

This excerpt from

Principles of Data Mining.  
David J. Hand, Heikki Mannila and Padhraic Smyth.  
© 2001 The MIT Press.

is provided in screen-viewable form for personal use only by members of MIT CogNet.

Unauthorized use or dissemination of this information is expressly forbidden.

If you have any questions about this material, please contact  
[cognetadmin@cognet.mit.edu](mailto:cognetadmin@cognet.mit.edu).

# *Index*

- A Priori algorithm, 143, 147, 157–160
- A priori relationships, 197
- Absolute error, 216
- Accurate measurements, 45
- Actual error rate, 359
- Acyclic directed graphical models, 290
- AD-Tree data structure, 425
- Additive form, 189
- Additive predictor, 393–394
- Advanced Scout* system, 14
- Agglomerative methods, 308, 311–314
- Aggregate, 414
- Aggregation, 414
- Akaike information criterion (AIC), 225
- Algebra, relational, 409
- Algorithm parameters, 267
- Algorithms. *See* Data mining algorithms; *specific types*
- Alternative hypothesis, 124
- ANNs, 391–393
- Apparent error rate, 359
- Approximation, 169, 265, 322–323
- Artificial neural networks (ANNs), 391–393
- ASCII characters and codes, 206–207
- Association analysis, 315
- Association rules, 14, 158–160, 433–435
- Assumption, 289
- AT&T, 13, 19
- Attributes, 4, 405
- Automated recommender systems, 471–472
- Autoregressive models, 199–200, 202, 478
- Average of points, 297
- B-trees, 402–403
- Backfitting algorithms, 394
- Backpropagation method, 256
- Backward elimination, 380
- Backward selection algorithms, 243
- Bandwidth, 176, 285, 350
- Basic algorithms for partition-based clustering, 302–308
- Basis functions, 195
- Basket data, 405–406
- Bayes assumption, first-order, 354
- Bayes error rate, 334
- Bayes factor, 130
- Bayes rule, 337–338
- Bayes theorem, 353
- Bayesian approximations, 322–323
- Bayesian estimation, 93, 96, 106, 116–124, 220, 283
- Bayesian Information Criterion (BIC), 225–227, 235, 292, 380
- Bayesian model, 120, 359, 361–362
- Beam search strategy, 246, 440
- Beam width search strategy, 246

- Belief networks, 290
- Bernoulli distributions, 487
- Best classification tree problem, 241
- Best unbiased estimators, 107
- Beta distributions, 119
- Beta posterior, 122–123
- Beta prior, 122–123
- Between-cluster variation, 297–298
- Bias of measurements, 45
- Bias-variance, 221–224
  - trade-off, 223
- Biased estimation, 106
- Biasing, 283–284
- BIC, 225–227, 235, 292, 380
- Binary data, 36–37
- Binomial distributions, 487
- Blind search, 245–246
- Bonferroni inequality, 131
- Boolean conditions and propositions, 213, 411, 429, 458, 461, 475
- Boosting methods, 358
- Bootstrap methods, 116, 360–361
- Box, George, 168
- Boxplots, 61–63
- Bracketing methods, 254–255
- Branch-and-bound concept, 246–247
- Breadth-first search, 245
- Brent's method, 254
- Brushing, 71
- Building models, 378–381
  
- Canonical discriminant function, 343
- Canonical parameter, 388
- CART algorithms, 145–151, 157, 228, 335, 345
- Cartesian product operations, 409, 412
- Cases, 4
- Categorical data, 187, 287–292
- Categorical variables, 6
- Causation, 101–102
- Central limit theorem, 115–116
- Centroid of points, 297
  
- Chaining, 312–313
- Chance, 93–97. *See also* Uncertainty
- Chernoff faces, 74
- Chi-squared distributions, 489–490
- Chomsky hierarchy of grammars, 202
- City-block metric, 36
- Class identifiers, 367
- Class of patterns, 204
- Class variables, 329
- Class-conditional approach, 335–339
- Classical hypothesis testing, 124–130
- Classical multidimensional scaling, 86
- Classification
  - document, 469–470
  - in high dimensions, feature selection for, 362–363
  - maximal predictive, 301
  - multilayer perceptrons for, 153–157
  - predictive models for, 327–366
    - classification models and, 329–339
    - evaluating and comparing, 359–362
    - feature selection for high-dimension, 362–363
    - linear discriminants and, 341–343
    - logistic discriminant analysis, 352–353
    - naive Bayes model, 353–356
    - nearest neighbor methods, 347–352
    - other methods, 356–359
    - overview, 180–182, 327–329
    - perceptrons and, 339–341
    - tree models, 343–347
  - text, 469–470
- Classification And Regression Trees (CART), 145–151, 153, 228, 335, 345
- Classification learning, 169, 328–329
- Classification models
  - background information, 329–330

- building real classifiers and, 335–339
- decision boundaries, 330–331
- discriminative, 330–331
- probabilistic models and, 331–334
- Classifiers
  - building real, 335–339
  - evaluating and comparing, 359–362
- Closed form methods, 249–250
- Cluster analysis, 12, 293–296, 323
- Cluster centers, 297
- Clustering techniques, 12–13, 279.
  - See also* Hierarchical clustering; Partition-based clustering algorithms; Probabilistic model-based clustering using mixture models
- Clusters, 137
- Co-occurrence patterns, 158
- Coding, image, 166–167
- Coefficients, 35, 37, 361
- Collaborative filtering, 471–472
- Collision, 404
- Column vectors, 36
- Combinatorial optimization, 236, 239
- Commensurability, 32
- Complete link method, 313
- Complexity of models
  - nesting and, 172
  - scoring and, 220–228
    - bias-variance, 221–224
    - concepts in comparing, general, 220–221
    - penalizing, 224–227
    - validation and, external, 227–228
  - in selecting predictive models, 183
- Compression, data, 166
- Computational methods, 141, 235, 291
- Computational resources, 268
- Concatenation property, 27
- Condensed nearest neighbor methods, 352
- Conditional density, 98
- Conditional error rate, 359
- Conditional independence (“naive”)
  - Bayes model, 191, 353–356
- Conditionally independent variables, 99–100, 289, 354
- Confidence
  - interval, 115
  - itemsets and, frequent, 430–431
  - limits, 115
- Confusion matrix, 361
- Conjugate directions, 258
- Conjugate families of distributions, 122–123
- Constrained optimization, 259–260
- Constraints, 10
- Content retrieval. *See* Retrieval by content
- Context-free grammar, 202
- Contingency table, 188
- Contour plots, 65–67
- Convenience samples, 21, 48
- “Cookbook” approach, 152–153
- Cosine distance, 459
- Counts, 31
- Covariance matrix, 78, 299
- Covariances, 33–35
- Coverage of a pattern, 214
- Coxcomb plot, 11
- Credibility interval, 123
- Critical region, 125
- Cross-validation, 148–149, 227–228, 322, 360
- Cumulative distribution function, 485
- Curse of dimensionality, 19, 193–196
- Customer transactions, 405–406
- Data. *See also* Databases; Graphical data exploration methods; Measurement and data analysis, 166–167
  - background information, 25–26

