
- pairwise alignment, 590
 - pairwise comparison, 372
 - PAM. *See* Partitioning Around Medoids algorithm
 - parallel and distributed data-intensive mining
 - algorithms, 31
 - parallel coordinates, 59, 62
 - parametric data reduction, 105–106
 - parametric statistical methods, 553–558
 - Pareto distribution, 592
 - partial distance method, 425
 - partial materialization, 159–160, 179, 234
 - strategies, 192
 - partition matrix, 538
 - partitioning
 - algorithms, 451–457
 - in Apriori efficiency, 255–256
 - bootstrapping, 371, 386
 - criteria, 447
 - cross-validation, 370–371, 386
 - Gini index and, 342
 - holdout method, 370, 386
 - random sampling, 370, 386

- partitioning (*Continued*)
 - recursive, 335
 - tuples, 334
- Partitioning Around Medoids (PAM) algorithm, 455–457
- partitioning methods, 448, 451–457, 491
 - centroid-based, 451–454
 - global optimality, 449
 - iterative relocation techniques, 448
 - k*-means, 451–454
 - k*-medoids, 454–457
 - k*-modes, 454
 - object-based, 454–457
 - See also* cluster analysis
- path-based similarity, 594
- pattern analysis, in recommender systems, 282
- pattern clustering, 308–310
- pattern constraints, 297–300
- pattern discovery, 601
- pattern evaluation, 8
- pattern evaluation measures, 267–271
 - all_confidence**, 268
 - comparison, 269–270
 - cosine, 268
 - Kulczynski, 268
 - max_confidence**, 268
 - null-invariant, 270–271
 - See also* measures
- pattern space pruning, 295
- pattern-based classification, 282, 318
- pattern-based clustering, 282, 516
- Pattern-Fusion, 302–307
 - characteristics, 304
 - core pattern, 304–305
 - initial pool, 306
 - iterative, 306
 - merging subpatterns, 306
 - shortcuts identification, 304
 - See also* colossal patterns
- pattern-guided mining, 30
- patterns
 - actionable, 22
 - co-location, 319
 - colossal, 301–307, 320
 - combined significance, 312
 - constraint-based generation, 296–301
 - context modeling of, 314–315
 - core, 304–305
 - distance, 309
 - evaluation methods, 264–271
 - expected, 22
 - expressed, 309
 - frequent, 17
 - hidden meaning of, 314
 - interesting, 21–23, 33
 - metric space, 306–307
 - negative, 280, 291–294, 320
 - negatively correlated, 292, 293
 - rare, 280, 291–294, 320
 - redundancy between, 312
 - relative significance, 312
 - representative, 309
 - search space, 303
 - strongly negatively correlated, 292
 - structural, 282
 - type specification, 15–23
 - unexpected, 22
 - See also* frequent patterns
- pattern-trees, 264
- Pearson's correlation coefficient, 222
- percentiles, 48
- perception-based classification (PBC), 348
 - illustrated, 349
 - as interactive visual approach, 607
 - pixel-oriented approach, 348–349
 - split screen, 349
 - tree comparison, 350
- phylogenetic trees, 590
- pivot (rotate) operation, 148
- pixel-oriented visualization, 153
- planning and analysis tools, 153
- point queries, 216, 217, 220
- pool-based approach, 433
- positive correlation, 55, 56
- positive tuples, 364
- positively skewed data, 47
- possibility theory, 428
- posterior probability, 351
- postpruning, 344–345, 346
- power law distribution, 592
- precision measure, 368–369
- predicate sets
 - frequent, 288–289
 - k*, 289
- predicates
 - repeated, 288
 - variables, 295
- prediction, 19
 - classification, 328
 - link, 593–594
 - loan payment, 608–609
 - with naive Bayesian classification, 353–355
 - numeric, 328, 385

