Index

Page numbers in **bold** indicate the primary source of information for the corresponding topic.

1-of-K coding scheme, 424

acceptance criterion, **538**, 541, 544
activation function, **180**, 213, 227
active constraint, 328, **709**
AdaBoost, 657, **658**
adaline, 196
adaptive rejection sampling, 530
ADF, *see* assumed density filtering
AIC, *see* Akaike information criterion
Akaike information criterion, **33**, 217
α family of divergences, 469
α recursion, 620
ancestral sampling, **365**, 525, 613
annular flow, 679
AR model, *see* autoregressive model
arc, 360
ARD, *see* automatic relevance determination
ARMA, *see* autoregressive moving average
assumed density filtering, 510
autoassociative networks, 592
automatic relevance determination, 259, 312, **349**, 485, 582
autoregressive hidden Markov model, 632
autoregressive model, 609
autoregressive moving average, 304

back-tracking, **415**, 630
backgammon, 3
backpropagation, 241
bagging, 656
basis function, **138**, 172, 204, 227
batch training, 240
Baum-Welch algorithm, 618
Bayes' theorem, 15
Bayes, Thomas, 21
Bayesian analysis, vii, 9, **21**
  hierarchical, 372
  model averaging, 654
Bayesian information criterion, 33, **216**
Bayesian model comparison, **161**, 473, 483
Bayesian network, 360
Bayesian probability, 21
belief propagation, 403
Bernoulli distribution, **69**, 113, 685
  mixture model, 444
Bernoulli, Jacob, 69
beta distribution, **71**, 686
beta recursion, 621
between-class covariance, 189
bias, **27**, 149
bias parameter, **138**, 181, 227, 346
bias-variance trade-off, 147
BIC, *see* Bayesian information criterion
binary entropy, 495
binomial distribution, **70**, 686

biological sequence, 610
bipartite graph, 401
bits, 49
blind source separation, 591
blocked path, 374, **378**, 384
Boltzmann distribution, 387
Boltzmann, Ludwig Eduard, 53
Boolean logic, 21
boosting, 657
bootstrap, **23**, 656
bootstrap filter, 646
box constraints, **333**, 342
Box-Muller method, 527

C4.5, 663
calculus of variations, 462
canonical correlation analysis, 565
canonical link function, 212
CART, *see* classification and regression trees
Cauchy distribution, **527**, 529, 692
causality, 366
CCA, *see* canonical correlation analysis
central differences, 246
central limit theorem, 78
chain graph, 393
chaining, 555
Chapman-Kolmogorov equations, 397
child node, 361
Cholesky decomposition, 528
chunking, 335
circular normal, *see* von Mises distribution
classical probability, 21
classification, 3
classification and regression trees, 663
clique, 385
clustering, 3
clutter problem, 511
co-parents, **383**, 492
code-book vectors, 429
combining models, 45, **653**
committee, 655
complete data set, 440
completing the square, 86
computational learning theory, 326, 344
concave function, 56
concentration parameter, **108**, 693
condensation algorithm, 646
conditional entropy, 55
conditional expectation, 20
conditional independence, 46, **372**, 383
conditional mixture model, *see* mixture model
conditional probability, 14
conjugate prior, 68, 98, **117**, 490
convex duality, 494
convex function, **55**, 493
convolutional neural network, 267
correlation matrix, 567
cost function, 41
covariance, 20
  between-class, 189
  within-class, 189
covariance matrix
  diagonal, 84
  isotropic, 84
  partitioned, **85**, 307
  positive definite, 308
Cox's axioms, 21
credit assignment, 3
cross-entropy error function, **206**, 209, 235, 631, 666
cross-validation, **32**, 161
cumulative distribution function, 18
curse of dimensionality, **33**, 36
curve fitting, 4