- basket, 405–406
- binary, 36–37
- categorical, 187, 287–292
- compression, 166
- cube, 419–420
- defined, 25
- experimental, 1
- flattened, 7, 20, 43, 358
- form of, 41–44
- geographic, 44
- high-dimensional, 194–196, 362–363
- image, 44
- market-basket, 158, 429–430
- maximum variability in, 77
- mode and, 56
- model, 405
- multirelational, 42–43
- observational, 1
- orthogonality of, 240
- “out-of-sample,” 227, 328, 372
- quality, 44–51
  - for collection of data, 47–51
  - for individual measurements, 44–47
  - poor, 51
- repeated measures, 349–350
- sequential, 477
- spatial, 44
- standard, 41
- structured, models for, 197–203
- summarizing, 54–57
- summary information, 52
- suspect, 50–51
- in table, 41
- time series, 476–481
- transforming, 38–41, 194–196, 363
- unordered categorical, joint distributions for, 187
- warehousing, 417–419
- Data management techniques, 17–18, 143, 296, 421–426. *See also* Databases
- Data matrix, 41, 203–206. *See also* Data sets
- Data mining. *See also* Data mining algorithms
  - analysis of, 144
  - background information, 1–4
  - data sets and, 4–9
  - databases and, 421–426
  - defined, 1
  - dredging and, 22–23
  - fishing and, 22–23
  - interactive, 11, 450
  - keyword spotting and, 479
  - knowledge discovery in databases and, 3
  - models and, 1–2, 10–11, 175, 271
  - patterns and, 1–2, 10–11, 271
  - samples and, 93
  - snooping and, 22–23
  - statistics and, 18–21
  - structures, 9–11, 142
  - summary information, 23–24
  - synthesis of, 144
  - tasks, 11–15, 142
  - visual techniques and, 11
- Data mining algorithms. *See also* Score functions for data mining algorithms
  - background information, 141–145
  - Classification And Regression Trees, 145–151, 153, 228, 335, 345
  - components of, 15–18, 142–145
  - defined, 141
  - nonscalable versions of, 424
  - reductionist viewpoint, 151–162
    - A Priori algorithm, 157–160
    - background information, 151–153
    - multilayer perceptrons for regres-

- sion and classification and, 153–157
  - vector-space for text retrieval and, 160–162
  - scalable versions of, 423–424
  - summary information, 162–164
  - tuple, 146, 151, 154–155
- Data sets. *See also* Databases
  - data mining and, 4–9
  - defined, 4, 7
  - heterogeneous, 279
  - likelihood of, 108–109
  - massive, 421–426
  - nature of, 4–9
  - pseudo, 425–426
- Data-driven hypothesis generation, 53
- Data-squashing, 425
- Databases. *See also* Data set
  - aggregation in, 414
  - background information, 399–400
  - data mining and, 421–426
  - data model in, 405
  - data warehousing, 417–419
  - index structures, 402–404
  - knowledge discovery in, 3
  - management of data and, 421–426
  - manipulating tables and, 409–412
  - massive data sets and, 421–426
  - memory hierarchy, 400–401
  - multidimensional indexing, 404–405
  - online analytical processing, 417–419
  - operational, 417
  - purpose of, 400
  - query execution and optimization, 415–417
  - relational, 405–409
  - strategic, 417
  - string, 420–421
  - Structured Query Language, 409, 413–415
- Deciles, 56
- Decision boundaries, 330–331
- Decision region, 330
- Decision surfaces, 330–331
- Decision trees, 422
- Degrees of freedom, 376–377, 489–490
- Dendrograms, 313
- Density estimation, 12, 184
- Density function, 97–98, 355, 485. *See also* Probability distribution and density function models
- Density mixtures, 279–281
- Density models, parametric, 275–279. *See also* Probability distribution and density function models
- Denumerable domain, 485
- Dependency modeling, 12
- Dependent variables, 35
- Depth-first search, 245
- Derived variables, 198
- Descriptive models
  - background information, 271–272
  - cluster analysis, 293–296
  - functions of, 12–13
  - goal of, 12
  - hierarchical clustering, 308–315
    - agglomerative methods, 308, 311–314
    - background information, 308–311
    - divisive methods, 308, 314–315
  - nonprobabilistic, 219
  - partition-based clustering algorithms, 296–308
    - background information, 296–297
    - basic algorithms for, 302–308
    - for nonprobabilistic predictive models, 219
    - score functions for, 296–302

