

INDEX

- L_2 norm, 188
- L_∞ norm, 188
- L_k norm, 188
- $O(\cdot)$, *see* big oh
- $\delta(\cdot)$, *see* Dirac delta
- \emptyset , *see* “don’t care” symbol
- $\lambda(\cdot, \cdot)$, *see* loss
- \mathbf{R}^d , *see* Euclidean space
- † , *see* matrix, pseudoinverse
- ω , 20, *see* state of nature
- ρ , *see* correlation, coefficient
- d -tuple, *see* tuple
- d' , 49, *see* discriminability
- k -nearest neighbor, 163, 174–177
 - Bayes relation, 177–178
- $\Theta(\cdot)$, *see* big theta
- $\boldsymbol{\theta}$, *see* vector, parameter
- \mathcal{A} , *see* alphabet
- $\mathcal{E}[\cdot]$, *see* expectation
- \mathcal{R}_i , *see* decision, region
- 20-questions, 395

- abstract computer, *see* computer, abstract
- abstract encoding of a string, 462
- accuracy-complexity tradeoff, 186
- action, 24
- activation function, 173, 285, 307
 - continuity, 307
 - linearity, 307
 - piecewise linear, 307
- active learning, *see* learning, active
- AdaBoost, 478
- adenine, 413
- adjoint, *see* matrix, adjoint
- agglomerative clustering, *see* clustering,
 - agglomerative
- Agglomerative hierarchical clustering
 - Algorithm*, 552
- algorithmic complexity, 462–463

- algorithmic entropy, *see* algorithmic
 - complexity
- alphabet, 415
- analysis by synthesis, 7
- annealing
 - deterministic
 - Algorithm*, 359
 - mean-field, *see* simulated annealing,
 - deterministic, *see* annealing, deterministic
 - schedule, 355
- anti-cross validation, *see* cross validation,
 - anti
- anticipatory coarticulation, *see*
 - coarticulation, anticipatory
- arcng, 475
- ART, *see* clustering, adaptive resonance
- artificial intelligence, 431
- ascender (character), 375
- asymptotic lower bound, *see* lower bound,
 - asymptotic
- asymptotic notation, 633
- asymptotic tight bound, *see* tight bound,
 - asymptotic
- asymptotic upper bound, *see* upper bound,
 - asymptotic
- augmented vector, 285
- auto-associator, 569
- auto-encoder, 568
 - network, 582
- average, *see* mean, *see also* expected value

- B-type unorganized machine, 333
- backpropagation, 283, 288–293
 - Algorithm*
 - stochastic, 294
 - stochastic with momentum, 314
 - and neurobiology, 285
 - Bayes Theory, 303

- backpropagation (*cont.*)
 - biological plausibility, 334
 - feature mapping, 299
- bad-character heuristic, 416
- bagging, 475–476
- Baldwin effect, 382
- Basic minimum-squared-error clustering
 - Algorithm*, 549
- batch training, 228
- Baum-Welch *Algorithm*, *see*
 - forward-backward *Algorithm*
- Bayes
 - decision rule, 23
 - density estimation, 97
 - formula, 22, 24, 91, 303, 463, 531
 - maximum-likelihood comparison, *see*
 - maximum-likelihood, Bayes comparison
 - nearest-neighbor relation, 178
- Bayes error
 - dependence on number of features, 110
- Bayes estimation
 - maximum-likelihood comparison, 100
- Bayes rule, 615, 616, 620
 - model, 486
 - vector, 617
- Bayesian
 - learning, *see* learning, Bayesian
- Bayesian belief networks, *see* Belief
 - networks
- Bayesian decision theory, *see* decision
 - theory, Bayesian
- Bayesian estimation, *see* learning,
 - Bayesian
 - Gaussian
 - multidimensional, 97
- beanbag chair
 - example, 8
- BEATS example, *see* subset/superset
- BEL(\cdot), *see* belief, function
- belief
 - function, 58
- belief net, 56
 - node, 57
- Bernoulli, *see* distribution, Bernoulli
- Berry paradox, 504
- Bertrand's paradox, 507
- Beta, *see* distribution, Beta
- between-class scatter, *see* scatter,
 - between-class
- Bhattacharyya
 - bound, 47
 - coefficient (ρ), 75
- bias, 37, 53, 466
 - boundary, 469
 - estimator, 473
 - preference, 464
- bias (offset), 216
- bias and variance, 465–471
 - classification, 468
 - regression, 466
- bias unit, 285
- Bienaymé-Chebyshev, *see* Chebyshev's
 - inequality
- big oh, 633
 - nonuniqueness, 111
 - notation, 111
- big omega, 633
- big theta, 111, 633
- binary feature, *see* feature, binary
- Binomial, *see* distribution, Binomial
- binomial coefficient, 78
- bit, 32, 400, 402
- blind source separation, 570, 571
- Boltzmann
 - chain, 370
 - constant, 353
 - factor, 353
 - zipper, 372
- Boltzmann learning, *see also* learning,
 - Boltzmann, 360–373
 - application, 382
 - deterministic
 - Algorithm*, 367
 - network, 455
 - stochastic, 360–366
 - weight initialization, 368
- boosting, 476–480
- bootstrap, 474–475
 - aggregation, *see* bagging
 - bias estimate, 475
 - classifier accuracy, 485
 - variance estimate, 475
- bound
 - Bhattacharyya, *see* Bhattacharyya,
 - bound
 - leave-one-out, 263
- boundary, *see* decision, boundary
- boundary bias, *see* bias, boundary
- boundary error, *see* error, boundary
- branch, 395
- branching factor, 397
 - ID3, 411
- branching ratio, *see* branching factor
- bump
 - network response, 287
- C4.5, *see* classifier, tree, C4.5
- camera
 - for pattern recognition, 2