D map, *see* dependency map
d-separation, 373, **378**, 443
DAG, *see* directed acyclic graph
DAGSVM, 339
data augmentation, 537
data compression, 429
decision boundary, **39**, 179
decision region, **39**, 179
decision surface, *see* decision boundary
decision theory, 38
decision tree, 654, **663**, 673
decomposition methods, 335
degrees of freedom, 559
degrees-of-freedom parameter, **102**, 693
density estimation, 3, **67**

density network, 597
dependency map, 392
descendant node, 376
design matrix, **142**, 347
differential entropy, 53
digamma function, 687
directed acyclic graph, 362
directed cycle, 362
directed factorization, 381
Dirichlet distribution, **76**, 687
Dirichlet, Lejeune, 77
discriminant function, 43, 180, **181**
discriminative model, **43**, 203
distortion measure, 424
distributive law of multiplication, 396
DNA, 610
document retrieval, 299
dual representation, **293**, 329
dual-energy gamma densitometry, 678
dynamic programming, 411
dynamical system, 548

E step, *see* expectation step
early stopping, 259
ECM, *see* expectation conditional maximization
edge, 360
effective number of observations, **72**, 101
effective number of parameters, 9, **170**, 281
elliptical K-means, 444
EM, *see* expectation maximization
emission probability, 611
empirical Bayes, *see* evidence approximation
energy function, 387
entropy, 49
  conditional, 55
  differential, 53
  relative, 55
EP, *see* expectation propagation
ϵ-tube, 341
ϵ-insensitive error function, 340
equality constraint, 709
equivalent kernel, **159**, 301
erf function, 211
error backpropagation, *see* backpropagation
error function, **5**, 23
error-correcting output codes, 339
Euler, Leonhard, 465
Euler-Lagrange equations, 705
evidence approximation, **165**, 347, 581
evidence function, 161
expectation, 19
expectation conditional maximization, 454
expectation maximization, 113, **423**, 440
  Gaussian mixture, 435
  generalized, 454
  sampling methods, 536
expectation propagation, 315, 468, **505**
expectation step, 437
explaining away, 378
exploitation, 3
exploration, 3
exponential distribution, **526**, 688
exponential family, 68, **113**, 202, 490
extensive variables, 490

face detection, 2
face tracking, 355
factor analysis, 583
  mixture model, 595
factor graph, 360, **399**, 625
factor loading, 584
factorial hidden Markov model, 633
factorized distribution, **464**, 476
feature extraction, 2
feature map, 268
feature space, **292**, 586
Fisher information matrix, 298
Fisher kernel, 298
Fisher's linear discriminant, 186
flooding schedule, 417
forward kinematics, 272
forward problem, 272
forward propagation, **228**, 243
forward-backward algorithm, 618
fractional belief propagation, 517
frequentist probability, 21
fuel system, 376
function interpolation, 299
functional, 462, **703**
  derivative, 463

gamma densitometry, 678
gamma distribution, 529, **688**
gamma function, 71
gating function, 672
Gauss, Carl Friedrich, 79
Gaussian, 24, **78**, 688
  conditional, **85**, 93
  marginal, **88**, 93
  maximum likelihood, 93
  mixture, 110, 270, 273, **430**
  sequential estimation, 94
  sufficient statistics, 93
  wrapped, 110
Gaussian kernel, 296
Gaussian process, 160, **303**
Gaussian random field, 305
Gaussian-gamma distribution, **101**, 690
Gaussian-Wishart distribution, **102**, 475, 478, **690**
GEM, *see* expectation maximization, generalized
generalization, 2
generalized linear model, **180**, 213
generalized maximum likelihood, *see* evidence approximation
generative model, **43**, 196, 297, 365, 572, 631
generative topographic mapping, 597
  directional curvature, 599
  magnification factor, 599
geodesic distance, 596
Gibbs sampling, 542
  blocking, 546
Gibbs, Josiah Willard, 543
Gini index, 666
global minimum, 237
gradient descent, 240
Gram matrix, 293
graph-cut algorithm, 390
graphical model, 359
  bipartite, 401
  directed, 360
  factorization, 362, 384
  fully connected, 361
  inference, 393
  tree, 398
  treewidth, 417
  triangulated, 416
  undirected, 360
Green's function, 299
GTM, *see* generative topographic mapping