- probabilistic model-based clustering, 315–323
  - advantages, 319
  - background information, 315–316
  - disadvantages, 319–321
  - examples, 316–319
  - techniques, 321–323
- probability distribution and density function models and, 272–292
  - background information, 272–274
  - Expectation Maximization algorithm for, 281–284
  - joint distributions for categorical data, 287–292
  - mixture distributions and densities, 279–281
  - nonparametric density estimation, 284–287
  - parametric density models, 275–279
  - score functions for, 274–275
  - score functions for, 212, 217–219
- Deviance of model, 389–390
- Diagnostic methods, 10, 338, 381–384
- Dice coefficient, 37
- Difference operation, 410
- Discovery task, pattern, 205
- Discriminant functions, 331
- Discriminative approach, 335–339
- Discriminative classification, 330–331
- Disk access, special-purpose algorithms for, 424
- Dispersion measurement, 56
- Dispersion parameter, 388
- Dissection, 293
- Distance
  - cosine, 459
  - distance, 32–33, 85
  - edit-distance, 312
  - Euclidean, 32–33, 85, 459, 480
  - Mahalanobis, 276–277
  - measurements, 31–38
  - minimum, 298
  - pairwise, 312
  - between queries and documents, 462
  - similarity and, 15, 451
  - weighted Euclidean, 33
- Distortion of samples, 49–50
- Distribution-free tests, 129
- Distributions. *See also* Probability distribution and density function models
  - Bernoulli, 487
  - Beta, 119
  - binomial, 487
  - chi-squared distribution, 489–490
  - conjugate families of, 122–123
  - exponential family of, 388
  - F, 490
  - finite mixture, 280
  - independently and identically distributed, 108
  - joint
    - for categorical data, 287–292
    - for unordered categorical data, 187
  - left-skewed, 57
  - mixture, 279–281
  - multimodal, 56, 60
  - multinomial, 487–488
  - multivariate normal, 490
  - Normal, 60, 113, 115–116, 118, 121–122, 127, 171, 276, 350, 488
  - Poisson, 280–281, 388, 488
  - posterior, 117, 122–123
  - predictive, 120–121
  - prior, 117, 122–123
  - probability, 485, 487–490
  - relative, 459
  - right-skewed, 57
  - skewness of, 56–57

- student's *t*-, 489
- Divisive methods, 308, 314–315
- Document, 456, 461–465, 469–470
- Dredging, 22–23
- Duplicates, 411
- EDA, 11–12
- Edit-distance, 312
- Edited nearest neighbor methods, 352
- EFFORT (software program), 29–30
- EM algorithms. *See* Expectation Maximization algorithms
- Entities, 4
- Episodes, 207–208, 436–438
- Epsem sample, 134
- Errors
  - absolute, 216
  - actual error rate, 359
  - apparent error rate, 359
  - Bayes error rate, 334
  - conditional error rate, 359
  - defined, 373
  - estimation, 216
  - family error rate, 131
  - mean squared, 107, 223–224
  - misclassification of objects and, 359–361
  - quadratic error function, 340
  - resubstitution error rate, 359
  - risk of, 45
  - squared, 216
  - true error rate, 359
- Estimation
  - Bayesian, 93, 96, 106, 116–124, 220, 283
  - biased, 106
  - cross-validation, 148–149
  - defined, 93
  - density, 12, 184
  - errors, 216
  - maximum likelihood, 96, 106, 108–116
  - nonparametric density, 284–287
  - over, 216
  - parameter, 240
  - probability distribution and density, 274–275
  - quasilielihood, 390
  - query selectivity, 273
  - regression and, 13
  - stochastic, 123, 265
  - unbiased, 106, 227
  - uncertainty and, 105–124
    - background information, 105–106
    - Bayesian, 93, 116–124
    - maximum likelihood and, 93, 108–116
    - properties of estimators and, desirable, 106–108
    - stochastic, 123, 265
  - under, 216
- Estimators, 106–109
- Euclidean distance, 32–33, 85, 459, 480
- Euclidean metric, 36
- Euclidean space, 298
- Evaluation
  - of classifiers, 359–362
  - of models and patterns, 229–231
  - of retrieval systems, 452–456
- Event-sequence, 43
- “Exclusive-or” structure, 71
- Expectation Maximization (EM) algorithms
  - function of, 21
  - for mixture models, 187, 281–284
  - optimization and, 260–265
  - red blood cell example, 317–318
- Expected value, 486
- Experimental data, 1
- Experimental design, 132
- Explainable variation, 179
- Explanatory variable, 168



- Exploratory data analysis (EDA), 11–12
- Exploring data. *See* Graphical data exploration methods
- Exponential family of distributions, 388
- Expressive power of model structure, 183
- F distributions, 490
- Factor analysis, 83
- Factor loadings, 83
- Factored form, 292
- Factorization, 187–193, 290
- Factors, 195
- Family error rate, 131
- Family of model structures, 238
- Fate, 93–97. *See also* Uncertainty
- Feasible region, 259
- Feature extraction approach, 197–198
- Feature selection for classification in high dimensions, 362–363
- Features, 4
- Feed-forward neural networks, 357, 391
- Fields, 4, 202
- File, inverted, 461
- Filtering, collaborative, 471–472
- Finite mixture distributions, 280
- Finite state machine (FSM), 202
- First normal form, 408
- First-order autoregressive model, 199–201
- First-order Bayes assumption, 354
- First-order Markov property, 101
- Fisher information, 122
- Fisher, R.A., 341
- Fisher's linear discriminant analysis method, 331, 353, 356, 362
- Fishing, 22–23
- Fitted model, 10–11
- Flattened data, 7, 20, 43, 358
- Forecasting, 133
- Form of data, 41–44
- Forward selection algorithms, 243, 379
- Freedom, degrees of, 376–377, 489–490
- Frequency of episode, 436–437
- Frequent itemsets, 429–433
- Frequent sets, 204, 431, 433–435
- Frequentist view of probability, 95
- From clause, 413
- FSM, 202
- Function approximation problems, 169
- Functional dependency, 206
- Furthest neighbor methods, 313
- Gaussian noise, 199
- Generalizations, 295, 377–378, 435–436, 476
- Generalized additive models, 393–395
- Generalized linear models, 173, 353, 384–390
- Generative models, 272
- Generic score functions, 16, 219
- Genetic search, 266–267
- Geographic data, 44
- GIGO, 44–45
- Gini coefficient of performance, 361
- Global models, 442–443, 478–480
- Global pattern, 9
- Goodness-of-fit tests, 126, 142, 372, 377
- Google system, 15
- Grades, 31
- Gradient descent method, 253
- Gradient-based methods, 250–251
- Grammars, 202
- Graphical data exploration methods
  - background information, 53–55
  - hypothesis testing and, 53
  - multidimensional scaling, 84–90

