

Index

- L^p norm, 35
- k -means, 296, 487
- k -nearest neighbors,
 boldindex135, 487
- 0-1 loss, 97
- , 50, 298
- Absolute value rectification, 162
- Active constraint, 90
- Adagrad, 254
- ADALINE, *see* Adaptive Linear Element
- Adaptive Linear Element, 13, 21, 24
- Adversarial example, 232
- Affine, 102
- AIS, *see* annealed importance sampling
- Almost everywhere, 67
- Ancestral sampling, 437
- ANN, *see* Artificial neural network
- Annealed importance sampling, 533, 569
- Approximate inference, 430
- Artificial intelligence, 1
- Artificial neural network, *see* Neural network
- Asymptotically unbiased, 115
- Audio, 293
- Autoencoder, 4
- Automatic differentiation, 182
- Back-propagation, 172
- Back-Propagation Through Time, 311
- Bagging, 223
- Bayes error,
 boldindex108
- Bayes' rule, 66
- Bayesian hyperparameter optimization, 369
- Bayesian network, *see* directed graphical model
- Bayesian statistics,
 boldindex126
- Beam Search, 359
- Beam search, 347
- Belief network, *see* directed graphical model
- Bernoulli distribution, 59
- Bias, 115
- Boltzmann distribution, 420
- Boltzmann machine, 420
- Boltzmann Machines, 549
- BPTT, *see* Back-Propagation Through Time
- CAE, *see* contractive auto-encoder
- Calculus of variations, 545
- Categorical distribution, *see* multinoulli distribution60
- CD, *see* contrastive divergence
- Centering trick (DBM), 573
- Central limit theorem, 62
- Chain rule of probability, 54
- Chess, 2
- Chord, 426
- Chordal graph, 426
- Classical dynamical system, 306
- Classical regularization, 197
- Classification, 93
- Cliffs, 241
- Clipping the gradient, 342
- Clique potential, *see* factor (graphical model)
- CNN, *see* convolutional neural network
- Collider, *see* explaining away
- Color images, 293
- Computer vision, 381
- Concept drift, 478

- Conditional computation, *see* dynamic structure
- Conditional independence, vi, 55
- Conditional probability, 53
- Connectionism, 15, 373
- Connectionist temporal classification, 346
- consistency, 122
- Constrained optimization, 88
- Context-specific independence, 423
- Continuation methods, 269
- Contractive auto-encoder, 465, 514
- Contractive autoencoders, 445
- Contrast, 382
- Contrastive divergence, 520, 569, 572
- Convolution, 272, 576
- Convolutional network, 14
- Convolutional neural network, 222, **boldindex**272
- Coordinate descent, 261, 572
- Correlation, 56
- Cost function, *see* objective function
- Covariance, vi, 55
- Covariance matrix, 56
- Cross entropy, **boldindex**59, 163
- Cross-correlation, 274
- Cross-validation, 113
- CTC, *see* connectionist temporal classification
- Curriculum-learning, 271
- curse of dimensionality, 143
- Cyc, 2
- D-separation, 422
- DAE, *see* denoising auto-encoder
- Data generating distribution, **boldindex**103, 122
- Data generating process, 103
- Data parallelism, 376
- Dataset, 97
- Dataset augmentation, 382, 387
- DBM, *see* deep Boltzmann machine
- Decision tree, **boldindex**136
- Decision trees, 487
- Decoder, 4
- Deep belief network, 24, 538, 550, 559, 577
- Deep Blue, 2
- Deep Boltzmann machine, 21, 24, 538, 550, 562, 572, 577
- Deep learning, 1, 5
- Denoising auto-encoder, 459
- Denoising autoencoders, 186
- Denoising score matching, 528
- Density estimation, 96
- Derivative, vi, 79
- Design matrix, **boldindex**99
- Detector layer, 280
- Diagonal matrix, 36
- Dirac delta function, 63
- Directed graphical model, 69, 414
- Directional derivative, 83
- Distributed Representation, 486
- Distributed representation, 15
- domain adaptation, 476
- Dot product, 31
- Double exponential distribution, *see* Laplace distribution
- Doubly block circulant matrix, 276
- Dream sleep, 519, 548
- DropConnect, 229
- Dropout, 186, 226, 364, 365, 572
- Dynamic structure, 378
- E-step, 541
- Early stopping, 215–219
- EBM, *see* energy-based model
- Echo state network, 21, 24, 331
- Effective number of parameters, 201
- Efficiency, 125
- Eigendecomposition, 37
- Eigenvalue, 38
- Eigenvector, 38
- ELBO, *see* evidence lower bound
- Element-wise product, *see* Hadamard product, *see* Hadamard product
- EM, *see* expectation maximization
- Embedding, 502
- Empirical distribution, 63
- Empirical risk, 235
- Empirical risk minimization, 235