Hamilton, William Rowan, 549
Hamiltonian dynamics, 548
Hamiltonian function, 549
Hammersley-Clifford theorem, 387
handwriting recognition, 1, 610, 614
handwritten digit, 565, 614, **677**
head-to-head path, 376
head-to-tail path, 375
Heaviside step function, 206
Hellinger distance, 470
Hessian matrix, 167, 215, 217, 238, **249**
  diagonal approximation, 250
  exact evaluation, 253
  fast multiplication, 254
  finite differences, 252
  inverse, 252
  outer product approximation, 251
heteroscedastic, **273**, 311
hidden Markov model, 297, **610**
  autoregressive, 632
  factorial, 633
  forward-backward algorithm, 618
  input-output, 633
  left-to-right, 613
  maximum likelihood, 615
  scaling factor, 627
  sum-product algorithm, 625
  switching, 644
  variational inference, 625
hidden unit, 227
hidden variable, 84, **364**, 430, 559
hierarchical Bayesian model, 372
hierarchical mixture of experts, 673
hinge error function, 337
Hinton diagram, 584
histogram density estimation, 120
HME, *see* hierarchical mixture of experts
hold-out set, 11
homogeneous flow, 679
homogeneous kernel, 292
homogeneous Markov chain, **540**, 608

Hooke’s law, 580
hybrid Monte Carlo, 548
hyperparameter, **71**, 280, 311, 346, 372, 502
hyperprior, 372

I map, *see* independence map
i.i.d., *see* independent identically distributed
ICA, *see* independent component analysis
ICM, *see* iterated conditional modes
ID3, 663
identifiability, 435
image de-noising, 387
importance sampling, 525, **532**
importance weights, 533
improper prior, **118**, 259, 472
imputation step, 537
imputation-posterior algorithm, 537
inactive constraint, 328, **709**
incomplete data set, 440
independence map, 392
independent component analysis, 591
independent factor analysis, 592
independent identically distributed, **26**, 379
independent variables, 17
independent, identically distributed, 605
induced factorization, 485
inequality constraint, 709
inference, 38, **42**
information criterion, 33
information geometry, 298
information theory, 48
input-output hidden Markov model, 633
intensive variables, 490
intrinsic dimensionality, 559
invariance, 261
inverse gamma distribution, 101
inverse kinematics, 272
inverse problem, 272
inverse Wishart distribution, 102
IP algorithm, *see* imputation-posterior algorithm
IRLS, *see* iterative reweighted least squares
Ising model, 389
isomap, 596
isometric feature map, 596
iterated conditional modes, **389**, 415
iterative reweighted least squares, **207**, 210, 316, 354, 672

Jacobian matrix, **247**, 264
Jensen’s inequality, 56
join tree, 416
junction tree algorithm, 392, **416**

K nearest neighbours, 125
K-means clustering algorithm, **424**, 443
K-medoids algorithm, 428
Kalman filter, 304, **637**
    extended, 644
Kalman gain matrix, 639
Kalman smoother, 637
Karhunen-Loève transform, 561
Karush-Kuhn-Tucker conditions, 330, 333, 342, **710**
kernel density estimator, **122**, 326
kernel function, 123, **292**, 294
    Fisher, 298
    Gaussian, 296
    homogeneous, 292
    nonvectorial inputs, 297
    stationary, 292
kernel PCA, 586
kernel regression, 300, **302**
kernel substitution, 292
kernel trick, 292
kinetic energy, 549
KKT, *see* Karush-Kuhn-Tucker conditions
KL divergence, *see* Kullback-Leibler divergence
kriging, *see* Gaussian process
Kullback-Leibler divergence, **55**, 451, 468, 505