- principal components analysis, 74–84
- summarizing data, 54–57
- visual techniques
  - for more than two variables, 70–74
  - for relationships between two variables, 62–70
  - for single variables, 57–62
- Graphical models, 189–190
- Greedy heuristic search methods, 241
- Hash indices, 403–404
- Hazard, 93–97. *See also* Uncertainty
- Heterogeneous data set, 279
- Heteroscedasticity, 381
- Heuristic search methods, 241, 244–246, 439–440
- Hidden Markov models (HMMs), 201–202, 291
- Hidden variables, 187, 190–191, 195
- Hierarchical clustering
  - agglomerative methods, 308, 311–314
  - background information, 308–311
  - divisive methods, 308, 314–315
- Hierarchical structure, 44
- High-dimensional data, 194–196, 362–363
- “Hill-climbing” algorithm, 244
- Histograms, 57–59, 61, 284
- HMMs, 201–202, 291
- Homoscedasticity, 381
- Horseshoe effect, 88
- Hypertetrahedron, 258
- Hypothesis testing
  - graphical data exploration methods and, 53
  - random variables and, 99
  - uncertainty and, 124–132
    - background information, 124
    - classical, 124–130
    - in context, 130–132
- IBM, 474
- Icon plot, 74
- Icons, 74
- Idealization, 95
- IDF, 463
- iid, 108
- Image
  - coding, 166–167
  - form of data and, 44
  - invariants, 475–476
  - local part of, 166
  - queries, 473–474
  - representation, 473
  - retrieval, 472–476
  - understanding, 473
  - whole, 166
- Improper priors, 122
- Independence in high dimensions, 187–193
- Independent variables, 99, 188–189
- Independently and identically distribution (iid), 108
- Indicator matrix, 429–430
- Individual contribution, 170
- Individual preferences, modeling, 470–472
- Individual X variables, 194–195
- Individuals, 4
- Inference, 377–378
- Information retrieval (IR). *See* Text retrieval
- Input variable, 329
- Inspection, model, 381–384
- Interactive techniques, 11, 456
- Interestingness, criteria for, 440–441
- Interquartile range, 56
- Intersection operation, 410
- Interval scale, 28–29
- Inverse-document-frequency (IDF), 463

- Inverted file, 461
- IR. *See* Text retrieval
- ISODATA algorithm, 307
- Itemsets, frequent, 429–433
- Iteratively weighted least square method, 258, 389
- Jaccard coefficient, 37
- Jackknife methods, 360–361
- Jeffrey’s prior, 122
- Join operations, 412
- Joint density function, 97–98
- Joint distributions
  - for categorical data, 287–292
  - for unordered categorical data, 187
- $K$ -means algorithms, 298, 305
- $k$ -nearest neighbor method, 348–349
- Kalman filters, 201–202
- KDD, 3
- Kernel density method, 284
- Kernel estimates, 59–62, 176
- Kernel function, 285
- Kernel methods, 176–178
- Kernel models, 287
- Kernel plots, 61
- Keyword spotting, 479
- Knowledge discovery in databases (KDD), 3
- Kolmogorov-Smirnov test statistic, 129–130
- $k$ th mixing proportion, 281
- $k$ th-order Markov model, 200
- Kuhn-Tucker conditions, 260
- Lagrange multipliers, 259–260
- Laplace approximation, 323
- Latent semantic indexing (LSI), 465–469
- Latent variables, 187, 190–191, 195
- Least squares fitting
  - computational issues in, 370–372
  - defined, 370
  - diagnostic methods and, 381–384
  - generalization and, 377–378
  - inference and, 377–378
  - interpreting, 375–377
  - model building and, 378–381
  - model inspection and, 381–384
- Least squares method, 114, 211, 370
- Leaving-one-out method, 360
- Lee, M., 425
- Left-skewed distributions, 57
- Length variables, 32
- Letters, 206. *See also* Strings
- Likelihood function, 105, 108–109, 274–275
- Likelihood ratio, 125–126
- Linear algebra methods, 249–250
- Linear correlation, 35
- Linear covariance, 35
- Linear dependencies, 35
- Linear discriminants, 341–343
- Linear function, 9
- Linear models
  - background information, 368–370
  - diagnostic methods and, 381–384
  - generalization and, 377–378
  - generalized, 384–390
  - global, 478
  - inference and, 377–378
  - inspection, 381–384
  - model building and, 378–381
  - probabilistic interpretation of, 372–375
- Linear predictor, 388
- Linear programming, 259
- Linear regression models. *See* Linear models
- Linear structure, regression models with, 169–173
- Local exploration, 243
- Local extremum, finding, 251
- Local improvement, 241
- Local part of image, 166

- Local piecewise model structures for regression, 174–175
- Locally linear, 174
- Locally weighted regression model, 175–176
- Location measurements, 55
- Location parameters, 184
- Loess regression model, 175–176
- Log-likelihood, 122, 274–275
- Log-linear models, 292
- Logistic discriminant analysis, 352–353
- Logistic link function, 385
- Logistic regression, 384–385
- Logit link function, 385
- Logit transformation, 40
- “Lower resolution” data samples, 11
- LSI, 465–469
- Luck, 93–97. *See also* Uncertainty
  
- Mahalanobis distance, 276–277
- Manhattan metric, 36
- Manipulation of variables, 168
- MAP method, 117, 226, 283, 291
- Marginal density, 98
- Marginal likelihoods, 130, 226
- Market-basket data, 158, 429–430
- Markov chain model, 189–190, 202, 290
- Markov Chain Monte Carlo (MCMC) methods, 123, 268
- Markov linear-switching model, 479–480
- Markov random fields, 202
- Massive data sets, 421–426
- Mathematical programming, 259
- Maximal predictive classification, 301
- Maximum likelihood estimation, 93, 106, 108–116
- Maximum likelihood estimator (MLE), 109, 113
  
- Maximum a posteriori (MAP) method, 117, 226, 283, 291
- Maximum variability in data, 77
- MCMC methods, 123, 268
- MDL method, 226
- Mean squared error (MSE), 107, 223–224
- Measurements. *See also* Data
  - accurate, 45–46
  - amounts and, 31
  - background information, 25–26
  - balances and, 31
  - bias of, 45
  - counted fractions and, 31
  - counts versus, 31
  - dispersion, 56
  - distance, 31–38
  - grades and, 31
  - individual data quality for, 44–47
  - location, 55
  - metrical versus categorical, 31
  - pairs of, 327
  - precise, 45
  - qualitative versus quantitative, 31
  - ranks and, 31
  - reliability of, 46
  - representational, 29–31
  - summary information, 52
  - types of, 26–31
  - validity of, 46–47
  - variability, 56
- Median, 55
- Memory hierarchy, 400–401
- Minimum description length (MDL) method, 226
- Minimum distance, 298
- Minkowski metric, 36
- Missing data, optimization with, 260–265
- Mixture distributions and densities, 279–281
- Mixture models