- canonical variate, *see* Fisher linear discriminant
- capacity
 - hyperplane, 495
- cardinality, 58
- cascade-correlation, 329, 335
 - Algorithm*, 330
- category symbol (ω), 20
- category unit
 - PNN, 172
- Cauchy distribution, *see* distribution, Cauchy
- Cauchy-Schwarz inequality, 606, 614
 - vector analog, 618
- causal network, *see* belief network
- Central Limit Theorem, 33, 621
- chain rule, *see* derivative, chain rule
- character, 413
- charge
 - physical, 453
- Chebyshev's inequality, 612
- Chernoff bound, 47
- chi-square, 629–630
 - statistic, *see* statistic, chi-squared, 629
 - table, 630
- child (belief net), 58
- Chomsky normal form, 425
- chromosome, 373
- clamp, 360, 362
- class
 - independence, 85, 91
- class-conditional probability, *see* probability, class-conditional
- classification, *see* pattern recognition, 12
 - cost, 3
 - fish example, 20
 - food, 517
 - model, 3
- Classification and regression tree, *see* classifier, tree, CART
- classifier
 - Bayes, 91, 531
 - unsupervised, 530
 - combinations, 495–500
 - complex, 461
 - component, 476
 - expert, 454
 - Gaussian, 220
 - linear, 53
 - margin, 224
 - polynomial, 307
 - representation, 375
 - simple, 215, 453, 461
 - stability, *see* stability, classifier
- tree
 - binary, 397
 - C4.5, 411
 - CART, 396–412
 - computational complexity, 406
 - ID3, 411
- cluster
 - chaining, 554
 - criterion
 - chi-squared, 557
 - determinant, 545
 - invariant, 546
 - Kolmogorov-Smirnov, 557
 - local minimum, 548
 - trace, 545
 - dendrogram, 551
 - diameter
 - path, 566
 - level, 551
 - mean, 549
 - membership fuzzy, 528
 - similarity measure, 551
 - singleton, 549
 - tree
 - minimal spanning, 567
 - validity, 551, 581
- clustering, *see* learning, unsupervised, 17
 - Adaptive Resonance, 582, 563–565
 - agglomerative, 552–556
 - Bayesian, 581
 - bottom-up, *see* clustering, agglomerative
 - chaining, 582
 - complete-linkage, 554, 566
 - criterion
 - sum-of-squared error, 542, 581
 - trace, 581
 - criterion function, 542
 - decision directed, 536–537
 - divisive, 552
 - farthest-neighbor, 554–555
 - fuzzy k -means, 528–530, 581
 - gradient ascent, 548
 - graph theoretic, 566–567, 582
 - hierarchical, 550–551, 581
 - agglomerative, 581
 - divisive, 581
 - stepwise-optimal, 555–556
 - hypothesis
 - null, 557
 - iterative algorithms, 581
 - k -means, 526–528, 581
 - leader-follower, 561–565, 582
 - maximum algorithm, 554
 - maximum-likelihood, 581
 - minimum algorithm, 553

- clustering (*cont.*)
 - motivations, 517
 - nearest-neighbor, 553–554
 - nonparametric method, 537
 - optimization
 - iterative, 548
 - single-linkage, 553, 566
 - small sample, 532
 - solution
 - unique, 519
 - splitting, *see* clustering, divisive
 - starting point, 550
- coarticulation
 - anticipatory, 10
- coefficient
 - Bhattacharyya, *see* Bhattacharyya, coefficient
- cofactor
 - matrix, *see* matrix, cofactor
- competitive learning, 559–561, 582
 - Algorithm*, 560
- complexity
 - classifier, 265
 - computational, 16, 111–114, 633–634
 - maximum-likelihood classifier, 111
 - descriptive, *see* algorithmic complexity
 - exponential, 112
 - Kolmogorov, *see* algorithmic complexity
 - polynomial, 112
 - space, 112, 634
 - time, 112, 634
- component analysis, 568–573, 582
- component classifier, *see* classifier, component
- compression technique, 462
- computational complexity, *see* complexity, computational
 - accuracy tradeoff, *see* accuracy, computational complexity tradeoff
 - and feature dimensions, 16
 - and number of categories, 16
 - and number of patterns, 16
 - of estimation, 100
- computer
 - abstract, 462
- Concept Learning System, 435
- conditional independence, 52
- conditional probability, *see* probability, conditional
- conditional risk, *see* risk, conditional
- confidence based query selection, *see* query, selection, confidence based
- confidence level, 403, 630
- configuration, 351
 - hidden, 361
- conjugate gradient, 321
- conjugate prior, *see* prior, conjugate
- conjunction rule, 192
- connected component, 566
- conservation law, 453
- consistency, 505
- constant (logical), 432
- constraint
 - imposition, 365
 - risk, 29
- context, 13
 - statistical dependence, 62
- convergence
 - Ho-Kashyap, 251
- convolution, 621
 - Parzen, 167
- cooling schedule, *see* annealing, schedule
- correct rejection, *see* rejection, correct
- correlation, 362
 - coefficient, 614, 627, 628
 - spurious, 362
 - unit, 367
- cost, 455, *see* error
- cost matrix, 409
- covariance, 34, 613, 618
 - matrix, *see* matrix, covariance
 - sample, 90
 - normalized, 614
 - of sum distribution, 97
- covariance matrix, *see* matrix, covariance
- Cox axiom, 194
- credit assignment problem, 289, 291
- criterion
 - Neyman-Pearson, *see* Neyman-Pearson criterion
 - scatter, 544
 - stopping, *see* stopping criterion
 - sum squared error, 542
- criterion function
 - minimizing, 215
 - second-order expansion, 225
- cross entropy, *see* entropy, cross, *see* distance, Kullback-Leibler
- cross moment, *see* covariance
- cross validation, 296, 454
 - anti, 484
 - m*-fold, 483
- crossover, 378
- curse of dimensionality, 170, 221, 557
- cytosine, 413