Lagrange multiplier, 707
Lagrange, Joseph-Louis, 329
Lagrangian, 328, 332, 341, **708**
laminar flow, 678
Laplace approximation, **213**, 217, 278, 315, 354
Laplace, Pierre-Simon, 24
large margin, *see* margin
lasso, 145
latent class analysis, 444
latent trait model, 597
latent variable, 84, **364**, 430, 559

lattice diagram, **414**, 611, 621, 629
LDS, *see* linear dynamical system
leapfrog discretization, 551
learning, 2
learning rate parameter, 240
least-mean-squares algorithm, 144
leave-one-out, 33
likelihood function, 22
likelihood weighted sampling, 534
linear discriminant, 181
    Fisher, 186
linear dynamical system, 84, **635**
    inference, 638
linear independence, 696
linear regression, 138
    EM, 448
    mixture model, 667
    variational, 486
linear smoother, 159
linear-Gaussian model, 87, **370**
linearly separable, 179
link, 360
link function, 180, 213
Liouville's Theorem, 550
LLE, *see* locally linear embedding
LMS algorithm, *see* least-mean-squares algorithm
local minimum, 237
local receptive field, 268
locally linear embedding, 596
location parameter, 118
log odds, 197
logic sampling, 525
logistic regression, **205**, 336
    Bayesian, 217, 498
    mixture model, 670
    multiclass, 209
logistic sigmoid, 114, 139, **197**, 205, 220, 227, 495
logit function, 197
loopy belief propagation, 417
loss function, 41
loss matrix, 41
lossless data compression, 429
lossy data compression, 429
lower bound, 484

M step, *see* maximization step
machine learning, vii
macrostate, 51
Mahalanobis distance, 80
manifold, **38**, 590, 595, 681
MAP, *see* maximum posterior
margin, 326, **327**, 502
    error, 334
    soft, 332
marginal likelihood, **162**, 165
marginal probability, 14
Markov blanket, **382**, 384, 545
Markov boundary, *see* Markov blanket
Markov chain, 397, **539**
    first order, 607
    homogeneous, **540**, 608
    second order, 608
Markov chain Monte Carlo, 537
Markov model, 607
    homogeneous, 612
Markov network, *see* Markov random field
Markov random field, 84, 360, **383**
max-sum algorithm, **411**, 629
maximal clique, 385
maximal spanning tree, 416
maximization step, 437
maximum likelihood, 9, **23**, 26, 116
    Gaussian mixture, 432
    singularities, 480
    type 2, *see* evidence approximation
maximum margin, *see* margin
maximum posterior, **30**, 441
MCMC, *see* Markov chain Monte Carlo
MDN, *see* mixture density network
MDS, *see* multidimensional scaling
mean, 24
mean field theory, 465
mean value theorem, 52
measure theory, 19
memory-based methods, 292
message passing, 396
    pending message, 417
    schedule, 417
    variational, 491
Metropolis algorithm, 538
Metropolis-Hastings algorithm, 541

microstate, 51
minimum risk, 44
Minkowski loss, 48
missing at random, **441**, 579
missing data, 579
mixing coefficient, 111
mixture component, 111
mixture density network, **272**, 673
mixture distribution, *see* mixture model
mixture model, 162, **423**
    conditional, 273, **666**
    linear regression, 667
    logistic regression, 670
    symmetries, 483
mixture of experts, 672
mixture of Gaussians, 110, 270, 273, **430**
MLP, *see* multilayer perceptron
MNIST data, 677
model comparison, 6, 32, **161**, 473, 483
model evidence, 161
model selection, 162
moment matching, **506**, 510
momentum variable, 548
Monte Carlo EM algorithm, 536
Monte Carlo sampling, 24, **523**
Moore-Penrose pseudo-inverse, *see* pseudo-inverse
moralization, **391**, 401
MRF, *see* Markov random field
multidimensional scaling, 596
multilayer perceptron, 226, **229**
multimodality, 272
multinomial distribution, 76, 114, **690**
multiplicity, 51
mutual information, 55, **57**