- autoregressive models, 202
- parametric, 185–187
- probabilistic model-based clustering using, 315–323
  - advantages, 319
  - background information, 315–316
  - disadvantages, 319–321
  - examples, 316–319
  - techniques, 321–323
- and radial basis function approaches, 357
- MLE, 109, 113
- MLPs, 153–157, 357, 391
- Mode, 56
- Model averaging methods, 346
- Models. *See also* Complexity of models; Patterns; *specific types*
  - background information, 165–167
  - building, 378–381
  - classes of structure, 235, 238
  - curse of dimensionality and, 193–196
  - data, 405
  - data mining and, 1–2, 10–11, 175, 271
  - defined, 165
  - deviance of, 389–390
  - evaluation of, 229–231
  - expressive power of, 183
  - fundamentals, 167–168
  - generalized linear, 173, 353, 384–390
  - generative, 272
  - global, 442–443, 478–480
  - goal of, 102
  - for individual preferences, 470–472
  - inspection of, 381–384
  - $k$ th order Markov, 200
  - Markov chain, 189–190, 202, 290
  - parameters of, 167, 276
  - for prediction, 168–183
  - background information, 168–169
  - local piecewise model structures for regression, 174–175
  - nonparametric “memory-based” local models, 175–178
  - regression models with linear structure, 169–173
  - selecting, of appropriate complexity, 183
  - stochastic components of, 178–180
- for probability distributions and density, 184–193
  - background information, 184
  - concepts, general, 184–185
  - factorization and independence in high dimensions, 187–193
  - joint distributions for unordered categorical data, 187
  - mixtures of, 185–187
- search methods for, 238–241, 378–381
  - background information, 238–241
  - branch-and-bound, 246–247
  - heuristic search, 244–246
  - simple greedy search algorithm, 243–244
  - state-space formulation, 241–243
  - systematic search, 244–246
- for structured data, 197–203
- Momentum-based methods, 254
- Monothetic divisive methods, 315
- Monotonic regression, 87
- Monte Carlo Markov Chain (MCMC)
  - methods, 123, 268
- Monte Carlo sampling techniques, 123, 226
- Morse codes, 85
- MSE, 107, 223–224
- Multicollinearity, 371

- Multidimensional indexing, 404–405
- Multidimensional scaling, 84–90
- Multidimensional scaling plot, 88
- Multilayer perceptrons (MLPs), 153–157, 357, 391
- Multimodal distributions, 56, 60
- Multinomial distributions, 487–488
- Multiple regression, 368–369
- Multirelational data, 42–43
- Multivariate function, 113–114
- Multivariate gradient descent method, 256
- Multivariate normal distributions, 490
- Multivariate parameter optimization, 255–259
- Multivariate random variables, 97–102
  
- Naive Bayes model, 191, 353–356
- NASA Earth Observing System, 19
- Natural language processing (NLP), 457
- Natural parameter, 388
- Nearest neighbor methods
  - agglomerative methods and, 312–313
  - condensed, 352
  - edited, 352
  - nonparametric “memory-based” local models and, 176, 178
  - pairwise distances of the members of each cluster and, 312–313
  - parametric models and, 351
  - predictive models for classification and, 347–352
  - reduced, 352
- Nelder and Mead variant, 259
- Nesting, 172
- Neural networks, 173
- Newton-Raphson (NR) method, 252–253, 255, 389
- Newton’s method, 256–257
  
- NIST, 456
- NLP, 457
- Nominal scales, 28, 31
- Non-metric multidimensional scaling, 87
- Nonlinear function, 10, 154
- Nonlinear global models, 478–479
- Nonparametric density estimation, 284–287
- Nonparametric “memory-based” local models, 175–178
- Nonparametric models, 185
- Nonparametric test, 130
- Nonprobabilistic descriptive models, 219
- Nonrepresentational procedures, 30
- Nonscalable versions of data mining algorithms, 424
- Nonsystematic variation, 179–180
- Normal density, 197, 355
- Normal distribution, 60, 113, 115–116, 118, 121–122, 127, 171, 276, 350, 488
- Normal posterior, 122–123
- Normal prior, 122–123
- NR method, 252–253, 255, 389
- Null hypothesis, 124–126
- Numerical scales, 31
  
- Objects, 4
- Observational data, 1
- Odds ratio, 352–353
- OLAP, 417–419
- OLTP, 417–419
- One-tailed test, 125
- Online algorithms, 265–266
- Online analytical processing (OLAP), 417–419
- Online approximation, 265
- Online transaction processing (OLTP), 417–419
- Operational databases, 417

- Operational procedures, 30
- Opportunity samples, 21, 48
- Optimization
  - background information, 235–238
  - combinatorial, 236, 239
  - as component of data mining algorithms, 16–17, 142–143
  - constrained, 259–260
  - Expectation Maximization algorithm and, 260–265
  - maximum likelihood estimation and, 114
  - with missing data, 260–265
  - online algorithm and, 265–266
  - parameter optimization methods, 247–260
    - background information, 247–249
    - closed form, 249–250
    - constrained, 259–260
    - gradient-based, 250–251
    - linear algebra, 249–250
    - multivariate, 255–259
    - univariate, 251–255
  - query, 415–417
  - single-scan algorithms and, 265–266
  - stochastic, 266–268
- Ordinal scales, 28, 31
- Organization of data. *See* Databases
- Orthogonality of data, 240
- “Out-of-sample” data, 227, 328, 372
- Overestimation, 216
- Overfitting, 19, 183, 223
  
- $p$ -dimensional space, 10, 12, 165, 180, 277, 479
- $p$ -dimensional vector, 9, 36, 174, 329–330, 399
- PageRank, 15
- Pairs of measurements, 327
- Pairwise distance, 312
  
