Index

L^p norm, 35
k-means, 321, 512
k-nearest neighbors,
boldindex136, 512
n-gram,
boldindex406
Freebase, 430
GeneOntology, 430
Wikibase, 430
WordNet, 430
0-1 loss,
boldindex97, 224, 241

, 322

Absolute value rectification, 165
Active constraint, 90
Adagrad, 266
ADALINE, *see* Adaptive Linear Element
Adaptive Linear Element, 13, 21, 24
Adversarial example, 238
Affine, 102
AIS, *see* annealed importance sampling
Almost everywhere, 68
Ancestral sampling, 462
ANN, *see* Artificial neural network
Annealed importance sampling, 559, 610
Approximate inference, 456
Artificial intelligence, 1
Artificial neural network, *see* Neural network
Asymptotically unbiased, 116
Audio, 317
Autoencoder, 4
Automatic differentiation, 185
Back-propagation, 175
Back-Propagation Through Time, 336
Bagging, 230
Batch gradient descent, 259
Bayes error,
boldindex109
Bayes' rule, 67
Bayesian hyperparameter optimization, 382
Bayesian network, *see* directed graphical model
Bayesian probability, 50
Bayesian statistics,
boldindex127
Beam Search, 583
Beam search, 570
Belief network, *see* directed graphical model
Bernoulli distribution, 59
Bias, 116
Bigram, 408
Binary relation, 429
Boltzmann distribution, 445
Boltzmann machine, 445
Boltzmann Machines, 590
BPTT, *see* Back-Propagation Through Time

CAE, *see* contractive auto-encoder
Calculus of variations, 586
Categorical distribution, see multinoulli distribution60
CD, *see* contrastive divergence
Centering trick (DBM), 614
Central limit theorem, 62
Chain rule of probability, 54

Chess, 2
Chord, 451
Chordal graph, 451
Class-based language models, 409
Classical dynamical system, 331
Classical regularization, 204
Classification, 93
Cliffs, 252
Clipping the gradient, 368
Clique potential, *see* factor (graphical model)
CNN, *see* convolutional neural network
Collaborative Filtering, 425
Collider, *see* explaining away
Color images, 317
Computer vision, 396
Concept drift, 503
Conditional computation, *see* dynamic structure
Conditional independence, xi, 55
Conditional probability, 53
Connectionism, 15, 388
Connectionist temporal classification, 570
Consistency, 123
Constrained optimization, 88
Content-based addressing, 365
Content-Based Recommender Systems, 427
Context-specific independence, 448
Contextual Bandits, 427
Continuation methods, 292, 294
Contractive auto-encoder, 490, 539
Contractive autoencoders, 470
Contrast, 398
Contrastive divergence, 545, 610, 613
Convolution, 296, 617
Convolutional network, 14
Convolutional neural network, 228, boldindex296
Coordinate descent, 283, 613
Correlation, 56
Cost function, *see* objective function
Covariance, xi, 55
Covariance matrix, 56
Cross entropy, boldindex59, 166
Cross-correlation, 298
Cross-validation, 114
CTC, *see* connectionist temporal classification
Curriculum learning, 294
curse of dimensionality, 145
Cyc, 2

D-separation, 447
DAE, *see* denoising auto-encoder
Data generating distribution, boldindex103, 124
Data generating process, 103
Data parallelism, 391
Dataset, 98
Dataset augmentation, 398, 401
DBM, *see* deep Boltzmann machine
Decision tree, boldindex138
Decision trees, 512
Decoder, 4
Deep belief network, 24, 564, 591, 600, 618
Deep Blue, 2
Deep Boltzmann machine, 21, 24, 564, 591, 603, 613, 618
Deep learning, 1, 5
Denoising auto-encoder, 485
Denoising autoencoders, 190
Denoising score matching, 554
Density estimation, 96
Derivative, xi, 79
Design matrix, boldindex99
Detector layer, 306
Deterministic gradient descent, 259
Diagonal matrix, 37
Dirac delta function, 63
Directed graphical model, 69, 439
Directional derivative, 83
Distributed Representation, 511
Distributed representation, 15
domain adaptation, 501
Dot product, 31
Double exponential distribution, *see* Laplace distribution
Doubly block circulant matrix, 300