- prediction cubes, 227–230, 235
 - example, 228–229
 - Probability-Based Ensemble, 229–230
 - predictive analysis, 18–19
 - predictive mining tasks, 15
 - predictive statistics, 24
 - predictors, 328
 - prepruning, 344, 346
 - prime relations
 - contrasting classes, 175, 177
 - deriving, 174
 - target classes, 175, 177
 - principle components analysis (PCA), 100, 102–103
 - application of, 103
 - correlation-based clustering with, 511
 - illustrated, 103
 - in lower-dimensional space extraction, 578
 - procedure, 102–103
 - prior probability, 351
 - privacy-preserving data mining, 33, 621, 626
 - distributed, 622
 - k*-anonymity method, 621–622
 - l*-diversity method, 622
 - as mining trend, 624–625
 - randomization methods, 621
 - results effectiveness, downgrading, 622
 - probabilistic clusters, 502–503
 - probabilistic hierarchical clustering, 467–470
 - agglomerative clustering framework, 467, 469
 - algorithm, 470
 - drawbacks of using, 469–470
 - generative model, 467–469
 - interpretability, 469
 - understanding, 469
 - See also* hierarchical methods
 - probabilistic model-based clustering, 497–508, 538
 - expectation-maximization algorithm, 505–508
 - fuzzy clusters and, 499–501
 - product reviews example, 498
 - user search intent example, 498
 - See also* cluster analysis
 - probability
 - estimation techniques, 355
 - posterior, 351
 - prior, 351
 - probability and statistical theory, 601
 - Probability-Based Ensemble (PBE), 229–230
 - PROCLUS, 511
 - profiles, 614
 - proximity measures, 67
 - for binary attributes, 70–72
 - for nominal attributes, 68–70
 - for ordinal attributes, 74–75
 - proximity-based methods, 552, 560–567, 581
 - density-based, 564–567
 - distance-based, 561–562
 - effectiveness, 552
 - example, 552
 - grid-based, 562–564
 - types of, 552, 560
 - See also* outlier detection
 - pruning
 - cost complexity algorithm, 345
 - data space, 300–301
 - decision trees, 331, 344–347
 - in *k*-nearest neighbor classification, 425
 - network, 406–407
 - pattern space, 295, 297–300
 - pessimistic, 345
 - postpruning, 344–345, 346
 - prepruning, 344, 346
 - rule, 363
 - search space, 263, 301
 - sets, 345
 - shared dimensions, 205
 - sub-itemset, 263
 - pyramid algorithm, 101
- ## Q
- quality control, 600
 - quantile plots, 51–52
 - quantile-quantile plots, 52
 - example, 53–54
 - illustrated, 53
 - See also* graphic displays
 - quantitative association rules, 281, 283, 288, 320
 - clustering-based mining, 290–291
 - data cube-based mining, 289–290
 - exceptional behavior disclosure, 291
 - mining, 289
 - quartiles, 48
 - first, 49
 - third, 49
 - queries, 10
 - intercuboid expansion, 223–225
 - intracuboid expansion, 221–223
 - language, 10
 - OLAP, 129, 130
 - point, 216, 217, 220
 - processing, 163–164, 218–227
 - range, 220
 - relational operations, 10

queries (*Continued*)

- subcube, 216, 217–218
- top-*k*, 225–227

query languages, 31

query models, 149–150

query-driven approach, 128

querying function, 433

R

rag bag criterion, 488

RainForest, 385

random forests, 382–383

random sampling, 370, 386

random subsampling, 370

random walk, 526

- similarity based on, 527

randomization methods, 621

range, 48

- interquartile, 49

range queries, 220

ranking

- cubes, 225–227, 235
- dimensions, 225
- function, 225
- heterogeneous networks, 593