- DAG, *see* directed acyclic graph
- data
 - collection, 14
 - nominal, 413
 - training, 84
- data description, 537
 - flat, 551
 - hierarchical, 551
- data matrix, 580
- data mining, 517
- decision, 24
 - Bayes, 23, 25, 51
 - binary features, 52
 - bias, 49
 - boundary, 4, 5, 30
 - complex, 5
 - hyperquadric, 41
 - simple, 7
 - compound, 63
 - missing feature, 54
 - noisy feature, 54
 - primary, 409
 - randomized, 68
 - region, 30
 - rule, *see* rule
 - sequential, 63
- decision boundary
 - hyperplane, 217
 - bias, 218
 - negative side, 217
 - positive side, 217
 - linear
 - convex, 219
 - smooth, 453
- decision function
 - direct estimation, 161
- decision surface
 - hyperellipsoid, 220
 - hyperhyperboloid, 220
 - hypersphere, 220
- decision theory, 4
 - Bayes, 20, 24
 - discrete features, 51
 - continuous features, 24
- decision tree, 395–412, 455
 - interpretability, 396
- deficient pattern, *see* pattern, deficient
- deformations
 - non-rigid, 12
- degree of freedom, 403
 - network, 311
- dendrogram, *see* cluster, dendrogram, 581
- density
 - class-conditional, 92
 - estimation, 84
- component, 518
- estimation, 92
 - convergence, 163
 - divergence, 163
 - nonparametric, 161
 - Parzen, 164–174
 - sequence, 98
- function
 - average, 162
- Gaussian, 84
 - bivariate, 626
 - conditional mean, 628
 - marginal, 627
 - mean, 621
 - univariate, 621
 - variance, 621
- joint, 520
 - estimate, 92
 - singular, 627
- mixture, 518, 533
- multimodal, 161
- space average, 163
- unimodal, 161
- density estimation, 9
- derivation tree, 424
- derivative
 - chain rule, 291
- descendent node, *see* node, descendent
- descender (character), 375
- descent procedures
 - table, 260
- design sample, *see* sample, design
- determinant
 - complexity, 111
- deterministic annealing, *see* simulated annealing, deterministic
- dichotomizer, 30
- dimensionality
 - curse of, *see* curse of dimensionality
- Dirac delta, 94, 98
- directed acyclic graph (DAG), 57
- discriminability, 48, 75
- discriminant
 - generalized linear, 220
 - homogeneous linear, 220
 - hyperplane, 282
 - hyperquadratic, 220
 - linear, 283
 - hyperplane, 217
 - multicategory, 218
 - optimal, 215
 - multiple, 121–124
 - quadratic, 220
 - regularized, 114

- discriminant function, 52
 - discrete, 52
 - neural network, 285, 286
- discrimination versus representation, 581
- dissimilarity, 556, 575
 - clustering, 553
- distance, *see also* metric
 - city block, *see* distance, Manhattan
 - edit, 418–420
 - Euclidean, 36, 606
 - Kullback-Leibler, 632
 - Manhattan, 188
 - Mahalanobis, 36, 107, 622, 626
 - tangent
 - two-sided, 205
 - Tanimoto, 188
- distance function
 - as dissimilarity measure, 538
- distributed representation, *see* representation, distributed
- distribution, *see also* probability, distribution
 - and missing data, 54
 - Bernoulli, 109
 - Beta, 109
 - Binomial, 109, 162
 - Cauchy, 148
 - cumulative, 440
 - exponential, 108
 - Gamma, 108
 - Gaussian, 108, 621
 - area, 612
 - covariance, 626
 - eigenvector, 626
 - moment, 624
 - multivariate, 624
 - principal axes, 626
 - univariate, 621
 - identifiable, 100
 - joint, 617
 - marginal, 55, 616
 - maximum entropy, 631
 - Maxwell, 108
 - multinomial, 109
 - normal, *see* distribution, Gaussian, 108
 - Poisson, 78, 109
 - prior, 616
 - Rayleigh, 108
 - separable, 161
 - triangle, 76
 - uniform, 98
- divisive clustering, *see* clustering, divisive
- DNA
 - base pair, 413
 - sequence identification, 1
- dogmatism, 145
- “Don’t-Care” Symbol, 413
- dot product, *see* inner product
- dual functional, 276
- dyadic product, *see* matrix product
- Eddington, Sir Arthur, 381
- edit distance, *see* distance, edit
 - Algorithm*, 418
- eigenfunction expansion, 198
- eigenvalue, 609–610
 - Fisher discriminant, 120
- eigenvector, 609–610
- EM algorithm, *see* Expectation-Maximization
- empty string, *see* string, null
- energy, 351
 - filter, 326
 - interaction, 351
 - landscape, 352, 358
 - physical, 453
- entropy, 32, 339, 462, 630
 - continuous distribution, 631
 - cross, 318
 - discrete, 631
 - for independent component analysis, 582
 - maximum, 501
 - relative, *see* Kullback-Leibler distance, 361, 632
 - surprise, 631
- epoch, 294
- error
 - Bayes, 22, 101, 453
 - Gaussian, 107
 - boundary, 469
 - dependence on number of features, 110
 - estimation, 101
 - generalization, 454
 - indistinguishability, *see* error, Bayes
 - Minkowski, 318
 - model, 90, 101
 - off-training set, 455
 - probability, 21
 - discrete case, 53
 - minimal, 22
 - sum-of-squared, 573
 - test, 169
 - training, 215, 290
 - Parzen window, 168
 - problems, 215
 - validation, 296, 483
- error function, 66, 559, 623
- error rate, 13
 - estimate, 482–486

- error surface, 296–299
 - plateau, 296
- estimate
 - maximum-likelihood clustering, 520
- estimation
 - complexity, 100
 - nonparametric, 215
 - Parzen, 163, 177
 - convergence, 166
 - illustration, 168
- estimation error, *see* error, estimation
- estimator
 - absolutely unbiased, 90
 - asymptotically unbiased, 90
 - consistent, 505
 - unbiased, 90
- Euclidean norm, *see* distance, Euclidean
- Euclidean space (\mathbf{R}^d), 51
- evaluation, 15
- events
 - mutually exclusive, 615
- evidence, 23, 616
 - for model, 486
- evolution
 - strategies, 382
- evolutionary method, 373–380
- exchangeability
 - invariance, 146
- exclusive-OR, 264
- expansion
 - eigenfunction, 198
 - series, 199
 - Taylor series, 198
- expectation
 - continuous, 619
 - entropy, 631
 - linearity, 611, 613
 - vector, 617
- Expectation-Maximization, 55, 124–128
 - Algorithm*, 125
 - Example*, 126
 - generalized, 137
- expected value, 31, *see* mean, 611
 - feature, 32
 - probability, 162
 - two variables, 613
- expert systems, 431
- explanations
 - multiple, *see* indifference, principle
- exponential, *see* distribution, exponential
- expressive power
 - network, 286
- factor, *see* string, factor
- factor analysis, 580
- factorial, 624
- Factorization Theorem, 104, 533
- false alarm, 49
- family (taxonomic), 550
- feature, 1
 - binary, 51
 - choice, 14
 - extraction, 2, 8, 11
 - good (uncorrupted), 55
 - independence, 53
 - independent, 107
 - integer valued, 51
 - missing, 12, 54, 365
 - noisy, 54–63
 - related to error, 110
 - robust, 7
 - space, 4, 24
 - continuous, 5
 - isotropic, 539
 - rescaling, 539
 - ternary, 51
 - transform, 85
 - vector, 4
 - binary, 52
 - continuous, 24
- feature selection, 12
- filter
 - matched, *see* matched filter
- fingerprint identification, 1
- finite state machine, 429
- first-order logic, *see* logic, first-order
- fish
 - categorization example, 1–7, 62
 - occlusion, 54
 - tooth, 394
- Fisher linear discriminant, *see* linear discriminant, Fisher
- fitness, 373, 373, 376
- fittest
 - survival of, 373
- fixed increment, 260
 - training
 - Algorithm*, 230, 233
- flat data description, 550
- Fletcher-Reeves equation, 322
- force
 - magnet, 358
- forward-backward *Algorithm*, 138
- Fourier's Theorem, 287, 335
- free parameter, 283
- frequency ratios, 526
- function, 433
 - Dirac delta, 94, 533, 631

