Index

0-1 loss, **104**, 276

Absolute value rectification, 192
Accuracy, 425
Activation function, 170
Active constraint, 95
AdaGrad, 307
ADALINE, *see* adaptive linear element
Adam, 308, 427
Adaptive linear element, 15, 24, 27
Adversarial example, 268
Adversarial training, 268, 271, 532
Affine, 110
AIS, *see* annealed importance sampling
Almost everywhere, 71
Almost sure convergence, 130
Ancestral sampling, 582, 597
ANN, *see* Artificial neural network
Annealed importance sampling, 627, 670, 719
Approximate Bayesian computation, 718
Approximate inference, 585
Artificial intelligence, 1
Artificial neural network, *see* Neural network
ASR, *see* automatic speech recognition
Asymptotically unbiased, 124
Audio, 102, 360, 460
Autoencoder, 4, 356, **504**
Automatic speech recognition, 460

Back-propagation, 203
Back-propagation through time, **384**
Backprop, *see* back-propagation
Bag of words, 473
Bagging, 256
Batch normalization, 268, 427
Bayes error, **117**
Bayes' rule, 70
Bayesian hyperparameter optimization, 438
Bayesian network, *see* directed graphical model
Bayesian probability, 55
Bayesian statistics, **135**
Belief network, *see* directed graphical model
Bernoulli distribution, 62
BFGS, 316
Bias, 124, 229
Bias parameter, 110
Biased importance sampling, 595
Bigram, 464
Binary relation, 484
Block Gibbs sampling, 601
Boltzmann distribution, 572
Boltzmann machine, 572, 656
BPTT, *see* back-propagation through time
Broadcasting, 34
Burn-in, 599

CAE, *see* contractive autoencoder
Calculus of variations, 179
Categorical distribution, *see* multinoulli distribution
CD, *see* contrastive divergence
Centering trick (DBM), 675
Central limit theorem, 63
Chain rule (calculus), 206
Chain rule of probability, 59

Chess, 2
Chord, 581
Chordal graph, 581
Class-based language models, 465
Classical dynamical system, 375
Classification, 100
Clique potential, *see* factor (graphical model)
CNN, *see* convolutional neural network
Collaborative Filtering, 480
Collider, *see* explaining away
Color images, 360
Complex cell, 365
Computational graph, 204
Computer vision, 454
Concept drift, 540
Condition number, 279
Conditional computation, *see* dynamic structure
Conditional independence, xiii, 60
Conditional probability, 59
Conditional RBM, 687
Connectionism, 17, 445
Connectionist temporal classification, 462
Consistency, 130, 515
Constrained optimization, 93, 237
Content-based addressing, 421
Content-based recommender systems, 482
Context-specific independence, 575
Contextual bandits, 482
Continuation methods, 327
Contractive autoencoder, 523
Contrast, 456
Contrastive divergence, 291, 612, 674
Convex optimization, 141
Convolution, 330, 685
Convolutional network, 16
Convolutional neural network, 254, **330**, 427, 462
Coordinate descent, 321, 673
Correlation, 61
Cost function, *see* objective function
Covariance, xiii, 61
Covariance matrix, 62
Coverage, 426
Critical temperature, 605
Cross-correlation, 332
Cross-entropy, **75**, 132
Cross-validation, 122
CTC, *see* connectionist temporal classification
Curriculum learning, 328
Curse of dimensionality, 154
Cyc, 2

D-separation, 574
DAE, *see* denoising autoencoder
Data generating distribution, **111**, 131
Data generating process, 111
Data parallelism, 449
Dataset, 105
Dataset augmentation, 271, 459
DBM, *see* deep Boltzmann machine
DCGAN, 553, 554, 703
Decision tree, **145**, 550
Decoder, 4
Deep belief network, 27, 531, 633, 659, 662, 686, 694
Deep Blue, 2
Deep Boltzmann machine, 24, 27, 531, 633, 654, 659, 665, 674, 686
Deep feedforward network, 167, 427
Deep learning, 2, 5
Denoising autoencoder, 512, 691
Denoising score matching, 621
Density estimation, 103
Derivative, xiii, 83
Design matrix, **106**
Detector layer, 339
Determinant, xii
Diagonal matrix, 41
Differential entropy, 74, 648
Dirac delta function, 65
Directed graphical model, 77, 509, 565, 694
Directional derivative, 85
Discriminative fine-tuning, *see* supervised fine-tuning
Discriminative RBM, 688
Distributed representation, 17, 150, 548
Domain adaptation, 538

Dot product, 34, 141
Double backprop, 271
Doubly block circulant matrix, 333
Dream sleep, 611, 654
DropConnect, 266
Dropout, **258**, 427, 432, 433, 674, 691
Dynamic structure, 450, 451