rare patterns, 280, 283, 320

- example, 291–292
- mining, 291–294

ratio-scaled attributes, 43–44, 79

reachability density, 566

reachability distance, 565

recall measure, 368–369

recognition rate, 366–367

recommender systems, 282, 615

- advantages, 616
- biclustering for, 514–515
- challenges, 617
- collaborative, 610, 615, 616, 617, 618
- content-based approach, 615, 616
- data mining and, 615–618
- error types, 617–618
- frequent pattern mining for, 319
- hybrid approaches, 618
- intelligent query answering, 618
- memory-based methods, 617
- use scenarios, 616

recursive partitioning, 335

reduced support, 285, 286

redundancy

- in data integration, 94
- detection by correlations analysis, 94–98

redundancy-aware top-*k* patterns, 281, 311, 320

- extracting, 310–312
- finding, 312
- strategy comparison, 311–312
- trade-offs, 312

refresh, in back-end tools/utilities, 134

regression, 19, 90

- coefficients, 105–106
- example, 19
- linear, 90, 105–106
- in statistical data mining, 599

regression analysis, 19, 328

- in time-series data, 587–588

relational databases, 9

- components of, 9
- mining, 10
- relational schema for, 10

relational OLAP (ROLAP), 132, 164, 165, 179

relative significance, 312

relevance analysis, 19

repetition, 346

replication, 347

- illustrated, 346

representative patterns, 309

retail industry, 609–611

RIPPER, 359, 363

robustness, classification, 369

ROC curves, 374, 386

- classification models, 377
- classifier comparison with, 373–377
- illustrated, 376, 377
- plotting, 375

roll-up operation, 11, 146

rough set approach, 428–429, 437

row enumeration, 302

rule ordering, 357

rule pruning, 363

rule quality measures, 361–363

rule-based classification, 355–363, 386

- IF-THEN rules, 355–357
- rule extraction, 357–359
- rule induction, 359–363
- rule pruning, 363
- rule quality measures, 361–363

rules for constraints, 294

S

sales campaign analysis, 610

samples, 218

- cluster, 108–109
- data, 219

- simple random, 108
- stratified, 109–110
- sampling
 - in Apriori efficiency, 256
 - as data redundancy technique, 108–110
 - methods, 108–110
 - oversampling, 384–385
 - random, 386
 - with replacement, 380–381
 - uncertainty, 433
 - undersampling, 384–385
- sampling cubes, 218–220, 235
 - confidence interval, 219–220
 - framework, 219–220
 - query expansion with, 221
- SAS Enterprise Miner, 603, 604
- scalability
 - classification, 369
 - cluster analysis, 446
 - cluster methods, 445
 - data mining algorithms, 31
 - decision tree induction and, 347–348
 - dimensionality and, 577
 - k*-means, 454
- scalable computation, 319
- SCAN. *See* Structural Clustering Algorithm for Networks
 - core vertex, 531
 - illustrated, 532
- scatter plots, 54
 - 2-D data set visualization with, 59
 - 3-D data set visualization with, 60
 - correlations between attributes, 54–56
 - illustrated, 55
 - matrix, 56, 59
- schemas
 - integration, 94
 - snowflake, 140–141
 - star, 139–140
- science applications, 611–613
- search engines, 28
- search space pruning, 263, 301
- second guess heuristic, 369
- selection dimensions, 225
- self-training, 432
- semantic annotations
 - applications, 317, 313, 320–321
 - with context modeling, 316
 - from DBLP data set, 316–317
 - effectiveness, 317
 - example, 314–315
 - of frequent patterns, 313–317
 - mutual information, 315–316
 - task definition, 315
- Semantic Web, 597
- semi-offline materialization, 226
- semi-supervised classification, 432–433, 437
 - alternative approaches, 433
 - cotraining, 432–433
 - self-training, 432
- semi-supervised learning, 25
 - outlier detection by, 572
- semi-supervised outlier detection, 551
- sensitivity analysis, 408
- sensitivity measure, 367
- sentiment classification, 434
- sequence data analysis, 319
- sequences, 586
 - alignment, 590
 - biological, 586, 590–591
 - classification of, 589–590
 - similarity searches, 587
 - symbolic, 586, 588–590
 - time-series, 586, 587–588
- sequential covering algorithm, 359
 - general-to-specific search, 360
 - greedy search, 361
 - illustrated, 359
 - rule induction with, 359–361
- sequential pattern mining, 589
 - constraint-based, 589
 - in symbolic sequences, 588–589
- shapelets method, 590
- shared dimensions, 204
 - pruning, 205
- shared-sorts, 193
- shared-partitions, 193
- shell cubes, 160
- shell fragments, 192, 235
 - approach, 211–212
 - computation algorithm, 212, 213
 - computation example, 214–215
 - precomputing, 210
- shrinking diameter, 592
- sigmoid function, 402
- signature-based detection, 614
- significance levels, 373
- significance measure, 312
- significance tests, 372–373, 386
- silhouette coefficient, 489–490
- similarity
 - asymmetric binary, 71
 - cosine, 77–78

