Index

k-nearest neighbors, 13

active learning, 17
AD, 176
AdaBoost, 212
ADAM, 192
adaptive boosting, 214
AI, 1
approximate inference, 357
    expectation propagation, 357
    loopy belief propagation, 357
    Monte Carlo sampling, 357, 361
    variational inference, 357
approximation error, 105
artificial intelligence, 1
artificial neural networks, 151
artificial neuron, 152
autoencoder, 90
automatic differentiation, 176

back-propagation, 153, 176
bag-of-words, 77
bagging, 208
bandlimitedness, 13
batch normalization, 160
Baum–Welch algorithm, 280, 286
Bayes decision rule, 224
Bayes error, 226
Bayesian classification, 314
Bayesian decision theory, 222
Bayesian inference, 313
Bayesian learning, 311, 313
    evidence, 312
    hyperparameter, 317
    maximum a posteriori estimation, 315
    posterior distribution, 312
    prior distribution, 312
Bayesian network, 343, 346
    causal Bayesian network, 347
    conditional independence, 346
    latent Dirichlet allocation, 362
    naive Bayes classifier, 361
Bernoulli distribution, 34
beta distribution, 35
bias–variance trade-off, 10, 30
bidirectional recurrent neural networks, 172
bilinear function, 144
binomial distribution, 34
biological neuron, 152
    axon, 152
    dendrites, 152
    synapse, 152
blessing of nonuniformity, 15
blind source separation, 300
BN features, 91
boosted trees, 75
boosting, 209
bootstrap, 208
bootstrap aggregating, 208
bottleneck features, 91

c.d.f., 28
CART, 205
categorical distribution, 35
causal inference, 16
chain, 348
class-conditional distribution, 223
classification, 4
classification and regression tree, 205
clustering, 5, 270
CNN, 166
collaborative filtering, 141
collider, 349
colliding, 349
complementary slackness, 57
compressed sensing, 146
conditional dependence, 368
conditional distribution, 31
conditional entropy, 43
conditional random field, 368
confounder, 347
confounding, 347
conjugate prior, 318
continuous latent variables, 292
convergence rate, 61
    linear, 61
    sublinear, 61
    superlinear, 61

convex optimization, 50
convex set, 50
convolutional neural networks, 166
    feature maps, 168
    kernel, 167
    receptive field, 170
covariance, 32
covariance matrix, 33
CRF, 368
critical point, 51
cross-attention, 199
cumulative distribution function, 28
curse of dimensionality, 14

d-separation, 350
data augmentation, 195
decision trees, 7, 205
deep generative model, 303
    generative adversarial nets, 307
    variational autoencoder, 304
deep learning, 75, 151
density estimation, 231
dictionary learning, 74, 145
dimension reduction, 79
dimensionality reduction, 15, 68, 79
directed graphical model, 343
Dirichlet distribution, 36
Dirichlet process, 333
discriminative model, 68, 221, 236
disentangled representation learning, 295
distribution, 27
distribution-free model, 7
domain adaption, 16
dropout, 195

e-family, 259
element-wise multiplication, 119
EM, 261
EM algorithm, 265
    E-step, 266
    M-step, 266
empirical Bayes methods, 323
empirical distribution, 237
empirical loss, 98
empirical risk, 98
end-to-end learning, 3, 197
ensemble learning, 203
entangled model, 291
    deep generative model, 303
    factor, 293
    linear Gaussian model, 296
    mixing function, 293
    non-Gaussian model, 300
    residual, 293
entropy, 42
error back-propagation, 176
estimation error, 105
exact inference, 357
    belief propagation, 360
    belief propagation algorithm, 357
    forward–backward algorithm, 357, 358
    junction-tree algorithm, 357
    max-sum algorithm, 357
    message passing, 360
    sum-product algorithm, 357
expectation, 28
expectation-maximization method, 261
    auxiliary function, 262
expected risk, 98
explain away, 349
exponential family, 259
    natural parameter, 259
    sufficient statistics, 259
exponential loss, 135

factor analysis, 294, 298
feature engineering, 68, 77
feature extraction, 3, 67
feature selection, 78
feedforward sequential memory network, 202
finite mixture distribution, 257
finite-dimensional model, 7
FOFE, 78
forward–backward algorithm, 276, 279
FSMN, 202
fully connected deep neural networks, 165, 185
functional, 209

gamma distribution, 378
GAN, 294, 307
gated recurrent unit, 172
Gaussian Bayesian network, 345
Gaussian distribution, 38, 39
    covariance matrix, 39
    mean vector, 39
    precision matrix, 39
Gaussian kernel, 89, 124
Gaussian mixture model, 258, 268
Gaussian model, 240
Gaussian process, 332

classification, 338
covariance function, 334
kernel, 334
mean function, 334
regression, 335
GBM, 214
GBRT, 214
generalization bound, 100
generalization error, 105
generalized linear model, 250
link function, 251
generative adversarial nets, 294, 307
generative model, 68, 221, 234
GLM, 250
global maximum, 51
global minimum, 51
global optimization, 52
GMM, 258, 268
gradient, 51
gradient boosting, 210
gradient descent, 60
gradient tree boosting, 214
gradient-boosted regression tree, 214
gradient-boosting machine, 214
graphical model, 343
Bayesian network, 343
directed graphical model, 343
factor graph, 361
inference algorithm, 355
junction tree, 361
Markov random field, 344
parameter estimation, 354
structure learning, 354
undirected graphical model, 343
GRU, 172