Dream sleep, 544, 589
DropConnect, 235
Dropout, 190, *232*, 377, 378, 613
Dynamic structure, 393

E-step, 567
Early stopping, 220, 221, 223--225
EBM, *see* energy-based model
Echo state network, 21, 24, 354
Effective number of parameters, 208
Efficiency, 127
Eigendecomposition, 38
Eigenvalue, 38
Eigenvector, 38
ELBO, *see* evidence lower bound
Element-wise product, *see* Hadamard product, *see* Hadamard product
EM, *see* expectation maximization
Embedding, 526
Empirical distribution, 63
Empirical risk, 241
Empirical risk minimization, 241
Encoder, 4
Energy function, 445
Energy-based model, 445, 603
Ensemble methods, 230
Epoch, 220
Equality constraint, 89
Equivariance, 303
Error function, *see* objective function
ESN, *see* echo state network
Euclidean norm, 35
Euler-Lagrange equation, 586
Evidence lower bound, 566--569, 602
Example, 98
Excess error, 259
Expectation, 55
Expectation maximization, 567
Expected value, *see* expectation
Explaining away, 449
Exploitation, 428
Exploration, 428
Exponential distribution, boldindex63

Factor (graphical model), 442
Factor analysis, 475
Factor graph, 453
Factors of variation, 4
Feature, 98
Feedforward deep network, 158
Fine-tuning, 289
Finite differences, 386
Forward-Backward algorithm, 570
Fourier transform, 317, 320
Fovea, 323
Freebase, 430
Frequentist probability, 50
Frequentist statistics, boldindex127
Functional derivatives, 586

Gabor function, 325
Gaussian distribution, see Normal distribution(
Gaussian kernel, 135
Gaussian mixture, 64
GCN, *see* Global contrast normalization
Generalization, 103
Generalized Lagrange function, *see* Generalized Lagrangian
Generalized Lagrangian, 89
Generative adversarial networks, 190
Gibbs distribution, 443
Gibbs sampling, 463
Global contrast normalization, 398
GPU, *see* Graphics processing unit
Gradient, 83
Gradient clipping, 368
Gradient clipping, 254
Gradient descent, 83
Graph, x
Graph Transformer, 580
Graph transformer, 577
Graphical model, *see* structured probabilistic model
Graphics processing unit, 389
Greedy algorithm, 289
Greedy layer-wise unsupervised pre-training, 494
Greedy supervised pre-training, 289
Grid search, 379

Hadamard product, x, 31

Hard tanh, 165
Harmonium, see Restricted Boltzmann machine459
Harmony theory, 446
Helmholtz free energy, *see* evidence lower bound
Hessian matrix, xi, 84, 271
Hidden layer, 6
Hidden Markov model, 574
HMM, *see* hidden Markov model
Hyperbolic tangent, 164
Hyperparameter optimization, 379
Hyperparameters, 113, 377
Hypothesis space, 104, 110

i.i.d assumptions, 236
i.i.d., 115
i.i.d. assumptions, 103
Identity matrix, 32
Immorality, 451
Independence, xi, 54
Independent and identically distributed, 115
Independent component analysis, 476
Inequality constraint, 89
Inference, 438, 456, 564, 566--568, 584, 586, 588
Information retrieval, 424
Initialization, 283
Integral, xi
Invariance, 306
Isomap, 498

Jacobian matrix, xi, 69, 83
Joint probability, 52

Karush-Kuhn-Tucker conditions, 90
Karush{Kuhn{Tucker, 89
Kernel (convolution), 297, 298
Kernel machine, 512
Kernel trick, 135
KKT, *see* Karush{Kuhn{Tucker
KKT conditions, *see* Karush-Kuhn-Tucker conditions
KL divergence, *see* Kllback-Leibler divergence59
Knowledge base, 2
Kullback-Leibler divergence, xi, boldindex59