E-step, 636
Early stopping, 246, 247, 249, 250, 427
EBM, *see* energy-based model
Echo state network, 24, 27, 405
Effective capacity, 114
Eigendecomposition, 42
Eigenvalue, 42
Eigenvector, 42
ELBO, *see* evidence lower bound
Element-wise product, *see* Hadamard product, *see* Hadamard product
EM, *see* expectation maximization
Embedding, 518
Empirical distribution, 66
Empirical risk, 276
Empirical risk minimization, 276
Encoder, 4
Energy function, 571
Energy-based model, 571, 597, 656, 665
Ensemble methods, 256
Epoch, 246
Equality constraint, 94
Equivariance, 338
Error function, *see* objective function
ESN, *see* echo state network
Euclidean norm, 39
Euler-Lagrange equation, 648
Evidence lower bound, 635, 663
Example, 99
Expectation, 60
Expectation maximization, 636
Expected value, *see* expectation
Explaining away, 576, 633, 646
Exploitation, 483
Exploration, 483
Exponential distribution, **65**

F-score, 425
Factor (graphical model), 569
Factor analysis, 492
Factor graph, 581
Factors of variation, 4
Feature, 99
Feature selection, 236
Feedforward neural network, 167
Fine-tuning, 323
Finite differences, 441
Forget gate, 306
Forward propagation, 203
Fourier transform, 360, 362
Fovea, 366
FPCD, 616
Free energy, **573**, 682
`Freebase`, 485
Frequentist probability, 55
Frequentist statistics, **135**
Frobenius norm, 46
Fully-visible Bayes network, 707
Functional derivatives, 647
FVBN, *see* fully-visible Bayes network

Gabor function, 368
GANs, *see* generative adversarial networks
Gated recurrent unit, 427
Gaussian distribution, *see* normal distribution
Gaussian kernel, 142
Gaussian mixture, 67, 188
GCN, *see* global contrast normalization
`GeneOntology`, 485
Generalization, 110
Generalized Lagrange function, *see* generalized Lagrangian
Generalized Lagrangian, 94
Generative adversarial networks, 691, 702
Generative moment matching networks, 705
Generator network, 695
Gibbs distribution, 570
Gibbs sampling, 583, 601
Global contrast normalization, 456
GPU, *see* graphics processing unit
Gradient, 84

Gradient clipping, 289, 416
Gradient descent, 83, 85
Graph, xii
Graphical model, *see* structured probabilistic model
Graphics processing unit, 446
Greedy algorithm, 323
Greedy layer-wise unsupervised pretraining, 530
Greedy supervised pretraining, 323
Grid search, 434

Hadamard product, xii, 34
Hard tanh, 196
Harmonium, *see* restricted Boltzmann machine
Harmony theory, 573
Helmholtz free energy, *see* evidence lower bound
Hessian, 223
Hessian matrix, xiii, 87
Heteroscedastic, 187
Hidden layer, 6, 167
Hill climbing, 86
Hyperparameter optimization, 434
Hyperparameters, 120, 432
Hypothesis space, 112, 118

i.i.d. assumptions, 111, 122, 268
Identity matrix, 36
ILSVRC, *see* ImageNet Large-Scale Visual Recognition Challenge
ImageNet Large-Scale Visual Recognition Challenge, 23
Immorality, 579
Importance sampling, 594, 626, 700
Importance weighted autoencoder, 700
Independence, xiii, 60
Independent and identically distributed, *see* i.i.d. assumptions
Independent component analysis, 493
Independent subspace analysis, 495
Inequality constraint, 94
Inference, 564, 585, 633, 635, 637, 640, 650, 653
Information retrieval, 527
Initialization, 301
Integral, xiii
Invariance, 342
Isotropic, 65

Jacobian matrix, xiii, 72, 86
Joint probability, 57

k-means, 364, 548
k-nearest neighbors, **143**, 550
Karush-Kuhn-Tucker conditions, 95, 237
Karush–Kuhn–Tucker, 94
Kernel (convolution), 331, 332
Kernel machine, 550
Kernel trick, 141
KKT, *see* Karush–Kuhn–Tucker
KKT conditions, *see* Karush-Kuhn-Tucker conditions
KL divergence, *see* Kullback-Leibler divergence
Knowledge base, 2, 485
Krylov methods, 223
Kullback-Leibler divergence, xiii, **74**

Label smoothing, 243
Lagrange multipliers, 94, 648
Lagrangian, *see* generalized Lagrangian
LAPGAN, 704
Laplace distribution, **65**, 498, 499
Latent variable, 67
Layer (neural network), 167
LCN, *see* local contrast normalization
Leaky ReLU, 192
Leaky units, 408
Learning rate, 85
Line search, 85, 86, 93
Linear combination, 37
Linear dependence, 38
Linear factor models, 491
Linear regression, **107**, 110, 140
Link prediction, 486
Lipschitz constant, 92
Lipschitz continuous, 92
Liquid state machine, 405