- function (*cont.*)
 - gamma, 624
 - harmonic, 287
 - Kronecker, 605
 - vector valued, 619
 - fuzzy *k*-means clustering
 - Algorithm*, 529
 - game theory, 28
 - Gamma, *see* distribution, Gamma
 - gamma function, *see* function, gamma
 - gap
 - Winnow algorithm, *see* Winnow
 - algorithm, gap
 - gating subsystem, 497
 - Gaussian, *see also* distribution, Gaussian
 - distribution, 64
 - multidimensional, 33
 - table, 623
 - unidimensional, 32, 621
 - univariate, 32, 33
 - Gaussian derivative, 623
 - Gaussian elimination, 111
 - GEM algorithm, *see*
 - expectation-maximization, generalized
 - generalization, 5
 - generalized additive model, 306
 - generalized delta rule, *see* backpropagation
 - generalized expectation maximization, *see*
 - Expectation-Maximization, generalized
 - genetic
 - Algorithm*, 374, 350
 - operator, 374
 - genetic programming, 378–380
 - genome
 - human, 414
 - genus, 550
 - Gibbs Algorithm, 102
 - Gini impurity, 399, 401
 - global error minimum, 296
 - golf course landscape, 352
 - good-suffix function, 417
 - good-suffix heuristic, 416
 - gradient, 606
 - operator, 87
 - gradient descent, 225, *see also*
 - backpropagation, 350
 - Algorithm*, 225
 - convergence, 229
 - fixed increment, 229, 230
- grammar, 7, 421–426
 - free, 424
 - inference, 429
 - learning
 - Algorithm*, 430
 - type 0, *see* grammar, free
 - unrestricted, *see* grammar, free
 - graph
 - similarity, 566
 - graphical models, 370
 - Gray code, 382
 - greedy method, 400
 - greedy search, *see* search, greedy
 - guanine, 413
 - Hamlet*, 381
 - hardware, 1
 - Harmonium, 382
 - Hessian, 488
 - Gaussian approximation, 488
 - singular, 226
 - Hessian matrix, 226, *see also* matrix,
 - Hessian
 - hidden layer, 284
 - hidden Markov model
 - Example*, 133
 - causal, 129
 - computation, 129
 - decoding, 135–136
 - ergodic, 129
 - evaluation, 131–135
 - learning, 137–138
 - Forward-backward *Algorithm*, 138
 - state
 - absorber, 129
 - final, 129
 - hidden unit exchange symmetry, *see*
 - symmetry, hidden unit exchange
 - hidden unit weights
 - representation, 302
 - hidden units
 - number of, 310
 - hierarchical data description, 551
 - Hierarchical dimensionality reduction
 - Algorithm*, 580
 - hint, 315
 - hit, 49
 - HMM
 - decoding
 - Example*, 136
 - left-to-right, 134
 - Ho-Kashyap
 - bound, 255
 - convergence, 255
 - nonseparability, 255
 - separability, 253
 - training
 - Algorithm*, 251

- Hopfield net, 355
- horizon effect, 403
- Horner's rule, 149
- hyperbolic tangent, 308
- hypercube, 53, 164, 357
 - Kolmogorov Theorem, 287
- hyperellipsoid, 41, 537
- hyperparaboloid, 41
- hyperplane, 41, 53
- hyperquadric, 41
- hypersphere, 41, 541
- hypothesis, 455
 - null, *see* null hypothesis
 - testing, 8, 403, 628
 - and clustering, 581
- i.i.d., 85, 162
- ID3, *see* classifier, tree, ID3
- identifiability, 100, 519, 533
 - discrete distribution, 519
- identity matrix, *see* matrix, identity
- idiot Bayes rule, *see* naive Bayes rule
- if-then rule, 431
- image
 - processing, 8
 - threshold, 3
- improper prior, *see* prior, improper
- impulse response, 325
- impurity
 - entropy, 398
 - gain ratio, 401
 - Gini, 399, 401
 - information, *see* impurity, entropy
 - Kolmogorov-Smirnov, 440
 - misclassification, 399
 - scaling, 400
 - statistical significance, 402
 - variance, 399
 - weighted, 409
- incompressibility, 462
- inconsistent edge, 566
- independence
 - class, 91
 - conditional, *see* conditional independence
 - statistical, 34, 62, 613
- independent component analysis, 570–573
- independent features, 107
- independent variables
 - sum, 620
- indifference
 - principle, 489, 500
- indistinguishability error, *see* error, Bayes
- inductive logic programming, 436
- infinite monkey theorem, 381
- information, 339
 - bit, *see* bit
 - divergence, *see* distance, Kullback-Leibler
 - for discrimination, *see* distance, Kullback-Leibler
 - loss, 8
 - mutual, 632–633
- information theory, 462, 630–633
- inner product, 33, 541, 605, 606
- in networks, 285
- PNN
 - pattern unit, 173
- input
 - scaling, *see* standarization, 308
- input layer, 284
- input unit
 - PNN, 172
- insertion, 375, 379
 - operator, 378
- instability, 476
- interchange operation, 418
- International Institute of Advanced Scientific Studies, xix
- interpolation, 9
 - Parzen, 164
- invariance
 - dilation, 541
 - exchangeability, 146
 - illumination, 12
 - line thickness, 12
 - rotation, 541
 - tangent distance method, 188
 - translation, 102, 539
- invariant, 11
- inversion (genetic operator), 375
- jackknife, 472–474
 - bias estimate, 473
 - classifier accuracy, 485
 - variance estimate, 473
- Jacobian, 607
- Jacobian matrix, *see* matrix, Jacobian
- “jeetyet” example, 13
- joining
 - node, *see* merging, node
- joint entropy, 572
- joint probability, *see* probability, joint
- k*-means clustering, *see* clustering, *k*-means
 - Algorithm, 527
- Kalman filtering, 334
- Karhunen-Loève transform, 568