Lagrange multipliers, 89, 90, 587
Lagrangian, *see* Gneralized Lagrangian89
Laplace distribution, boldindex63
Latent variable, 64, 471
LCN, *see* local contrast normalization
Leaky units, 357
Learning Relations, 429
Line search, 83
Linear combination, 33
Linear convergence, 259
Linear dependence, 34
Linear factor models, 474
Linear regression, boldindex100, 102, 134
Link Prediction, 430
Liquid state machine, 354
Local conditional probability distribution, 439
Local contrast normalization, 400
Logistic regression, 2, 135
Logistic sigmoid, 7, 65
Long short-term memory, 359
Loop, 451
Loss function, *see* objective function
LSTM, 23, *see* lng short-term memory359

M-step, 567
Machine learning, 2
Main diagonal, 30
Manifold, 152
Manifold hypothesis, 523
Manifold hypothesis, 154
Manifold learning, 152, 523
Manifold Tangent Classifier, 539
MAP inference, 569
Marginal probability, 53
Markov chain, 462, 571
Markov network, see undirected model441
Markov property, 571
Markov random field, see undirected model441
Matrix, ix, x, 29

Matrix inverse, 32
Matrix product, 30
Max pooling, 306
Maximum likelihood, boldindex124
Maxout, 165
Mean field, 610, 613
Mean squared error, 101
Measure theory, 67
Measure zero, 68
Memory network, 364
Memory Networks, 362
Method of steepest descent, *see* gradient descent
Minibatch, 144, 244
Missing inputs, 93
Mixing (Markov chain), 464
Mixture distribution, 64
Mixture of experts, 394, 512
MLP, *see* multilayer perception
MNIST, 19, 20, 613
Model averaging, 230
Model compression, 392
Model identifiability, 248
Model parallelism, 391
Moore-Penrose Pseudoinverse, 41
Moore-Penrose pseudoinverse, 214
Moralized graph, 451
MP-DBM, *see* multi-prediction DBM
MRF (Markov Random Field), see undirected model441
MSE, see mean squared error101
Multi-modal learning, 507
Multi-prediction DBM, 612, 614
Multi-task learning, 235, 503
Multilayer perception, 5
Multilayer perceptron, 24, boldindex158
Multinomial distribution, 60
Multinoulli distribution, 60

Naive Bayes, 2, 73
Nat, 57
natural image, 435
Nearest neighbor regression, boldindex106
Negative definite, 84
Negative phase, 542, 544
Neocognitron, 14, 21, 24
Nesterov momentum, 264
Netflix Grand Prize, 232
Neural network, 12
Neural Turing machine, 364
Neuroscience, 13
Noise-contrastive estimation, 554
Non-parametric model, boldindex106
Norm, xii, 35
Normal distribution, 60, 62
Normal equations, boldindex102, 102, 104, 208
Normalized initialization, 286
Numerical differentiation, 185, *see* finite differences

Object detection, 396
Object recognition, 396
Objective function, 79
Offset, 160
One-shot learning, 504
Orthodox statistics, *see* frequentist statistics
Orthogonal matrix, 37
Orthogonality, 37

Parallel distributed processing, 15
Parameter initialization, 283
Parameter sharing, 301
Parameter tying , Parameter sharing227
Parametric model, boldindex106
Partial derivative, 82
Partition function, 444, 540, 610
PCA, *see* principal components analysis
PCD, *see* stochastic maximum likelihood
Perceptron, 13, 24
Perplexity, 126
Persistent contrastive divergence, *see* stochastic maximum likelihood
Point Estimator, 115
Policy, 428
Pooling, 296, 617

Positive definite, 84
Positive phase, 542, 544
Pre-training, 289, 494
Precision (of a normal distribution), 62
Predictive sparse decomposition, 321, 469, 481, 483
Preprocessing, 397
Primary visual cortex, 322
Principal components analysis, 43, 475, 564
Principle components analysis, 140--142, 154
Prior probability distribution, boldindex127
Probabilistic max pooling, 617
Probability density function, 52
Probability distribution, 51
Probability function estimation, 96
Probability mass function, 51
Product rule of probability, *see* chain rule of probability
PSD, *see* predictive sparse decomposition
Pseudolikelihood, 550

