INDEX

A posterior distribution 146
A priori algorithm 283
  Partition-based 287
  Sampling-based 287
  Incremental updating 287
  Concept hierarchy 288
A prior distribution 146
A priori knowledge 5, 88
Approximating functions 90
Activation function 201
Agglomerative clustering algorithms 260
Aggregation 16
Allela 386
Alpha cut 419
Alternation 396
Analysis of variance (ANOVA) 150, 155
Anchored visualization 457
Andrews's curve 452
Approximate reasoning 430
Approximation by rounding 76
Artificial neural network (ANN) 105, 199
Artificial neural network, architecture 205
  feedforward 205
  recurrent 205
  Competitive 221
  Self-organizing map (SOM) 225
Artificial neuron 202
Association rules 105, 280
  Apriori 283
  FPgrowth 288
  Classification based on multiple association rules (CMAR) 290
Asymptotic consistency 94
Autoassociation 211
Authorities 305

Bar chart 450
Bayesian inference 146
Bayesian networks 370
Bayes theorem 147
Binary features 255
Bins 74
Bins cutoff 74
Bootstrap method 125
Boxplot 144
Building blocks 402

Candidate counting 283
Candidate generation 283
Cardinality 419
Cases reduction 104
Causality 369
Censoring 41
Centroid 252, 263
Chameleon 262
Change detection 3, 104
Chernoff's faces 453
ChiMerge technique 77
Chi-squared test 79, 161
Chromozome 386
Circular coordinates 458
City block distance 254
Classification 3
  CART 189
  C4.5 185
  ID3 172
  k-NN 118
  SVM 105
Classifier 118, 139, 240
CLS 173
Cluster analysis 105, 249

Data Mining: Concepts, Models, Methods, and Algorithms, Second Edition. Mehmed Kantardzic.

© 2011 by Institute of Electrical and Electronics Engineers. Published 2011 by John Wiley & Sons, Inc.

Cluster feature vector (CF) 268
Clustering 3, 250
  BIRCH 272
  DBSCAN 270
  Validation 275
  k-means 264
  k-medoids 266
  Incremental 266
  Using genetic algorithms 409
Clustering tree 252
Competitive learning rule 221
Complete-link method 260
Confidence 282, 291
Confirmatory visualization 450
Confusion matrix 126
Contingency table 77, 159
Control theory 5, 208
Core 419
Correlation coefficient 154
Correspondence analysis 159
Cosine correlation 255
Covariance matrix 45, 62
Crisp approximation 438
Crossover 392
Curse of dimensionality 29

Data cleansing 15
Data scrubbing 15
Data collection 7
Data constellations 454
Data cube 451
Data discovery 6
Data integration 15
Data mart 14
Data mining 2
  Privacy 377
  Security 379
  Regal aspects 378
Data mining process 6
Data mining roots 4
Data mining tasks 3
Data preprocessing 8
Data quality 13
Data set 2
  Iris 72
  messy 32
  preparation 33
  quality 13
  raw 32
  semistructured 11
  structured 11
  temporal 28
  time-dependent 37
  transformation 15
  unstructured 11
Data set dimensions 54
  cases 54
  columns 54
  feature values 54
Data sheet 454
Data smoothing 34
Data types,
  alphanumeric 11
  categorical 27
  dynamic 28
  numeric 11, 26
  symbolic 27
Data warehouse 14
Data representation 26, 395
Decimal scaling 33
Decision node 173
Decision rules 105, 185
Decision tree 105,183
Deduction 88
Default class 188
Defuzzification 433
Delta rule 208
Dendogram 262
Dependency modeling 3, 103
Descriptive accuracy 55
Descriptive data mining 2
Designed experiment 7
Deviation detection 3, 104
Differences 35
Dimensional stacking 454
Directed acyclic graph (DAG) 371
Discrete optimization 388
Discrete Fourier Transform 348
Discrete Wavelet Transform 348
Discriminant function 163
Distance error 75
Distance measure 254
Distributed data mining 360
  Distributed DBSCAN 366
Divisible clustering algorithms 260
Document visualization 319
Domain-specific knowledge 7
Don't care symbol 400