- Parallel coordinates plots, 74, 76
- Parameter optimization methods
  - background information, 247–249
  - closed form, 249–250
  - constrained, 259–260
  - gradient-based, 250–251
  - linear algebra, 249–250
  - multivariate, 255–259
  - univariate, 251–255
- Parameters
  - algorithm, 267
  - canonical, 388
  - defined, 47
  - dispersion, 388
  - estimation, 240
  - linear function of, 9
  - location, 184
  - of models, 167, 276
  - natural, 388
  - regression model, 173
  - scale, 184, 388
- Parametric models
  - density, 275–279
  - mixtures of, 185–187
  - nearest neighbor methods and, 351
  - overview, 184
- Parents of variables, 189
- Partition-based clustering algorithms
  - background information, 296–297
  - basic algorithms for, 302–308
  - for nonprobabilistic descriptive models, 219
  - score functions for, 296–302
- Pattern search, 259
- Patterns. *See also* Models
  - background information, 165–167
  - class of, 204
  - co-occurrence, 158
  - coverage of, 214
  - in data matrices, 203–206
  - data mining and, 1–2, 10–11, 271
  - defined, 165

- detection of, 102
- discovering, 13–14, 438–441
- discovery task, 205
- evaluation of, 229–231
- finding, 427–448
  - association rules, 433–435
  - background information, 427–428
  - episodes from sequences, 436–438
  - from local patterns to global models, 442–443
  - generalizations, 435–436
  - itemsets, frequent, 429–433
  - predictive rule induction and, 443–447
  - rule representations, 428–429
  - selective discovery, 438–441
- global, 9
- local, to global models, 442–443
- primitive, 204
- Q, 450, 454
- scoring, 212–215
- search methods for, 238–241, 378–381
  - background information, 238–241
  - branch-and-bound, 246–247
  - heuristic search, 241, 244–246
  - simple greedy search algorithm, 243–244
  - state-space formulation, 241–243
  - systematic search, 244–246
- for strings, 206–208
- structure of, 158
- structures, 203–208
  - in data matrices, 203–206
  - for strings, 206–208
- text retrieval, 14
- PCA. *See* Principal components analysis
- Penalized likelihood, 321–322
- Percentiles, 56
- Perceptrons, 153–157, 339–341, 357, 391
- Permutation tests, 129
- Piecewise model structures for regression, 174–175, 182
- Point estimates, 115, 119
- Poisson distributions, 280–281, 388, 488
- Poisson regression, 388
- Polysemy, 457
- Polythetic divisive methods, 315
- Population drift, 49
- Position, sequential, 477
- Posterior distributions, 117
- Precise functional form, 176
- Precise measurement, 45
- Precision, 121, 453–456
- Predicted intervals, 374–375
- Predictive distributions, 120–121
- Predictive models
  - background information, 168–169
  - for classification, 327–366
    - classification models and, 329–339
    - evaluating and comparing, 359–362
    - feature selection for high-dimension, 362–363
    - linear discriminants and, 341–343
    - logistic discriminant analysis, 352–353
    - naive Bayes model, 353–356
    - nearest neighbor methods, 347–352
    - other methods, 356–359
    - overview, 180–182, 327–329
    - perceptrons and, 339–341
    - tree models, 343–347
  - examples of, 14
  - goal of, 13



- local piecewise model structures for regression, 174–175
- nonparametric “memory-based” local models, 175–178
- for regression, 367–398
  - artificial neural networks, 391–393
  - background information, 367–368
  - generalized linear models, 384–390
  - least squares fitting, 368–384
  - linear models, 368–384
  - other highly parameterized models, 393–397
- regression models with linear structure, 169–173
- score functions for, 212, 215–217
- selecting, of appropriate complexity, 183
- stochastic components of, 178–180
- Predictive performance, 196
- Predictive rule induction, 443–447
- Predictor variables, 168, 367
- PREFERENCE property, 27
- Preferences, modeling individual, 470–472
- PRIM algorithms, 445–446
- Primitive patterns, 204
- Principal components, 195
- Principal components analysis (PCA)
  - graphical data exploration methods and, 74–84
  - high-dimensional data and, 196
- Principal coordinates method, 86
- Prior distributions, 117
- Priors, 122–123
- Probabilistic model-based clustering
  - using mixture models
    - advantages, 319
    - background information, 315–316
    - disadvantages, 319–321
    - examples, 316–319
    - techniques, 321–323
- Probabilistic models for classification, 331–334
- Probabilistic rule, 213–214, 428
- Probability, 93–97
- Probability calculus, 94–96
- Probability distribution and density function models
  - background information, 184
  - concepts, general, 184–185
  - descriptive models and
    - background information, 272–274
    - Expectation Maximization algorithm for, 281–284
    - joint distributions for categorical data, 287–292
    - mixture distributions and densities, 279–281
    - nonparametric density estimation, 284–287
    - parametric density models, 275–279
    - score functions for, 274–275
  - estimation, 274–275
  - factorization and independence in high dimensions, 187–193
  - joint distributions for unordered categorical data, 187
  - mixtures of, 185–187
- Probability distributions, 485, 487–490
- Probability mass function, 485
- Probability theory, 94–95
- Projection operation, 411
- Projection pursuit methods, 77, 195–196, 357, 395–397
- Proximity, 32
- Pruning, 153, 159
- Pseudo data sets, 425–426
- QBIC, 15, 474