- similarity (*Continued*)
 - measuring, 65–78, 79
 - nominal attributes, 70
- similarity measures, 447–448, 525–528
 - constraints on, 533
 - geodesic distance, 525–526
 - SimRank, 526–528
- similarity searches, 587
 - in information networks, 594
 - in multimedia data mining, 596
- simple random sample with replacement (SRSWR), 108
- simple random sample without replacement (SRSWOR), 108
- SimRank, 526–528, 539
 - computation, 527–528
 - random walk, 526–528
 - structural context, 528
- simultaneous aggregation, 195
- single-dimensional association rules, 17, 287
- single-linkage algorithm, 460, 461
- singular value decomposition (SVD), 587
- skewed data
 - balanced, 271
 - negatively, 47
 - positively, 47
 - wavelet transforms on, 102
- slice operation, 148
- small-world phenomenon, 592
- smoothing, 112
 - by bin boundaries, 89
 - by bin means, 89
 - by bin medians, 89
 - for data discretization, 90
- snowflake schema, 140
 - example, 141
 - illustrated, 141
 - star schema versus, 140
- social networks, 524–525, 526–528
 - densification power law, 592
 - evolution of, 594
 - mining, 623
 - small-world phenomenon, 592
 - See also* networks
- social science/social studies data mining, 613
- soft clustering, 501
- soft constraints, 534, 539
 - example, 534
 - handling, 536–537
- space-filling curve, 58
- sparse data, 102
- sparse data cubes, 190
- sparsest cuts, 539
- sparsity coefficient, 579
- spatial data, 14
- spatial data mining, 595
- spatiotemporal data analysis, 319
- spatiotemporal data mining, 595, 623–624
- specialized SQL servers, 165
- specificity measure, 367
- spectral clustering, 520–522, 539
 - effectiveness, 522
 - framework, 521
 - steps, 520–522
- speech recognition, 430
- speed, classification, 369
- spiral method, 152
- split-point, 333, 340, 342
- splitting attributes, 333
- splitting criterion, 333, 342
- splitting rules. *See* attribute selection measures
- splitting subset, 333
- SQL, as relational query language, 10
- square-error function, 454
- squashing function, 403
- standard deviation, 51
 - example, 51
 - function of, 50
- star schema, 139
 - example, 139–140
 - illustrated, 140
 - snowflake schema versus, 140
- Star-Cubing, 204–210, 235
 - algorithm illustration, 209
 - bottom-up computation, 205
 - example, 207
 - for full cube computation, 210
 - ordering of dimensions and, 210
 - performance, 210
 - shared dimensions, 204–205
- starnet query model, 149
 - example, 149–150
- star-nodes, 205
- star-trees, 205
 - compressed base table, 207
 - construction, 205
- statistical data mining, 598–600
 - analysis of variance, 600
 - discriminant analysis, 600
 - factor analysis, 600
 - generalized linear models, 599–600
 - mixed-effect models, 600
 - quality control, 600