- kernel, *see* Parzen window
- Kesler's construction, 265
 - fixed increment, 266
- keyword, 414
- Kiefer-Wolfowitz, 249
- kingdom, 550
- knowledge
 - incorporating, 7
 - prior, 20, 91
- Kohonen map, *see* self-organizing feature map
- Kolmogorov's Theorem, 287, 335
- Kolmogorov-Smirnov test, 440
- Kronecker delta, 418, 455, 525, *see also*
 - function, Kronecker
- Kuhn-Tucker construction, 263
- Kullback-Leibler distance, 361
 - weighted, 362
 - Gaussian, 143
- Lagrange optimization, *see* optimization, Lagrange
- Lagrange undetermined multiplier, 610
- Lagrangian method
 - network training, *see* network training, Lagrangian method
- language, 423
- last-occurrence function, 416
- Law of Large Numbers
 - Hessian calculation, 488
- Law of Total Probability, 615
- Leader-follower clustering
 - Algorithm*, 562
- leaf node, 395
- learner
 - weak, 476
- learning
 - active, 480
 - algorithm
 - best, 453
 - majority, 502
 - minority, 502
 - and adaptation, 16
 - Bayesian, 85, 98
 - pattern classification, 90
 - Bayesian vs. maximum-likelihood, 536
 - Boltzmann, 350
 - application, 382
 - competitive, *see* competitive learning
 - cost-based, 480
 - curve, 295, 492–494
 - degenerate, 94
 - evolution interaction, 382
 - incremental
 - recursive Bayes, 98
 - nonuniform, 312, 321
 - rate, 291
 - Boltzmann, 361
 - reinforcement, 17
 - supervised, 16, 85
 - supervised vs. unsupervised, 533
 - uniform, 311
 - unsupervised, 17, 85
 - batch protocol, 536
 - Bayesian, 532
 - computational complexity, 533
 - decision-directed, 536
 - with queries, 480–481
- learning component, 362
- learning logic, 334
- learning rate, 225
 - optimal, 312
- least-mean-square
 - Algorithm*, 246, 288, 303
- leave-one-out bound, *see* bound, leave-one-out
- leave-one-out mean, 472
- letter, *see* string
- level curves, 627
- likelihood, 22, 52, 86, 616
 - extremum, 87
 - gradient ascent solution, 522
 - in belief net, 59
 - ratio, 52
 - smoothness assumption, 92
- linear discriminant
 - dual, 259
 - Fisher, 85, 117–124
- linear independence, 606
- matrix columns, 609
- linear machine, 285, 333
 - fixed increment, 266
- linear programming
 - objective, 256
- linear separability, 223, 229, 251
- linear threshold unit, 285, 334
- link
 - tree, *see* branch
- Lisp*, 379
- little omega, 634
- LMS, *see* least-mean-square, 260, 282
 - training
 - Algorithm*, 246
- LMS error
 - projection pursuit, 305
- local minimum, 355
 - high dimensional, 299
- local representation, *see* representation, local

- log-likelihood, 86
 - function, 87
- logic
 - first-order, 432
- logical function, 333
- lookup table, 402
- loss, 24
 - classification, 26
 - expected, 24
 - function, 24
 - matrix, 24, 49
 - minimal, 25
 - symmetric, 26
 - zero-one, 26
- lower bound
 - asymptotic, 633
- LTU, *see* linear threshold unit
- Lyapunov function, 351

- machine perception, *see* perception, machine
- magnet
 - analogy for optimization, 351
- Mahalanobis distance, *see* distance, Mahalanobis
- Manhattan distance, *see* metric, Manhattan
- manufacturing data, 310
- MAP, *see* maximum a posteriori
- margin, *see* classifier, linear, margin
- marginal, 612
 - distribution, 612
- marginalize, 55
- MARS, *see* multivariate adaptive regression
- mass function
 - probability, *see* probability, mass function
- matched filter, 302
- matching
 - template, *see* template matching
- mating, *see* crossover
- matrix
 - addition, 605
 - adjoint, 609
 - antisymmetric, 605
 - covariance, 33, 526, 607
 - complexity, 111
 - determinant, 625, 626
 - diagonal, 526, 618, 619, 625
 - eigenvalues, 618
 - estimates, 90
 - inverse, 625, 626
 - data, *see* data matrix
 - derivative, 606–608
 - determinant, 608–609
 - hypervolume, 608
 - Hessian, 608
 - identity (I), 605
 - inverse
 - derivative, 607
 - inversion, 609–610
 - complexity, 111
 - Jacobian, 607
 - multiplication, 605
 - nonnegative, 605
 - positive semidefinite, 618
 - product, *see* outer product
 - pseudoinverse, 609
 - scatter, 116, 119, 544
 - total, *see also* scatter matrix, total, 544
 - similarity, 566
 - singular, 120
 - skew-symmetric, 605
 - square, 605
 - sweep methods, 111
 - symmetric, 605, 607
 - trace, 144, 609
- maximum-likelihood
 - solution
 - non-uniqueness, 524
- maximum a posteriori (MAP), 87
 - estimator, 87
- maximum algorithm, 559
- maximum entropy, *see* entropy, maximum, 631
- maximum-likelihood, 84, 86, 350
 - Bayes comparison, 85, 100
 - Gaussian
 - mean, 88
 - mean and covariance, 88
- maximum-likelihood
 - solution
 - non-uniqueness, 100
 - singular, 524, 525
 - unsupervised, 520
 - non-uniqueness, 522
- Maxwell distribution, *see* distribution, Maxwell
- MDL, *see* minimum description length, *see* minimum description length
- MDS, *see* multidimensional scaling
- mean, 32, 54, *see also* expected value
 - calculation
 - computational complexity, 633
 - Parzen estimate
 - convergence, 167
 - sample, *see* sample mean, 537
 - trimmed, 475, *see* trimmed mean