Nadaraya-Watson, *see* kernel regression
naive Bayes model, 46, **380**
nats, 50
natural language modelling, 610
natural parameters, 113
nearest-neighbour methods, 124
neural network, 225
    convolutional, 267
    regularization, 256
    relation to Gaussian process, 319
Newton-Raphson, **207**, 317
node, 360
noiseless coding theorem, 50
nonidentifiability, 585
noninformative prior, 23, **117**
nonparametric methods, 68, **120**
normal distribution, *see* Gaussian
normal equations, 142
normal-gamma distribution, **101**, 691
normal-Wishart distribution, **102**, 475, 478, 691
normalized exponential, *see* softmax function
novelty detection, 44
ν-SVM, 334

object recognition, 366
observed variable, 364
Occam factor, 217
oil flow data, 34, 560, 568, **678**
Old Faithful data, 110, 479, 484, **681**
on-line learning, *see* sequential learning
one-versus-one classifier, **183**, 339
one-versus-the-rest classifier, **182**, 338
ordered over-relaxation, 545
Ornstein-Uhlenbeck process, 305
orthogonal least squares, 301
outlier, 44, 185, 212
outliers, **103**
over-fitting, **6**, 147, 434, 464
over-relaxation, 544

PAC learning, *see* probably approximately correct
PAC-Bayesian framework, 345
parameter shrinkage, 144
parent node, 361
particle filter, 645
partition function, **386**, 554
Parzen estimator, *see* kernel density estimator
Parzen window, 123
pattern recognition, vii
PCA, *see* principal component analysis
pending message, 417
perceptron, 192
    convergence theorem, 194
    hardware, 196
perceptron criterion, 193
perfect map, 392

periodic variable, 105
phase space, 549
photon noise, 680
plate, 363
polynomial curve fitting, 4, 362
polytree, 399
position variable, 548
positive definite covariance, 81
positive definite matrix, 701
positive semidefinite covariance, 81
positive semidefinite matrix, 701
posterior probability, 17
posterior step, 537
potential energy, 549
potential function, 386
power EP, 517
power method, 563
precision matrix, 85
precision parameter, 24
predictive distribution, **30**, 156
preprocessing, 2
principal component analysis, **561**, 572, 593
  Bayesian, 580
  EM algorithm, 577
  Gibbs sampling, 583
  mixture distribution, 595
  physical analogy, 580
principal curve, 595
principal subspace, 561
principal surface, 596
prior, 17
  conjugate, 68, 98, **117**, 490
  consistent, 257
  improper, **118**, 259, 472
  noninformative, 23, **117**
probabilistic graphical model, *see* graphical model
probabilistic PCA, 570
probability, 12
  Bayesian, 21
  classical, 21
  density, 17
  frequentist, 21
  mass function, 19
  prior, 45
  product rule, 13, **14**, 359
  sum rule, 13, **14**, 359
  theory, 12
probably approximately correct, 344
probit function, **211**, 219
probit regression, 210
product rule of probability, 13, **14**, 359
proposal distribution, **528**, 532, 538
protected conjugate gradients, 335
protein sequence, 610
pseudo-inverse, **142**, 185
pseudo-random numbers, 526

quadratic discriminant, 199
quality parameter, 351

radial basis function, 292, **299**
Rauch-Tung-Striebel equations, 637
regression, 3
regression function, **47**, 95
regularization, 10
  Tikhonov, 267
regularized least squares, 144
reinforcement learning, 3
reject option, **42**, 45
rejection sampling, 528
relative entropy, 55
relevance vector, 348
relevance vector machine, 161, **345**
responsibility, 112, **432**, 477
ridge regression, 10
RMS error, *see* root-mean-square error
Robbins-Monro algorithm, 95
robot arm, 272
robustness, **103**, 185
root node, 399
root-mean-square error, 6
Rosenblatt, Frank, 193
rotation invariance, **573**, 585
RTS equations, *see* Rauch-Tung-Striebel equations
running intersection property, 416
RVM, *see* relevance vector machine