- Quadratic discriminant function, 343
- Quadratic error function, 340
- Quadratic function, 249
- Quadratic programming, 259
- Quality of data
  - for collection of data, 47–51
  - for individual measurements, 44–47
  - poor, 51
- QUALITY OF LIFE property, 29
- Quantitative variables, 6
- Quartiles, 56
- Quasi-likelihood methods, 180
- Quasi-Newton methods, 257–258
- Quasilikelihood estimation, 390
- Query
  - aggregation in, 414
  - execution, 415–417
  - image, 473–474
  - matching, 461–465
  - optimization, 415–417
  - pattern Q, 450, 454
  - rectangular range, 404
  - selectivity estimation, 273
  - Structured Query Language, 409, 413–415
  - text, 456–457
- Query by Image Content (QBIC), 15, 474
- Radial basis function networks, 393
- RAM, 17
- Random samples, 20, 54, 123
- Random variables, 97–102, 485–490
- Random variation, 179–180
- Random-access memory (RAM), 17
- Randomization tests, 129
- Randomness, 93–97. *See also* Uncertainty
- Range, 56, 404
- Ranks, 31
- Ratio scales, 28
- Recall, 453–456
- Receiver Operating Characteristic (ROC) curve, 361, 454
- Reciprocals of variances, 121
- Records, 4
- Rectangular range query, 404
- Reduced nearest neighbor methods, 352
- Reductionist viewpoint on data mining algorithms
  - A Priori algorithm, 157–160
  - background information, 151–153
  - multilayer perceptrons for regression and classification and, 153–157
  - vector-space for text retrieval and, 160–162
- Redundant variables, 194
- Reference prior, 122
- Regression
  - approach, 335–339
  - defined, 169, 328–329
  - estimation and, 13
  - line, 368
  - linear, probabilistic interpretation of, 372–375
  - local piecewise model structures for, 174–175
  - locally weighted model, 175–176
  - loess model, 175–176
  - logistic, 384–385
  - methods, 348
  - models with linear structure, 169–173
  - monotonic, 87
  - multilayer perceptrons for, 153–157
  - multiple, 368–369
  - plane, 368–369
  - Poisson, 388
  - predictive models for, 367–398
    - artificial neural networks, 391–393

- background information, 367–368
- generalized linear models, 384–390
- least squares fitting, 368–384
- linear models, 368–384
- other highly parameterized models, 393–397
- projection pursuit, 195–197, 395–397
- rule-based, 446
- simple, 368
- sum of squares, 376
- Regular expression E, 207
- Regular grammars, 202
- Regularities, 134
- Regularized discriminant analysis, 343
- Reject option, 350
- Rejection region, 125
- Relation schema, 405
- Relational algebra, 409
- Relational data model, 405
- Relational databases, 405–409
- Relations, 405
- Relative distributions, 459
- Relevance feedback, 462, 470–471
- Reliability of measurements, 46
- Repeated measures data, 349–350
- Representational measurements, 29–31
- Resampling techniques, 322
- Residual sum of squares, 376
- Residuals, 369
- Response variable, 168, 367
- Resubstitution error rate, 359
- Retesting, effective, 46
- Retrieval by content
  - applications of, 15
  - background information, 449–452
  - evaluation of systems, 452–456
  - goal of, 14
  - image retrieval, 472–476
  - sequence retrieval, 476–481
  - summary information, 481–482
  - for text, 456–470
    - background information, 456–457
    - classification of document and text, 469–470
    - latent semantic indexing, 465–469
    - matching queries and documents, 461–465
    - patterns, 14
    - representation of text, 457–461
  - time series, 476–481
- Right-skewed distributions, 57
- Risk of error, 45
- Robust methods, 231–232
- ROC curve, 361, 454
- Rocchio's algorithm, 470
- Root node, 244–245
- Rotations, random, 71
- Rothamsted Experimental Station, 11–12
- Rows, 36
- Rules
  - discovering, 13–14, 438–441
  - finding
    - association rules, 433–435
    - background information, 427–428
    - episodes from sequences and, 436–438
    - from local patterns to global models, 442–443
    - generalizations, 435–436
    - itemsets, frequent, 429–433
    - predictive rule induction and, 443–447
    - rule representations, 428–429
    - selective discovery of, 438–441
  - probabilistic, 213–214, 428
  - regression based on, 446

- representations of, 428–429
  - set of, 443
  - structure of, 158
- Sample correlation coefficient, 35
- Sample covariance, 35
- Sample mean, 33, 55
- Sample-based estimate of sample mean, 55
- Samples, 7. *See also* Data set
- convenience, 21, 48
  - data mining and, 93
  - distortion of, 49–50
  - epsem, 134
  - “lower resolution” data, 11
  - opportunity, 21, 48
  - random, 20, 54, 123
  - systematic, 133–134
  - uncertainty and, 102–105
- Sampling fraction, 133
- Sampling methods, 132–138, 338
- Sampling paradigm, 128
- Scalable versions of data mining algorithms, 423–424
- Scale parameter, 184, 388
- Scales, 28–29, 31
- Scatterplot matrix, 71–72
- Scatterplots, 64–65
- Schemas, 41–44, 405, 410
- Score functions for data mining algorithms
- background information, 211–212
  - decomposable, 240
  - defined, 211, 235
  - descriptive, 212, 217–219
  - with different complexities, 220–228
    - bias-variance, 221–224
    - concept in comparing, general, 220–221
    - penalizing, 224–227
    - validation and, external, 227–228
  - evaluating, 229–231
  - function of, 142
  - generic, 16, 219
  - for partition-based clustering algorithms, 296–302
  - patterns, scoring, 212–215
  - predictive, 212, 215–217
  - for probability distribution and density function models, estimating, 274–275
  - robust methods, 231–232
  - scoring method versus, 389
- Scoring method
- complexity of a model and, 220–228
    - bias-variance, 221–224
    - concepts in comparing, general, 220–221
    - penalizing, 224–227
    - validation and, external, 227–228
  - score functions versus, 389
- Scree plots, 79–80
- Search methods
- background information, 235–238
  - blind, 245–246
  - branch-and-bound, 246–247
  - breadth-first, 245
  - as component of data mining algorithms, 16–17, 142–143
  - depth-first, 245
  - genetic, 266–267
  - greedy heuristic, 241
  - heuristic, 241, 244–246, 439–440
  - for models and patterns, 238–241, 378–381
  - simple greedy search algorithm, 243–244
  - state-space formulation, 241–243
  - stochastic, 266–268
  - systematic, 244–246
- Search operators, 241–242
- Search tree, 244–245, 402
- Segmentation, 12, 293