- mean (*cont.*)
 - two variables, 613
 - vector
 - total, 121
- mean vector, *see* vector, mean
- median, 472
- Melville, Herman, 414
- memory
 - associative, 8
- mereology, 11
- merging
 - node, 403
 - see also* distance
- metric, 187
 - L_1 norm, 188
 - L_∞ norm, 188
 - L_k norm, 187
 - city block, *see* metric, Manhattan
 - clustering, 538
 - dissimilarity, 557
 - Euclidean, 187, 538
 - induced, 556
 - Mahalanobis, 541
 - Manhattan, 188
 - Minkowski, 187, 541
 - non-negativity, 187, 556
 - properties, 187, 556
 - reflexivity, 187, 556
 - symmetry, 187, 557
 - tangent distance, 188–192
 - Tanimoto, 188, 541
 - triangle inequality, 187, 557
- minimal spanning tree, *see* tree, minimal
 - spanning
- minimax risk, *see* risk, minimax
- minimum algorithm, 553
- minimum description length, 402,
 - 461–464, 501
 - principle, 463–464
 - Bayes relation, 463
- misclassification
 - impurity, *see* impurity, misclassification
- miss, 49
- missing attribute
 - tree, 409
- missing feature, *see* feature, missing
- mixing parameter, *see* parameter, mixing
- mixture
 - density, 518
 - discrete distribution, 519
- mixture model, 496
- mixture of experts, 496
- Moby Dick**, 414
- mode, 78, 472
 - MAP estimation, 87
- model, 2
 - choice, 15
- model error, *see* error, model
- model selection, 486, 489
 - heuristic, 283
 - maximum-likelihood, 486
- Modified Ho-Kashyap
 - training
 - Algorithm*, 254
- moment
 - cross, *see* covariance
 - second, 611
- momentum, 313
 - physical, 453
- monotonicity constraint, 575
- Monte-Carlo, 92
- motif, 414
- MSE algorithm
 - generalizations, 268
- multidimensional scaling, 573–576, 582
- multinomial, *see* distribution, multinomial
- multiple integral, 619
- multivariate adaptive regression, 306
- mutation, 379
- mutual information, *see* information,
 - mutual
- naive Bayes rule, 62
- nat, 32
- nearest-neighbor, 455
 - classifier, 402
 - condensing *see* nearest-neighbor, editing
 - 186
 - convergence, 179
 - editing, 185–186
 - Algorithm*, 186
 - error, 180–182
 - finite sample, 182, 204
 - variance, 181
 - partial distance, 185
 - prestructuring, 185
 - pruning *see* nearest-neighbor, editing
 - 186
- net activation, 173, 285
- network
 - expressive power, 335
- network pruning, 335
- network topology, 283
- neural network, 376
 - chromosome representation, 376
 - feedforward operation, 289
 - learning, 290
 - Probabilistic, 172–174
 - time delay, 326
- neuron, *see* unit

- Newton descent
 - Algorithm*, 226
- Neyman-Pearson criterion, 29
- Niels Bohr Institute, xix
- No Free Lunch Theorem, 454–458
- node
 - belief net, *see* belief net, node
 - child, 58
 - descendent, 395, 403
 - leaf
 - label, 404
 - neighboring, 403
 - parent, 58
 - root, 395
 - terminal, *see* node, leaf
- noise, 12
 - model, 55, 56
- nominal data, 394
- nonlinear component analysis, 569–570
- nonparametric method, 169
- nonuniform learning, *see* learning, nonuniform
- norm, *see* distance or metric
- normal, *see* distribution, normal
- null hypothesis, 8, 628
- null string, *see* string, null
- objective function, *see* Lyapunov function
- Occam
 - factor, 487
 - William of, 6
- Occam's razor, 398, 453, 464–465
- occlusion, 11
- OCR, *see* optical character recognition
- offspring, 373, 374
- omega (ω), *see* state of nature or category symbol
- on-line learning, *see* learning, incremental
- one-of- c representation, 360
- operating characteristic, 48–50
- optical character recognition, 1
 - exhaustive training, 16
 - handwritten, 8
 - rate variation, 12
- Optimal Brain Damage, 332, 335
- Optimal Brain Surgeon, 331, 332, 335
 - Algorithm*, 332
- optimal coding theorem, 464
- optimization
 - iterative, 527, 550
 - Lagrange, 610
- optimization problem, 351
- oracle, 480
- order (taxonomic), 550
- order of a function, 111
- outer product, 120, 606, 618
- outlier, 543, 555
- overdetermined solution, 113
- overfitting, 5, 16, 373, 464
 - and genetic algorithms, 377
- palindrome, 444
- Pandemonium, 333, 381
- parameter
 - estimation, 84
 - mixing, 518
 - space, 92
- parameter estimation
 - Bayesian
 - Gaussian case, 97
 - recursive Bayes, 98
- parent
 - genetic algorithm, 373
- parent (belief net), 58
- parity, *see* exclusive-OR three-bit, 299
- parse
 - table, 426
- parsing
 - bottom-up, 428
 - Algorithm*, 427
- partition
 - minimum variance, 542
- partition function, 353, 361
- Parzen window, 164, 367
 - classification, 168–171
 - estimate, 163
 - window width effect, 165
- prior, 174
- prior, 168
- pattern
 - completion, 382
 - Boltzmann network, 365–366
 - deficient, 365, 409
 - desired, *see* pattern, target
 - target, 289
 - teaching, *see* pattern, target
 - training, 361
- pattern classification, *see* pattern recognition
- pattern recognition, 1
 - general purpose, 7
 - information reduction, 9
 - neural, 7
 - statistical, 6
 - syntactic, 7, 395
- pattern unit
 - PNN, 172
- PCA, *see* principal component analysis
- penalty, 464