- regression, 599
- survival analysis, 600
- statistical databases (SDBs), 148
 - OLAP systems versus, 148–149
- statistical descriptions, 24, 79
 - graphic displays, 44–45, 51–56
 - measuring the dispersion, 48–51
- statistical hypothesis test, 24
- statistical models, 23–24
 - of networks, 592–594
- statistical outlier detection methods, 552, 553–560, 581
 - computational cost of, 560
 - for data analysis, 625
 - effectiveness, 552
 - example, 552
 - nonparametric, 553, 558–560
 - parametric, 553–558
 - See also* outlier detection
- statistical theory, in exceptional behavior disclosure, 291
- statistics, 23
 - inferential, 24
 - predictive, 24
- StatSoft, 602, 603
- stepwise backward elimination, 105
- stepwise forward selection, 105
- stick figure visualization, 61–63
- STING, 479–481
 - advantages, 480–481
 - as density-based clustering method, 480
 - hierarchical structure, 479, 480
 - multiresolution approach, 481
 - See also* cluster analysis; grid-based methods
- stratified cross-validation, 371
- stratified samples, 109–110
- stream data, 598, 624
- strong association rules, 272
 - interestingness and, 264–265
 - misleading, 265
- Structural Clustering Algorithm for Networks (SCAN), 531–532
- structural context-based similarity, 526
- structural data analysis, 319
- structural patterns, 282
- structure similarity search, 592
- structures
 - as contexts, 575
 - discovery of, 318
 - indexing, 319
 - substructures, 243
- Student's *t*-test, 372
- subcube queries, 216, 217–218
- sub-itemset pruning, 263
- subjective interestingness measures, 22
- subject-oriented data warehouses, 126
- subsequence, 589
 - matching, 587
- subset checking, 263–264
- subset testing, 250
- subspace clustering, 448
 - frequent patterns for, 318–319
- subspace clustering methods, 509, 510–511, 538
 - biclustering, 511
 - correlation-based, 511
 - examples, 538
- subspace search methods, 510–511
- subspaces
 - bottom-up search, 510–511
 - cube space, 228–229
 - outliers in, 578–579
 - top-down search, 511
- substitution matrices, 590
- substructures, 243
- sum of the squared error (SSE), 501
- summary fact tables, 165
- superset checking, 263
- supervised learning, 24, 330
- supervised outlier detection, 549–550
 - challenges, 550
- support, 21
 - association rule, 21
 - group-based, 286
 - reduced, 285, 286
 - uniform, 285–286
- support, rule, 245, 246
- support vector machines (SVMs), 393, 408–415, 437
 - interest in, 408
 - maximum marginal hyperplane, 409, 412
 - nonlinear, 413–415
 - for numeric prediction, 408
 - with sigmoid kernel, 415
 - support vectors, 411
 - for test tuples, 412–413
 - training/testing speed improvement, 415
- support vectors, 411, 437
 - illustrated, 411
 - SVM finding, 412
- supremum distance, 73–74
- surface web, 597
- survival analysis, 600
- SVMs. *See* support vector machines

symbolic sequences, 586, 588
 applications, 589
 sequential pattern mining in, 588–589
 symmetric binary dissimilarity, 70
 synchronous generalization, 175