sample mean, 27
sample variance, 27
sampling-importance-resampling, 534
scale invariance, 119, **261**

scale parameter, 119
scaling factor, 627
Schwarz criterion, *see* Bayesian information criterion
self-organizing map, 598
sequential data, 605
sequential estimation, 94
sequential gradient descent, 144, 240
sequential learning, **73**, 143
sequential minimal optimization, 335
serial message passing schedule, 417
Shannon, Claude, 55
shared parameters, 368
shrinkage, 10
Shur complement, 87
sigmoid, *see* logistic sigmoid
simplex, 76
single-class support vector machine, 339
singular value decomposition, 143
sinusoidal data, 682
SIR, *see* sampling-importance-resampling
skip-layer connection, 229
slack variable, 331
slice sampling, 546
SMO, *see* sequential minimal optimization
smoother matrix, 159
smoothing parameter, 122
soft margin, 332
soft weight sharing, 269
softmax function, 115, **198**, 236, 274, 356, 497
SOM, *see* self-organizing map
sparsity, 145, 347, **349**, 582
sparsity parameter, 351
spectrogram, 606
speech recognition, **605**, 610
sphereing, 568
spline functions, 139
standard deviation, 24
standardizing, 425, **567**
state space model, 609
    switching, 644
stationary kernel, 292
statistical bias, *see* bias
statistical independence, *see* independent variables
statistical learning theory, *see* computational learning theory, 326, 344
steepest descent, 240
Stirling's approximation, 51
stochastic, 5
stochastic EM, 536
stochastic gradient descent, 144, 240
stochastic process, 305
stratified flow, 678
Student's t-distribution, **102**, 483, 691
subsampling, 268
sufficient statistics, 69, 75, **116**
sum rule of probability, 13, **14**, 359
sum-of-squares error, **5**, 29, 184, 232, 662
sum-product algorithm, 399, **402**
    for hidden Markov model, 625
supervised learning, 3
support vector, 330
support vector machine, 225
    for regression, 339
    multiclass, 338
survival of the fittest, 646
SVD, *see* singular value decomposition
SVM, *see* support vector machine
switching hidden Markov model, 644
switching state space model, 644
synthetic data sets, 682

tail-to-tail path, 374
tangent distance, 265
tangent propagation, 262, **263**
tapped delay line, 609
target vector, 2
test set, 2, **32**
threshold parameter, 181
tied parameters, 368
Tikhonov regularization, 267
time warping, 615
tomography, 679
training, 2
training set, 2
transition probability, **540**, 610
translation invariance, 118, **261**
tree-reweighted message passing, 517
treewidth, 417

trellis diagram, *see* lattice diagram
triangulated graph, 416
type 2 maximum likelihood, *see* evidence approximation

undetermined multiplier, *see* Lagrange multiplier
undirected graph, *see* Markov random field
uniform distribution, 692
uniform sampling, 534
uniquenesses, 584
unobserved variable, *see* latent variable
unsupervised learning, 3
utility function, 41

validation set, 11, **32**
Vapnik-Chervonenkis dimension, 344
variance, **20**, 24, 149
variational inference, 315, **462**, 635
  for Gaussian mixture, 474
  for hidden Markov model, 625
  local, 493
VC dimension, *see* Vapnik-Chervonenkis dimension
vector quantization, 429
vertex, *see* node
visualization, 3
Viterbi algorithm, 415, **629**
von Mises distribution, **108**, 693

wavelets, 139
weak learner, 657
weight decay, 10, **144**, 257
weight parameter, 227
weight sharing, 268
  soft, 269
weight vector, 181
weight-space symmetry, **231**, 281
weighted least squares, 668
well-determined parameters, 170
whitening, 299, **568**
Wishart distribution, **102**, 693
within-class covariance, 189
Woodbury identity, 696
wrapped distribution, 110

Yellowstone National Park, 110, **681**