- Perceptron, 286, 333, 375, 482
 - batch
 - Algorithm*, 228, 233
 - criterion, 227, 258
 - phi function, 264, 282, *see* activation function, 285
 - phoneme, 10, 517
 - phylum, 550
 - π , algorithmic complexity, 462
 - PNN, *see* neural network, probabilistic
 - space complexity, 174
 - time complexity, 174
 - weight, 172
 - Pocket algorithm, 278
 - Poisson distribution, *see* distribution, Poisson
 - Polak-Ribiere equation, 322
 - poll, 354
 - POLOPONY example, *see* subset/superset
 - polychotomizer, 30
 - polynomial classifier, *see* classifier, polynomial
 - polynomial discriminant, 220
 - population, 373
 - post processing, 13
 - posterior
 - convergence, 100
 - delta function, 100
 - model, 486
 - posterior probability, *see* probability, posterior
 - and backpropagation, 304
 - postpruning, *see* pruning
 - potential function, 195
 - predicate, 432, 458
 - prefix, 416
 - empty, 420
 - preprocessing, 2, 517
 - prepruning, *see* stopped training
 - primal functional, 276
 - primary decision, 409
 - principal axes, *see* axes, principal
 - principal component, 539, 545
 - analysis, 568, 582
 - nonlinear, 582
 - principle of indifference, *see* indifference, principle
 - prior, 20–21, 24, 53, 616
 - conjugate, 93
 - determination, 91
 - estimation, 84
 - improper, 102
 - subjective, 486
 - prior distribution, *see* distribution, prior
 - prior knowledge, 15, 282
 - probabilistic neural network, *see* neural network, probabilistic
 - benefits, 174
 - recall
 - Algorithm*, 173
 - training
 - Algorithm*, 172
 - probability, 51
 - a posteriori, 22
 - a priori, 20
 - class-conditional, 52
 - conditional, 21, 23, 614, 615
 - density, 7, 21, 51, 618
 - estimation, 84
 - joint, 619
 - singular, 51
 - state-conditional, 24
 - distribution, 5
 - joint, 22, 612, 615
 - mass, 615, 618
 - joint, 612
 - mass function, 611
 - posterior, 520
 - nonparametric estimation, 161
 - prior, 22
 - subjective, 193
 - total
 - law, *see* Bayes rule
 - probability theory, 611–623
 - product space, 612
 - production, 423
 - projection pursuit, 305
 - projective distortion, 11
 - property d -tuple, *see* tuple
 - propositional logic, *see* logic, propositional
 - protein, 414
 - prototype, 174
 - pruning, 464
 - Boltzmann network, 382
 - neural network, 330
 - pseudoinverse, 260, 324
 - purity, 398
 - quadratic programming, 264
 - query, 480
 - selection
 - confidence based, 480
 - voting based, 480
 - random guess, 454
 - random variable, *see* variable, random
 - discrete, 611
 - vector, 616–618
 - randomized decision rule, *see* decision rule, randomized

- rank, 459
- rate variation, 12
- Rayleigh distribution, *see* distribution, Rayleigh
- Rayleigh quotient, 120
- RCE, *see* reduced Coulomb energy
- recall, *see* classification
- receiver operating characteristic (ROC), 49
- recognition
 - chair example, 8
- recursive Bayes, 98
 - Example*, 98
 - true, 98
- reduced Coulomb energy, 195
- regression, 9
 - ridge, 114
- regularization, 330–332, 464
- reinforcement learning, *see* learning, reinforcement
- rejection
 - correct, 49
 - option, 24, 69
- relative entropy, 361
- relaxation, 236
 - single-sample, 236
- replication, 378
- representation, 7, 581
 - distributed, 307
 - global, *see* representation, distributed
 - local, 308
- resampling, 454, 471
 - for classifier design, 475
- response function, 358
- Ricoh Silicon Valley, xix
- ridge function, 305
- ridge regression, *see* regression, ridge, 305
- risk, 13, 24
 - conditional, 24, 51
 - minimax, 28
 - overall, 51
 - sample, *see* error, training
- ROC, *see* receiver operating characteristic
- root node, *see* node, root
- root symbol, 422
- rotation, 11
- rule
 - decision, 21
- rule-based methods, 431–434
- saddle point, 524
- saliency, 331
- sample
 - design, 84
 - mean, 88
- sample independence, 532
- San Jose State University, xix
- Santa Fe Institute, xix
- satisficing, 465
- scalar product, *see* inner product
- scale, 11
- scatter
 - between-class, 120
 - within-class, 119, 120
- scatter matrix, 116, *see* matrix, scatter
 - eigenvector, 546
 - invariant, 546
- scatter plot, 4
- score (evolutionary methods), *see* fitness
- search
 - bias, 524
 - exhaustive, 350
 - greedy, 351
 - stochastic, 351
 - tree, 185
- second moment, *see* moment, second
- segment, *see* string, factor
- segmentation, 2, 9, 582
 - speech, 10
- selection
 - and genetic algorithm, 377
 - fitness-proportional, 377
- self-organizing feature map, 576–579, 582
- sensitivity
 - unit, 291
- sensor vector, 570
- sentence, 422
- sequential covering, 433
- set diagram, *see* Venn diagram
- shadow, 12
- Shannon coding theorem, 464
- Sherman-Morrison-Woodbury formula, 151
- shift
 - valid, 415
- shrinkage, 114, 151
- sigmoid, 358
 - parameter, 308
- sign function, 285
- signal detection theory, 48
- significance
 - level, *see* confidence level
 - statistical, 629
 - threshold, 8
- similarity, 394
- similarity function, 541
- similarity graph, *see* graph, similarity
- similarity measure, 538
- simplex algorithm, 256, 261
 - constraint, 258