Local conditional probability distribution, 566
Local contrast normalization, 458
Logistic regression, 3, **140**, 140
Logistic sigmoid, 7, 67
Long short-term memory, 18, 25, 306, **410**, 427
Loop, 581
Loopy belief propagation, 587
Loss function, *see* objective function
L^p norm, 39
LSTM, *see* long short-term memory

M-step, 636
Machine learning, 2
Machine translation, 101
Main diagonal, 33
Manifold, 160
Manifold hypothesis, 161
Manifold learning, 161
Manifold tangent classifier, 272
MAP approximation, 138, 507
Marginal probability, 58
Markov chain, 597
Markov chain Monte Carlo, 597
Markov network, *see* undirected model
Markov random field, *see* undirected model
Matrix, xi, xii, 32
Matrix inverse, 36
Matrix product, 34
Max norm, 40
Max pooling, 339
Maximum likelihood, **131**
Maxout, 192, 427
MCMC, *see* Markov chain Monte Carlo
Mean field, 640, 641, 674
Mean squared error, 108
Measure theory, 71
Measure zero, 71
Memory network, 418, 420
Method of steepest descent, *see* gradient descent
Minibatch, 279
Missing inputs, 100
Mixing (Markov chain), 603
Mixture density networks, 188
Mixture distribution, 66
Mixture model, 188, 512
Mixture of experts, 452, 550
MLP, *see* multilayer perception
MNIST, 21, 22, 674
Model averaging, 256
Model compression, 450
Model identifiability, 284
Model parallelism, 449
Moment matching, 705
Moore-Penrose pseudoinverse, 45, 239
Moralized graph, 579
MP-DBM, *see* multi-prediction DBM
MRF (Markov Random Field), *see* undirected model
MSE, *see* mean squared error
Multi-modal learning, 541
Multi-prediction DBM, 676
Multi-task learning, 244, 540
Multilayer perception, 5
Multilayer perceptron, 27
Multinomial distribution, 62
Multinoulli distribution, 62

n-gram, **463**
NADE, 710
Naive Bayes, 3
Nat, 73
Natural image, 561
Natural language processing, 463
Nearest neighbor regression, **115**
Negative definite, 89
Negative phase, 472, 608, 610
Neocognitron, 16, 24, 27, 367
Nesterov momentum, 300
Netflix Grand Prize, 258, 481
Neural language model, 465, 478
Neural network, 13
Neural Turing machine, 420
Neuroscience, 15
Newton's method, 89, 310
NLM, *see* neural language model
NLP, *see* natural language processing
No free lunch theorem, 116

Noise-contrastive estimation, 622
Non-parametric model, **114**
Norm, xiv, 39
Normal distribution, 63, 64, 125
Normal equations, **109**, 109, 112, 234
Normalized initialization, 303
Numerical differentiation, *see* finite differences

Object detection, 455
Object recognition, 455
Objective function, 82
OMP-k, *see* orthogonal matching pursuit
One-shot learning, 540
Operation, 204
Optimization, 80, 82
Orthodox statistics, *see* frequentist statistics
Orthogonal matching pursuit, 27, **255**
Orthogonal matrix, 42
Orthogonality, 41
Output layer, 167

Parallel distributed processing, 17
Parameter initialization, 301, 407
Parameter sharing, 253, 335, 373, 375, 389
Parameter tying, *see* Parameter sharing
Parametric model, **114**
Parametric ReLU, 192
Partial derivative, 84
Partition function, 570, 607, 671
PCA, *see* principal components analysis
PCD, *see* stochastic maximum likelihood
Perceptron, 15, 27
Persistent contrastive divergence, *see* stochastic maximum likelihood
Perturbation analysis, *see* reparametrization trick
Point estimator, 122
Policy, 482
Pooling, 330, 685
Positive definite, 89
Positive phase, 472, 608, 610, 658, 670
Precision, 425
Precision (of a normal distribution), 63, 65
Predictive sparse decomposition, 525
Preprocessing, 455
Pretraining, 323, 530
Primary visual cortex, 365
Principal components analysis, 48, 146–148, 492, 633
Prior probability distribution, **135**
Probabilistic max pooling, 685
Probabilistic PCA, 492, 493, 634
Probability density function, 58
Probability distribution, 56
Probability mass function, 56
Probability mass function estimation, 103
Product of experts, 572
Product rule of probability, *see* chain rule of probability
PSD, *see* predictive sparse decomposition
Pseudolikelihood, 617