Hessian matrix, 53
hidden Markov model, 271, 276
Baum–Welch algorithm, 280, 286
decoding problem, 279
evaluation problem, 276
forward–backward algorithm, 276, 279
training problem, 280
Viterbi algorithm, 279
higher-order recurrent neural networks, 172
hinge function, 134
hinge loss, 134, 135
HMM, 271
Hoeffding's inequality, 100
HOPE, 294, 302
HORNN, 172
hybrid orthogonal projection and estimation, 294, 302
hyperparameter, 16, 109
hypothesis space, 98

i.i.d. assumption, 232
ICA, 294, 300
IFA, 294, 301
imitation learning, 17
in-domain data, 3
in-sample error, 98
independent component analysis, 294, 300
independent factor analysis, 294, 301
independent random variables, 32
infinite mixture model, 288
information theory, 41
conditional entropy, 43
entropy, 42
information, 41
joint entropy, 43
mutual information, 44
input space, 97
inverse-gamma distribution, 319
inverse-Wishart distribution, 319, 378
Isomap, 88
isometric feature mapping, 88

Jacobian matrix, 40
Jensen's inequality, 46
joint distribution, 30
joint entropy, 43

K-means, 270
K-means clustering, 270
k-NN, 13, 18
kernel function, 124
kernel PCA, 125
kernel trick, 123, 125
keyword selection, 44
KL divergence, 46
Kullback–Leibler divergence, 46

L2 function, 151
Lagrange dual function, 56
Lagrange dual problem, 56
Lagrange multipliers, 54
Lagrangian function, 55, 56
language modeling, 248
Laplace's method, 324

LASSO, 73, 139
latent Dirichlet allocation, 74, 362
latent semantic analysis, 142
law of the smooth world, 12
layer normalization, 160
LDA, 74, 84, 362
learnability, 99
learning to learn, 16
least-square error, 112
likelihood function, 232
linear algebra, 19
    determinant, 22
    eigenvalue, 23
    eigenvector, 23
    identity matrix, 22
    inner product, 22
    inverse matrix, 22
    matrix, 19
    matrix multiplication, 20
    matrix transpose, 21
    symmetric matrix, 22
    trace, 23
    vector, 19
linear dimension reduction, 79
linear discriminant analysis, 84, 243
linear Gaussian model, 296, 345
    factor analysis, 298
    probabilistic PCA, 296
linear kernel, 124
linear programming, 50
linear regression, 72, 112, 251
linear SVM, 73, 116
linear transformation, 20
linear-chain conditional random field, 369
linearly nonseparable, 107
linearly separable, 107
Lipschitz continuous, 13, 18
LLE, 87
local extreme, 51
local maximum, 51
local minimum, 51
local optimization, 52
locality modeling, 158
locally linear embedding, 87
log-linear model, 251, 253
log-sum, 262
logistic loss, 135
logistic regression, 73, 114, 251
long short-term memory, 172
loss function, 98
Lp function, 151
Lp norm, 137
LSA, 142
LSTM, 172

machine learning, 2
manifold, 86
manifold learning, 15, 87
MAP estimation, 315
MAP rule, 224
margin, 109
marginal distribution, 31
marginal likelihood, 323
marginalization, 31
Markov assumption, 246
Markov chain model, 245
Markov random field, 344, 366
    Boltzmann distribution, 367
    clique, 366
    conditional random field, 368
    energy function, 367
    maximum clique, 366
    partition function, 367
    potential function, 367
    restricted Boltzmann machine, 370
matrix calculus, 25
matrix completion, 142
matrix factorization, 24, 74, 140
maximum a posteriori estimation, 315
maximum a posteriori rule, 224
maximum-entropy model, 254
maximum-likelihood classifier, 234
maximum-likelihood estimation, 231
maximum-marginal-likelihood estimation, 323
MCE, 113
MDL, 11
MDS, 88
mean, 28
mean field theory, 327
mediator, 348
Mercer's condition, 124
meta-learning, 16
    meta-learner, 16
minimum classification error, 113
minimum description length, 11
mixture model, 257, 261
ML, 231
MLE, 231

model space, 98
moments, 28
multiclass SVM, 127
multidimensional scaling, 88
multimodality, 257
multinomial distribution, 34
multinomial mixture model, 258
multinomial model, 243
multiplication rule of probability, 33
multivariate Gaussian distribution, 39
mutual information, 44