T

tables, 9
 attributes, 9
 contingency, 95
 dimension, 136
 fact, 165
 tuples, 9
 tag clouds, 64, 66
 Tanimoto coefficient, 78
 target classes, 15, 180
 initial working relations, 177
 prime relation, 175, 177
 targeted marketing, 609
 taxonomy formation, 20
 technologies, 23–27, 33, 34
 telecommunications industry, 611
 temporal data, 14
 term-frequency vectors, 77
 cosine similarity between, 78
 sparse, 77
 table, 77
 terminating conditions, 404
 test sets, 330
 test tuples, 330
 text data, 14
 text mining, 596–597, 624
 theoretical foundations, 600–601, 625
 three-layer neural networks, 399
 threshold-moving approach, 385
 tilted time windows, 598
 timeliness, data, 85
 time-series data, 586, 587
 cyclic movements, 588
 discretization and, 590
 illustrated, 588
 random movements, 588
 regression analysis, 587–588
 seasonal variations, 588
 shapelets method, 590
 subsequence matching, 587
 transformation into aggregate approximations, 587
 trend analysis, 588
 trend or long-term movements, 588
 time-series data analysis, 319
 time-series forecasting, 588
 time-variant data warehouses, 127
 top-down design approach, 133, 151
 top-down subspace search, 511
 top-down view, 151
 topic model, 26–27
 top-*k* patterns/rules, 281
 top-*k* queries, 225
 example, 225–226
 ranking cubes to answer, 226–227
 results, 225
 user-specified preference components, 225
 top-*k* strategies
 comparison illustration, 311
 summarized pattern, 311
 traditional, 311
 TrAdaBoost, 436
 training
 Bayesian belief networks, 396–397
 data, 18
 sets, 328
 tuples, 332–333
 transaction reduction, 255
 transactional databases, 13
 example, 13–14
 transactions, components of, 13
 transfer learning, 430, 435, 434–436, 438
 applications, 435
 approaches to, 436
 heterogeneous, 436
 negative transfer and, 436
 target task, 435
 traditional learning versus, 435
 treemaps, 63, 65
 trend analysis
 spatial, 595
 in time-series data, 588
 for time-series forecasting, 588
 trends, data mining, 622–625, 626
 triangle inequality, 73
 trimmed mean, 46
 trimodal, 47
 true negatives, 365
 true positives, 365
t-test, 372
 tuples, 9
 duplication, 98–99
 negative, 364
 partitioning, 334, 337
 positive, 364
 training, 332–333
 two sample *t*-test, 373

two-layer neural networks, 399
two-level hash index structure, 264

U

ubiquitous data mining, 618–620, 625
uncertainty sampling, 433
undersampling, 384, 386
 example, 384–385
uniform support, 285–286
unimodal, 47
unique rules, 92
univariate distribution, 40
univariate Gaussian mixture model, 504
univariate outlier detection, 554–555
unordered attributes, 103
unordered rules, 358
unsupervised learning, 25, 330, 445, 490
 clustering as, 25, 445, 490
 example, 25
 supervised learning versus, 330
unsupervised outlier detection, 550
 assumption, 550
 clustering methods acting as, 551
upper approximation, 427
user interaction, 30–31

V

values
 exception, 234
 expected, 97, 234
 missing, 88–89
 residual, 234
 in rules or patterns, 281
variables
 grouping, 231
 predicate, 295
 predictor, 105
 response, 105
variance, 51, 98
 example, 51
 function of, 50
variant graph patterns, 591
version space, 433
vertical data format, 260
 example, 260–262

frequent itemset mining with, 259–262,
 272

video data analysis, 319
virtual warehouses, 133
visibility graphs, 537
visible points, 537
visual data mining, 602–604, 625
 data mining process visualization, 603
 data mining result visualization, 603
 data visualization, 602–603
 as discipline integration, 602
 illustrations, 604–607
 interactive, 604, 607
 as mining trend, 624
Viterbi algorithm, 591

W

warehouse database servers, 131
warehouse refresh software, 151
waterfall method, 152
wavelet coefficients, 100
wavelet transforms, 99, 100–102
 discrete (DWT), 100–102
 for multidimensional data, 102
 on sparse and skewed data, 102
web directories, 28
web mining, 597, 624
 content, 597
 as mining trend, 624
 structure, 597–598
 usage, 598
web search engines, 28, 523–524
web-document classification, 435
weight arithmetic mean, 46
weighted Euclidean distance, 74
Wikipedia, 597
WordNet, 597
working relations, 172
 initial, 168, 169
World Wide Web (WWW), 1–2, 4, 14
Worlds-with-Worlds, 63, 64
wrappers, 127

Z

z-score normalization, 114–115