- Select clause, 413
- Selection operation, 411
- Selectivity, 273
- Sequence retrieval, 476–481
- Sequences, episodes from, 436–438
- Sequential data, 477
- Sequential position, 477
- Set operations, 410
- Set of rules, 443
- SEVERITY property, 27
- Severity scale, 28
- Significance level, 105, 125
- Similarity, 15, 449, 451, 480
- Simple greedy search algorithm, 243–244
- Simple regression models, 368
- Simplex algorithm, 258
- Simplex search method, 258
- Simpson's paradox, 100–101
- Simulated annealing, 267–268
- Simultaneous test procedures, 131
- Single link method, 312–313
- Single-link criterion, 298
- Single-scan algorithm, 265
- Singular-value decomposition (SVD), 415, 466
- Skewness, 56–57
- SKICAT system, 13
- Sloan Digital Sky Survey, 19
- Snooping, 22–23
- Spatial data, 44
- Special-purpose algorithms for disk access, 424
- Spline function, 174
- Splines, 174–175
- Splitting a node, 344–345
- SQL, 409, 413–415
- Squared error, 216
- SRM approach, 226
- SSE, 155–156, 235
- Standard data, 41
- Standard deviation, 56, 60
- Standardization, 38
- Star icons, 74
- Star plot, 75
- State space representation, 241
- State variables, 200–201
- State-space formulation for search methods, 241–243
- Stationarity, 198–199
- Statistical inference, 102–105
- Statistics, 18–21, 47, 425–426
- Stepwise model, 130
- Stochastic approximation, 265
- Stochastic components of model structures, 178–180
- Stochastic estimation, 123, 265
- Stochastic search methods, 266–268
- Strategic databases, 417
- Stratified random sampling, 135
- Strings, 43, 206–208, 420–421
- Structural risk minimization (SRM) approach, 226
- Structured data models, 197–203
- Structured Query Language (SQL), 409, 413–415
- Structures, data mining, 9–11, 142
- Student's *t*-distributions, 489
- Subsamples, 360
- Subsets problem, 241
- "Sufficient statistic" concept, 112–113
- Sufficient statistics, 19–20, 425–426
- Suffix tree data structure, 421
- Sum of squared errors (SSE), 155–156, 235
- Sum of squared residuals, 376
- Summarizing data, 54–57
- Supervised classification, 169, 328–329
- Support, 430
- Support vector machines, 357
- Surrogate document, 461
- Suspect data, 50–51
- SVD, 415, 466

- Synonymy, 457
- Systematic sampling, 133–134
- Systematic search methods, 244–246
- Systematic variation, 179
- T*-dimensional “term space,” 461
- Tables, 41, 188, 408–412
- Tasks, data mining, 11–15, 142
- Taylor series, 227, 257, 369
- Temperature schedule, 267
- Ten-fold cross-validation, 322
- Term, 456
- Term frequency (TF), 463
- Test set, 360
- Text retrieval
  - background information, 456–457
  - classification of document and text, 469–470
  - latent semantic indexing, 465–469
  - matching queries and documents, 461–465
  - patterns, 14
  - representation of text, 457–461
- Text retrieval Conferences (TREC), 456
- TF, 463
- Time series data, 476–481
- Total sum of squares, 376
- Training data, 7. *See also* Data set
- Training data points, 346
- Transactions, 405–406
- Transforming data, 38–41, 195–196, 363
- TREC, 456
- Tree models, 174, 343–347
- Tree-structured rule sets, 443
- Trellis plotting, 71, 73–74
- Trimmed mean, 231–232
- True error rate, 359
- True value concept, 45
- Tuple, algorithm, 146, 151, 154–155
- Unbiased estimation, 106, 227
- Uncertainty
  - background information, 93
  - dealing with, 94–97
  - estimation and, 105–124
    - background information, 105–106
    - Bayesian, 93, 116–124
    - maximum likelihood and, 93, 108–116
    - properties of estimators and, desirable, 106–108
    - stochastic, 123, 265
  - hypothesis testing and, 124–132
    - background information, 124
    - classical, 124–130
    - in context, 130–132
  - multivariate random variables and, 97–102
  - probability and, 93–97
  - random variables and, 97–102
  - samples and, 102–105
  - sampling method and, 132–138
  - statistical inference and, 102–105
  - summary information, 138
- Underestimation, 216
- Unexplainable variation, 179–180
- Union operation, 410
- Univariate parameter optimization, 251–255
- Univariate random variables, 485–487
- Universal table, 408
- Unordered categorical data, joint distributions for, 187
- U.S. National Institute of Standards and Technology (NIST), 456
- Validation, 227–228
- Validation log-likelihood, 275
- Validation subset, 148–149
- Validity of measurements, 46–47
- Variability measurements, 56
- Variables

- categorical, 6
- class, 329
- conditionally independent, 99–100, 289, 354
- defined, 4
- dependent, 35
- derived, 198
- explanatory, 168
- frequent sets of, 204
- hidden, 187, 190–191, 195
- independent, 99, 188–189
- individual X, 194–195
- input, 329
- latent, 187, 190–191, 195
- length, 32
- linear dependencies between, 35
- manipulating, 168
- multivariate, 97–102
- parents of, 189
- predictor, 168, 367
- quantitative, 6
- random, 97–102, 485–490
- redundant, 194
- response, 168, 367
- selecting, 362–363
- selection for high-dimensional data, 194–195
- state, 200–201
- transforming, 363
- univariate random, 485–487
- visual techniques for displaying
  - more than two, 70–74
  - relationships between two, 62–70
  - single, 57–62
- weight, 32
- Variance function, 388
- Variances, 56, 78, 121, 221–224
- Variations, 297–298
- Vector space representation, 458
- Vector-space algorithms, 160–162
- Visual techniques
  - data mining and, 11
  - for more than two variables, 70–74
  - for relationships between two variables, 62–70
  - for single variables, 57–62
- Warehousing, data, 417–419
- WEIGHT property, 26–28
- Weight variables, 32
- Weighted Euclidean distance, 33
- Weighted least squares solution, 382
- Where clause, 413
- Whole image, 166
- Wilcoxon test statistic, 129–130
- Within-cluster sum-of-squares, 298
- Within-cluster variation, 297–298
- Zero skewness, 57

This excerpt from

Principles of Data Mining.  
David J. Hand, Heikki Mannila and Padhraic Smyth.  
© 2001 The MIT Press.

is provided in screen-viewable form for personal use only by members of MIT CogNet.

Unauthorized use or dissemination of this information is expressly forbidden.

If you have any questions about this material, please contact  
[cognetadmin@cognet.mit.edu](mailto:cognetadmin@cognet.mit.edu).