Quadrature pair, 327
Quasi-Newton condition, 278
Quasi-Newton methods, 278

Radial basis function, 165
Random search, 381
Random variable, 51
Ratio matching, 553
RBF, 165
RBM, *see* restricted Boltzmann machine
Receptive field, 302
Recommender Systems, 424
Rectified linear unit, 164
Rectifier, 164
Recurrent network, 24
Recurrent neural network, 333
Recursive neural networks, 331
Regression, 94
Regularization, boldindex112, 112, 201, 377
Reinforcement Learning, 99, 427
Reinforcement learning, 190
ReLU, 164
Representation learning, 3
Restricted Boltzmann machine, 459, 564, 591, 593, 613, 614, 616, 617
Ridge regression, see weight decay205
Risk, 241

Saddle points, 249
Sample mean, 117
Scalar, ix, x, 28
Score matching, 552
Secant condition, 278
Second derivative, 83
Second derivative test, 84
Self-information, 57
Semantic hashing, 424
Semi-supervised learning, 508
Separable convolution, 320
Separation (probabilistic modeling), 447
Set, x
SGD, *see* stochastic gradient descent
Shannon entropy, xi, 57, 586
Shortlist, 412
Sigmoid, xii, *see* logistic sigmoid, 164
Sigmoid belief network, 24
Simple cell, 322
Simulated annealing, 294
Singular value, *see* singular value decompositio
Singular value decomposition, 40, 141, 142
Singular vector, *see* singular value decompositi
SML, *see* stochastic maximum likelihood
Softmax, 164, 168
Softplus, xii, 65, 165
Spam detection, 2
Sparse coding, 469, 478, 564
Sparse initialization, 287
Sparse representations, 228, 480
Spearmint, 382
spectral radius, 354
Speech recognition, 401
Sphering, *see* Whitening, 399
Spike and slab restricted Boltzmann machine, 616

Square matrix, 34
ssRBM, *see* spike and slab restricted Boltzmann machine
Standard deviation, 55
Standard error, 119
Statistic, 115
Statistical learning theory, 103
Steepest descent, *see* gradient descent
Stochastic gradient descent, 13, boldindex144, 244, boldindex260, 613
Stochastic maximum likelihood, 546, 610, 613
Stochastic pooling, 235
Structure learning, 455
Structured output, 95
Structured probabilistic model, 69, 434
Student-t, 470
Sum rule of probability, 53
Sum-product network, 518
Supervised learning, boldindex98
Support vector machine, 135
Surrogate loss function, 242
SVD, *see* singular value decomposition
Symbolic differentiation, 186
Symmetric matrix, 37, 40

t-SNE, 498
Tangent Distance, 536
Tangent plane, 526
Tangent-Prop, 537
Tanh, 164
TDNN, *see* time-delay neural network
Teacher forcing, 335
Tensor, ix, x, 30
Test set, 103
Tikhonov regularization, *see* weight decay
Tiled convolution, 313
Time-delay neural network, 324, 330
Time-delay neural networks, 356
Toeplitz matrix, 300
Trace operator, 42
Training error, 103
Transcription, 94
Transfer learning, 501
Translation, 94
Transpose, x, 30
Triangle inequality, 35
Triangulated graph, *see* chordal graph
Trigram, 408

Unbiased, 116
Undirected graphical model, 69
Undirected model, 441
Uniform distribution, 52
Unigram, 408
Unit norm, 37
Unit vector, 37
Universal approximation theorem, 192
Universal approximator, 517
Unnormalized probability distribution, 443
Unsupervised learning, boldindex98, 139
Unsupervised pre-training, 494

V-structure, *see* explaining away
V1, 322
Variance, xi, 55
Variational autoencoder, 190
Variational derivatives, *see* functional derivatives
Variational free energy, *see* evidence lower bound
Vector, ix, x, 29
Visible layer, 6
Viterbi algorithm, 570
Viterbi decoding, 574
Volumetric data, 317

Weight decay, 111, boldindex205, 378
Weight space symmetry, 248
Weights, 13, 100
Whitening, 399
Wikibase, 430
Word-Sense Disambiguation, 431

Zero-data learning, 504
Zero-shot learning, 504