Quadrature pair, 369
Quasi-Newton methods, 316

Radial basis function, 196
Random search, 436
Random variable, 56
Ratio matching, 620
RBF, 196
RBM, *see* restricted Boltzmann machine
Recall, 425
Receptive field, 337
Recommender Systems, 480
Rectified linear unit, 171, 192, 427, 509
Recurrent network, 27
Recurrent neural network, 378
Regression, 101
Regularization, **120**, 120, 177, 228, 432
Regularizer, 119
REINFORCE, 691
Reinforcement learning, 25, 106, 482, 691
Relational database, 485
Reparametrization trick, 690
Representation learning, 3
Representational capacity, 114
Restricted Boltzmann machine, 356, 461, 481, 589, 633, 658, 659, 674, 678, 680, 682, 685

Ridge regression, *see* weight decay
Risk, 275
RNN-RBM, 687

Saddle points, 285
Sample mean, 125
Scalar, xi, xii, 31
Score matching, 515, 619
Second derivative, 86
Second derivative test, 89
Self-information, 73
Semantic hashing, 527
Semi-supervised learning, 243
Separable convolution, 362
Separation (probabilistic modeling), 574
Set, xii
SGD, *see* stochastic gradient descent
Shannon entropy, xiii, 73
Shortlist, 468
Sigmoid, xiv, *see* logistic sigmoid
Sigmoid belief network, 27
Simple cell, 365
Singular value, *see* singular value decomposition
Singular value decomposition, 44, 148, 481
Singular vector, *see* singular value decomposition
Slow feature analysis, 495
SML, *see* stochastic maximum likelihood
Softmax, 183, 420, 452
Softplus, xiv, 68, 196
Spam detection, 3
Sparse coding, 321, 356, 498, 633, 694
Sparse initialization, 304, 407
Sparse representation, 146, 226, 254, 507, 558
Spearmint, 438
Spectral radius, 406
Speech recognition, *see* automatic speech recognition
Sphering, *see* whitening
Spike and slab restricted Boltzmann machine, 682
SPN, *see* sum-product network
Square matrix, 38
ssRBM, *see* spike and slab restricted Boltzmann machine
Standard deviation, 61
Standard error, 127
Standard error of the mean, 128, 278
Statistic, 122
Statistical learning theory, 110
Steepest descent, *see* gradient descent
Stochastic back-propagation, *see* reparametrization trick
Stochastic gradient descent, 15, 150, 279, **294**, 674
Stochastic maximum likelihood, 614, 674
Stochastic pooling, 266
Structure learning, 584
Structured output, 101, 687
Structured probabilistic model, 77, 560
Sum rule of probability, 58
Sum-product network, 555
Supervised fine-tuning, 531, 664
Supervised learning, **105**
Support vector machine, 140
Surrogate loss function, 276
SVD, *see* singular value decomposition
Symmetric matrix, 41, 43

Tangent distance, 270
Tangent plane, 518
Tangent prop, 270
TDNN, *see* time-delay neural network
Teacher forcing, 382, 383
Tempering, 605
Template matching, 141
Tensor, xi, xii, 33
Test set, 110
Tikhonov regularization, *see* weight decay
Tiled convolution, 352
Time-delay neural network, 368, 374
Toeplitz matrix, 333
Topographic ICA, 495
Trace operator, 46
Training error, 110
Transcription, 101
Transfer learning, 538
Transpose, xii, 33

Triangle inequality, 39
Triangulated graph, *see* chordal graph
Trigram, 464

Unbiased, 124
Undirected graphical model, 77, 509
Undirected model, 568
Uniform distribution, 57
Unigram, 464
Unit norm, 41
Unit vector, 41
Universal approximation theorem, 197
Universal approximator, 555
Unnormalized probability distribution, 569
Unsupervised learning, **105**, 146
Unsupervised pretraining, 461, 530

V-structure, *see* explaining away
V1, 365
VAE, *see* variational autoencoder
Vapnik-Chervonenkis dimension, 114
Variance, xiii, 61, 229
Variational autoencoder, 691, **698**
Variational derivatives, *see* functional derivatives
Variational free energy, *see* evidence lower bound
VC dimension, *see* Vapnik-Chervonenkis dimension
Vector, xi, xii, 32
Virtual adversarial examples, 269
Visible layer, 6
Volumetric data, 360

Wake-sleep, 653, 663
Weight decay, **118**, 177, 231, 433
Weight space symmetry, 284
Weights, 15, 107
Whitening, 457
Wikibase, 485
`Wikibase`, 485
Word embedding, 466
Word-sense disambiguation, 486
`WordNet`, 485

Zero-data learning, *see* zero-shot learning
Zero-shot learning, 540