- simplex algorithm (*cont.*)
 - feasible solution, 257
 - nonseparability, 257
- simulated annealing, 351–360
 - deterministic, 357–360
 - stochastic
 - Algorithm*, 355
- skeleton, *see* tree, spanning, minimal
- “sklee”
 - coarticulation in, 10
- “skloo”
 - coarticulation in, 10
- softmax, 305, 498
- solution region, 224
- SOM, *see* self-organizing feature map
- source separation
 - blind, 582
- space
 - Euclidean (\mathbb{R}^d), 24
- space-time tradeoff, 634
- species, 550
- specification method, 462
- speech recognition
 - rate variation, 12
 - rounding, 10
- spline function, 306
- split, 396, 397
 - surrogate, 409
- stability
 - classifier, 406
- stability-plasticity dilemma, 559
- standard deviation, 612, 622
- standardize, 308, 311
 - data, 338, 539
- Stanford University, xix
- state of nature (ω), 20, 85
- state-conditional probability density, *see*
 - probability density, state
 - conditional
- statistic
 - chi-squared, 403, *see* chi-squared
 - statistic
 - sufficient, 462
- statistical
 - dependence, 62, 614
 - independence, 34, 614, 618
 - expectation, 614
 - noise, 55
 - Gaussian, 627
 - significance, 402, *see also* significance,
 - statistical
 - vector, 616
- statistical mechanics, 350
- Stepwise optimal hierarchical clustering
 - Algorithm*, 555
- stochastic descent, 260
- stopped splitting, 402–403
- stopping criterion, 138, 294, 299
 - and genetic algorithms, 377
- string, 394, 413
 - factor, 413
 - matching, 415–418
 - Algorithm*, 415
 - Boyer-Moore *Algorithm*, 416
 - with “don’t-care” symbol, 421
 - with errors, 420
 - null, 422
 - operations, 413
 - random, 463
- student component, *see* unlearning
- component
- subcluster, 550, 581
- subfamily (taxonomic), 550
- subgraph
 - complete, 554
 - maximal complete, 566
- subjective probability, *see* probability,
 - subjective
- suborder (taxonomic), 550
- subphylum, 550
- subset-superset problem, 418
- substring, *see* string, factor
- subtree, 396
- sufficient statistics, 98, 102–109, 533, 536
 - in unsupervised learning, 537
 - integral nature, 105, 148
- suffix, 416
 - matching, *see* good suffix
- supervised learning, *see* learning,
 - supervised
- support vector, 262
- Support Vector Machines, 259–265
- surprise, 631
- surrogate split, *see* split, surrogate, 412
- survival of the fittest, 373
- symmetry
 - hidden unit exchange, 337
- synapse, 285
- syntactic pattern recognition, *see* pattern
 - recognition, syntactic
- tangent vector, 190
- tanh, *see* hyperbolic tangent
- Tanimoto metric, *see* metric, Tanimoto
- target
 - value, 309
- target propagation, 334
- taxonomy, 550
- Taylor series, 607
- TDNN, *see* neural network, time delay

- teacher component, *see* learning component
- temperature
 - annealing, 352
 - in genetic algorithms, 377
- temperature (randomness), 352
- template matching, 39
 - string, 414
- temporality, 128
- term (logical), 433
- ternary feature, *see* feature, ternary
- text
 - string matching, 413
- threshold, 37, *see* bias
- threshold function, 285
- threshold weight, 53
- thymine, 413
- tight bound
 - asymptotic, 633
- time delay neural network, *see* neural network, time delay
- Tolstoy, Leo, 414
- topologically ordered map, *see* self-organizing feature map
- topology
 - Boltzmann net, 367
 - network, *see also* network topology
- trace, 144, *see* matrix, trace
- trace criterion, *see* cluster criterion, trace
- training, 15
- training data, *see* data, training
 - limited, 163
- training procedure
 - stochastic, 293
- training protocol
 - batch, 293
 - on-line, 293
- transform
 - whitening, 34
- transition probability
 - Markov model, 128
- translation, 11
 - axis, 220
- transpose, 604
- transposition, *see* interchange
- tree, *see* decision tree
 - balanced, 402
 - binary, 397
 - derivation, 424
 - interpretation of, 396
 - minimal spanning, 554
 - missing attribute, 409–410
 - multivariate, 408
 - polythetic, 400
 - priors, 409
 - pruning, 403–404
 - simplicity, 398
 - spanning, 554
 - minimal, 566
 - unbalanced, 403
- tree (graph), 553
- triangle inequality, *see* metric, triangle inequality
- trimmed mean, *see* mean, trimmed, 510
- tuple, 394
- Turing machine, 462
- twiddle, *see* interchange
- two-joint arm
 - self-organizing map example, 576
- twoing criterion, 401
- Ugly Duckling Theorem, 458–461
- ultrametric, 557
- unidentifiable
 - complete, 519
- uniform distribution, *see* distribution, uniform
- uniform learning, *see* learning, uniform
- unit, 285
 - hidden, 330
- universal Turing machine, *see* Turing machine
- unlearning component, 362
- unpredictability, *see* entropy
- unsupervised learning, *see* learning, unsupervised
 - convergence rate, 536
- upper bound
 - asymptotic, 633
- valid shift, 415
- validation
 - error, *see* error, validation
 - set, 295, 483
- validity, 557
- variable
 - random, 62
 - continuous, 618–620
 - discrete, 612
 - standardized, 622, 626
- variable (logical), 432
- variable increment, 260
- variable-free logic, *see* logic, propositional
- variables
 - uncorrelated, 614
- variance, 32, 466, 611
 - addition, 95
 - bias, 89
 - nonlinearity, 612

- variance (*cont.*)
 - Parzen estimate
 - convergence, 167
 - two variables, 613
 - variance impurity, *see* impurity, variance
- vector, 604
 - addition, 605
 - augmented, *see* augmented vector
 - colinearity, 606
 - linearly independent, 606
 - mean, 617
 - total, *see* mean vector, total
 - orthogonal, 606
 - parameter, 85, 92
 - true, 92
 - separating, *see* vector, solution
 - solution, 223
 - space, 606
 - span, 606
 - support, *see* support vector
 - target, 289
 - weight, 216
- vector product, *see* outer product
- vector quantization, 582
- Venn diagram, 551
- vigilance, 565
- virtual value, 410
- visible unit, 361
- Voronoi
 - cell, 178
 - tessellation, 178, 528
- voting-based query selection, *see* query, selection, voting-based
- VQ, *see* vector quantization
- Wald statistic, 331, 332
- Wald test, 335
- War and Peace**, 414
- weak learner, *see* learner, weak
- weight
 - decay, 314–315, 335
 - excitatory, 284
 - inhibitory, 284
 - initialization, 311
 - sharing, 328
 - normalization, 560
 - space, 224
 - vector, 230
- whitening transform, *see* transform, whitening, 539
- wild shot pattern, *see* outlier pattern
- William of Occam, *see* Occam, William of
- window function, 577
 - Gaussian, 198
- winner-take-all, 305, 497
- Winnow algorithm, 234
 - balanced, 234, 279
 - gap, 235
- within-class scatter, *see* scatter, within-class
- word, 413
- wrapper, 380
- XOR problem
 - error surface, 298
 - network, 285
- z score, 622
- zero-information distribution, 182
- zero-one loss, *see* loss, zero-one