Eigenvalue 72
Eigenvector 72
Empirical risk 94
Empirical risk minimization (ERM) 94
Encoding 8
Encoding scheme 389
Ensemble learning 235
  Bagging 241
  Boosting 242
  AdaBoost 243
Entropy 70
Error back-propagation algorithm 214
Error energy 208
Error-correction learning 208
Error rate 126
Euclidean distance 65, 69, 120, 254
Exponential moving average 39
Exploratory analysis 2
Exploratory visualizations 450
Extension principle 426

False acceptance rate (FAR) 130
False reject rate (FRT) 130
Fault tolerance 201
Feature discretization 79
Features composition 58
Features ranking 59
Features reduction 56
Features selection 57
  Relief 66
Filtering data 212
First-principle models 1
Fitness evaluation 390
Free parameters 100, 207
F-list 291
FP-tree 281
Function approximation 211
Fuzzy inference systems 105
Fuzzy logic 429
Fuzzy number 420
Fuzzy relation 425
  containment 425
  equality 425
Fuzzy rules 430
Fuzzy set 415
Fuzzy set operation 420
  complement 421
  cartesian product 421
  concentration 424
  dilation 424
  intersection 420
  normalization 424
  union 420
Fuzzification 433

Gain function 175
Gain-ratio function 180
Gaussian membership function 418
Gene 386
Generalization 95, 122, 219, 288, 301
Generalized Apriori 171
Generalized modus ponens 431
Genetic algorithm 105, 386
Genetic operators 387, 390
  crossover 392
  mutation 393
  selection 390
Geometric projection visualization 451
GINI index 189
Glyphs 453
Gradviz 460
Graph mining 329
  Centrality 336
  Closeness 336
  Betweenness 336
Graph compression 341
Graph clustering 341
Gray coding 390
Greedy optimization 97
Grid-based rule 436
Growth function 95

Hamming distance 422
Hamming networks 223
Hard limit function 203
Heteroassociation 211
Hidden node 217
Hierarchical clustering 252
Hierarchical visualization techniques 454
Histogram 450
Holdout method 125
Hubs 305
Hyperbolic tangent sigmoid 203
Hypertext 317

Icon-based visualization 453
Induction 88
Inductive-learning methods 97

Inductive machine learning 89
Inductive principle 93
Info function 175
Information visualization 450
Information retrieval (IR) 304
Initial population 395
Interesting association rules 286
Internet searching 316
Interval scale 27
Inverse document frequency 317
Itemset 283

Jaccard coefficient 256

Kernel function 114
Knowledge distillation 318

Large data set 9
Large itemset 283
Large reference sequence 312
Lateral inhibition 222
Latent semantic analysis (LSA) 320
Learning machine 89
Learning method 89
Learning process 88
Learning tasks 101
Learning theory 93
Learning rate 209
Learning system 99
Learning with teacher 99
Learning without teacher 99
Leave-one-out method 125
Lift chart 126
Line chart 450
Linear discriminant analysis (LDA) 162
Linguistic variable 424
Local gradient 216
Locus 386
Logical classification models 173
Log-linear models 158
Log-sigmoid function 203
Longest common sequence (LCS) 349
Loss function 92

Machine learning 4
Mamdani model 436
Manipulative visualization 450
Multivariate analysis of variance (MANOVA) 156
Market basket analysis 281
Markov Model (MM) 351
  Hidden Markov Model (HMM) 351
Max-min composition 428
MD-pattern 294
Mean 34, 44, 60, 143
Median 143
Membership function 416
Metric distance measure 254
Minkowski metric 255
Min-max normalization 33
Misclassification 92
Missing data 36
Mode 143
Model 6
  estimation 126
  selection 122
  validation 126
  verification 126
Momentum constant 218
Moving average 32
Multidimensional association rules 293
Multifactorial evaluation 433
Multilayer perceptron 213
Multiple discriminant analysis 164
Multiple regression 152
Multiscape 454
Mutual neighbor distance (MND) 258

Naïve Bayesian classifier 147
N-dimensional data 101
N-dimensional space 101
N-dimensional visualization 105
N-fold cross-validation 125
Necessity measure 423
Negative border 285
Neighbor number (NN) 258
Neuro-Fuzzy system 442
Nominal scale 27
Normalization 33
NP hard problem 47
Null hypothesis 142