N-gram model, 249
naive Bayes classifier, 361
nearest neighbors, 13
neural networks, 75, 90, 151
    attention, 162
        attention function, 163
        key, 163
        query, 163
        value matrix, 164
    batch normalization, 160
    convolution, 157, 180
        kernel, 157
        locality modelling, 158
        weight sharing, 158
    cross entropy, 175
    error signal, 176
    full connection, 156, 178
    layer, 156
    layer normalization, 160
    max-pooling, 159, 184
    mean-square error, 175
    nonlinear activation, 158, 179
    normalization, 159, 183
    SGD, 189
        epoch number, 190
        initialization, 190
        learning rate, 191
        mini-batch size, 190
    softmax, 159, 180
    tapped delay line, 161
    time-delayed feedback, 161
    universal approximator, 154
    weight decay, 194
    weight normalization, 194
neuron, 152
Newton boosting, 211
Newton method, 63
no-free-lunch theorem, 11
nonlinear dimension reduction, 86, 90
    manifold learning, 87
    neural networks, 90
nonlinear SVM, 73, 123
nonlinear transformation, 21
nonparametric Bayesian method, 333
    Dirichlet process, 333
    Gaussian process, 333
nonparametric model, 7

Occam's razor, 11
one-versus-all strategy, 127
one-versus-one strategy, 127
online learning, 17
optimization, 48
    convex optimization, 50
    equality constraint, 49
    first-order method, 60
    inequality constraint, 49
    linear programming, 50
    second-order method, 63
    zero-order method, 59
output space, 97
overfitting, 8

p.d.f., 28
p.m.f., 27
parametric model, 7
PCA, 80
Pearson's correlation coefficient, 79
perceptron, 108
plug-in MAP rule, 229
Poisson distribution, 377
Poisson regression, 251, 252
polynomial kernel, 124
positive definite matrix, 24
positive semidefinite matrix, 24
predictive distribution, 314
principal component, 80
principal component analysis, 80
prior probability, 223
prior specification, 313
probabilistic functions of Markov chains, 276
probabilistic PCA, 294, 296
probability density function, 28
probability distribution, 27
probability function, 27
probability mass function, 27
probit regression, 251, 252

product rule of probability, 33
product space, 30
projected gradient descent, 59, 127

QDA, 242
quadratic discriminant analysis, 242
quadratic programming, 126
quasi-Newton methods, 63

radial basis function, 124
random forests, 208
random variable, 27
    transformation of random variables, 40
RBF kernel, 124
RBM, 370
recommendation, 141
rectified linear loss, 135
rectified linear unit, 153
recurrent neural networks, 170
regression, 4
    curve fitting, 6
regularization, 10, 134
reinforcement learning, 15
    deep Q-learning, 15
    deep reinforcement learning, 15
    Q-learning, 15
ReLU, 153
restricted Boltzmann machine, 370
ridge regression, 72, 139
RNN, 170
rule of sum in probability, 31

saddle point, 51
Sammon mapping, 88
self-attention, 172
semisupervised learning, 5
separation margin, 109
seq2seq, 198
sequence-to-sequence learning, 198
sequential Bayesian Learning, 315
sequential minimization optimization, 127, 131
SGD, 61, 62
shattering, 102
shrinkage, 215
sigmoid function, 114
sigmoid loss, 135
singular value decomposition, 24, 140
SMO, 127, 131
SNE, 89
soft margin, 121
soft SVM, 73, 121
softmax function, 115, 159
sparse representation learning, 145
sparse sampling, 146
square loss, 135
stationary point, 51
statistical data modeling, 229
steepest descent, 60
stochastic gradient descent, 61
    mini-batch, 62
stochastic neighborhood embedding, 89
strong duality, 57
structured learning, 4
structured prediction, 4
sufficient statistics, 259
supervised learning, 5
support of a distribution, 34
support vector machine, 116
SVD, 24, 140
SVM, 116
symbolic approach, 2
    expert system, 2
    knowledge base, 1
    rule, 1

t-SNE, 89
target function, 97
tensor, 20
text categorization, 254
tf-idf, 78
topic modeling, 74
transfer learning, 16
transformer, 172
tree boosting, 75

unconstrained optimization, 50
uncorrelated random variables, 32
underfitting, 8
undirected graphical model, 343, 366
uniform distribution, 377
unimodal model, 239
unimodality, 239
universal approximator, 154
unsupervised learning, 5

VAE, 294, 304
Vapnik–Chervonenkis dimension, 102
variance, 28
variational autoencoder, 294, 304
variational Bayesian method, 326

variational distribution, 327
VB, 326
VC dimension, 102
Viterbi algorithm, 279
    token-passing algorithm, 280
    Viterbi path, 279
von Mises–Fisher distribution, 379

weakly supervised learning, 5
weight decay, 194
weight sharing, 158