Objective function 110, 388
Observational approach 7
OLAP (Online analytical processing) 17
Optimization 98
Ordinal scale 28
Outlier analysis 41

Outlier detection 7, 42
Outlier detection, distance based 45
Overfitting (overtraining) 97

PageRank algorithm 313
Parabox 454
Parallel coordinates 455
Parameter identification 5
Partially matched crossover (PMC) 403
Partitional clustering 263
Pattern 6
Pattern association 211
Pattern recognition 211
Pearson correlation coefficient 61
Perception 488
Perceptron 200
Pie chart 451
Piecewise aggregate approximation (PAA) 346
Pixel-oriented visualization 453
Population 141, 390
Possibility measure 423
Postpruning 184
Prediction 2
Predictive accuracy 124
Predictive data mining 2
Predictive regression 149
Prepruning 184
Principal Component Analysis (PCA) 70
Principal components 64
Projected database 292
Pruning decision tree 184

Radial visualization (Radviz) 460
Random variable 150
Rao's coefficient 256
Ratio scale 27
Ratios 35
Receiver operating characteristic (ROC) 130
Receiver operating characteristic (ROC) curve 130
Regression 3,102
  Logistic 157
  Linear 150
  Nonlinear 153
  Multiple 152
Regression equation 150
Resampling methods 124, 150
Resubstitution method 125
Return on investment (ROI) chart 129
Risk functional 92
Rotation method 125
RuleExchange 408
RuleGeneralization 408
RuleSpecialization 408
RuleSplit 408

Sample 6
Sampling 81
  average 82
  incremental 82
  inverse 83
  random 82
  stratified 83
  systematic 82
Saturating linear function 203
Scaling 8
Scatter plot 451
Schemata 399
  fitness 401
  length 401
  order 400
Scientific visualization 449
Scrubbing 15
Sensitivity 98
Sequence 311
Sequence mining 311
Sequential pattern 351
Similarity measure 68, 253, 349
Simple matching coefficient (SMC) 256
Single-link method 260
Smoothing data 212
Spatial data mining 357
  Autoregressive model 359
  Spatial outlier 359
Specificity 131
Split-info function 182
SQL (Structured query language) 17
SSE (Sum of squares of the errors) 150
Standard deviation 34, 43, 144
Star display 453
Statistics 4
Statistical dependency 91
Statistical inference 140
Statistical learning theory (SLT) 93
Statistical methods 105
Statistical testing 142
Stochastic approximation 97

Stopping rules 98
Strong rules 282
Structure identification 5
Structural risk minimization (SRM) 96
Summarization 3, 103
Supervised learning 99
Support 282, 291, 339, 354, 419
Survey plot 452
Survival data 41
Synapse 201
System identification 5

Tchebyshev distance 422
Temporal data Mining 343
  Sequences 344
  Time series 344
Test of hypothesis 142
Testing sample 119, 192
Text analysis 316
Text database 316
Text mining 316
Text-refining 319
Time lag (time window) 37
Time series, multivariate 40
Time series, univariate 40
Training sample 94, 214, 239
Transduction 88
Traveling salesman problem (TSP) 402
Trial and error 6
True risk functional 94

Ubiquitous data mining 356
Underfitting 97
Unobserved inputs 13
Unsupervised learning 99

Value reduction 73
Variables 12
  continuous 27
  discrete 27
  categorical 27
  dependent 12
  independent 12
  nominal 27
  numeric 26
  ordinal 28
  periodic 28
  unobserved 13
Variance 61
Variogram cloud technique 359
Vapnik-Chervonenkis (VC) theory 93
Vapnik-Chervonenkis (VC) dimension 95
Visual clustering 466
Visual data mining 449
Visualization 448
Visualization tool 450
Voronoi diagram 119

Web mining 300
  content 302
  HITS(Hyperlink-Induced Topic Search) algorithm 306
  LOGSOM algorithm 308
  path-traversal patterns 310
  structure 304
  usage 304
Web page content 301
Web page design 301
Web page quality 302
Web site design 301
Web site structure 302
Widrow-Hoff rule 208
Winner-take-all rule 222, 227

XOR problem 206