- Encoder, 4
- Energy function, 420
- Energy-based model, 420, 562
- Ensemble methods, 223
- Equality constraint, 89
- Equivariance, 279
- Error function, *see* objective function
- ESN, *see* echo state network
- Euclidean norm, 35
- Euler-Lagrange equation, 546
- Evidence lower bound, 540–543, 561
- Example, 97
- Expectation, 55
- Expectation maximization, 541
- Expected value, *see* expectation
- Explaining away, 424
- Exponential distribution,
 - [boldindex63](#)
- Factor (graphical model), 417
- Factor analysis, 450
- Factor graph, 428
- Factors of variation, 4
- Feature, 97
- Feedforward deep network, 155
- Finite differences, 372
- Forward-Backward algorithm, 347
- Fourier transform, 293, 295
- Fovea, 299
- Frequentist probability, 50
- Frequentist statistics,
 - [boldindex126](#)
- Functional derivatives, 545
- Gabor function, 300
- Gaussian distribution, *see* Normal distribution
 - [60](#)
- Gaussian kernel, 134
- Gaussian mixture, 64
- GCN, *see* Global contrast normalization
- Generalization, 102
- Generalized Lagrange function, *see* Generalized Lagrangian
- Generalized Lagrangian, 89
- Generative adversarial networks, 186
- Gibbs distribution, 418
- Gibbs sampling, 438
- Global contrast normalization, 383
- GPU, *see* Graphics processing unit
- Gradient, 83
- Gradient clipping, 342
- Gradient descent, 83
- Graph, v
- Graph Transformer, 356
- Graph transformer, 353
- Graphical model, *see* structured probabilistic model
- Graphics processing unit, 374
- Greedy layer-wise unsupervised pre-training,
 - [469](#)
- Grid search, 364
- Hadamard product, v, 31
- Hard tanh, 162
- Harmonium, *see* Restricted Boltzmann machine
 - [433](#)
- Harmony theory, 421
- Helmholtz free energy, *see* evidence lower bound
- Hessian matrix, vi, 84, 259
- Hidden layer, 6
- Hidden Markov model, 305
- HMM, *see* hidden Markov model
- Hyperbolic tangent, 161
- Hyperparameter optimization, 364
- Hyperparameters, 112, 362, 364
- Hypothesis space, 104, 110
- i.i.d assumptions, 232
- i.i.d., 114
- i.i.d. assumptions, 103
- Identity matrix, 32
- Immorality, 426
- Independence, vi, 54
- Independent and identically distributed, 114
- Independent component analysis, 451
- Inequality constraint, 89
- Inference, 413, 430, 538, 540–544, 547
- Initialization, 262
- Integral, vi
- Invariance, 280
- Isomap, 473

- Jacobian matrix, vi, 68, 83
- Joint probability, 52
- Karush-Kuhn-Tucker conditions, 90
- Karush-Kuhn-Tucker, 88
- Kernel (convolution), 273, 274
- Kernel machine, 487
- Kernel trick, 133
- KKT, *see* Karush-Kuhn-Tucker
- KKT conditions, *see* Karush-Kuhn-Tucker conditions
- KL divergence, *see* Kullback-Leibler divergence59
- Knowledge base, 2
- Kullback-Leibler divergence, vi, **boldindex**59
- Lagrange multipliers, 88, 90, 546
- Lagrangian, *see* Generalized Lagrangian89
- Laplace distribution, **boldindex**63
- Latent variable, 446
- LCN, *see* local contrast normalization
- Leaky units, 334
- Line search, 83
- Linear combination, 33
- Linear dependence, 34
- Linear factor models, 449
- Linear regression, **boldindex**100, 102, 132
- Liquid state machine, 331
- Local conditional probability distribution, 414
- Local contrast normalization, 384
- Logistic regression, 2, 133
- Logistic sigmoid, 7, 65
- Long short-term memory, 335
- Loop, 426
- Loss function, *see* objective function
- LSTM, 22, *see* long short-term memory335
- M-step, 541
- Machine learning, 2
- Main diagonal, 30
- Manifold, 150
- Manifold hypothesis, 498
- Manifold hypothesis, 151
- Manifold learning, 150, 498
- Manifold Tangent Classifier, 513
- MAP inference, 543
- Marginal probability, 53
- Markov chain, 348, 437
- Markov network, *see* undirected model416
- Markov property, 348
- Markov random field, *see* undirected model416
- Matrix, iv, v, 29
- Matrix inverse, 32
- Matrix product, 30
- Max pooling, 280
- Maximum likelihood, **boldindex**122
- Maxout, 162
- Mean field, 569, 572
- Mean squared error, 101
- Measure theory, 67
- Measure zero, 67
- Method of steepest descent, *see* gradient descent
- Missing inputs, 93
- Mixing (Markov chain), 439
- Mixture distribution, 64
- Mixture of experts, 487
- MLP, *see* multilayer perception
- MNIST, 18, 19, 572
- Model averaging, 223
- Model capacity, 363
- Model compression, 377
- Model parallelism, 376
- Moore-Penrose Pseudoinverse, 41
- Moore-Penrose pseudoinverse, 207
- Moralized graph, 426
- MP-DBM, *see* multi-prediction DBM
- MRF (Markov Random Field), *see* undirected model416
- MSE, *see* mean squared error101
- Multi-modal learning, 482
- Multi-prediction DBM, 571, 573
- Multi-task learning, 230, 478
- Multilayer perception, 5
- Multilayer perceptron, 24, **boldindex**155
- Multinomial distribution, 60
- Multinoulli distribution, 60

- Naive Bayes, 2, 71
- Nat, 57
- natural image, 410
- Nearest neighbor regression, **boldindex107**
- Negative definite, 84
- Negative phase, 517, 519
- Neocognitron, 14, 21, 24
- Nesterov momentum, 252
- Netflix Grand Prize, 226
- Neural network, 12
- Neuroscience, 13
- Noise-contrastive estimation, 529
- Non-parametric model, **boldindex107**
- Norm, vii, 35
- Normal distribution, 60, 62
- Normal equations, **boldindex101, 102, 104, 201**
- Normalized initialization, 264
- Numerical differentiation, 182, *see* finite differences

- Object detection, 381
- Object recognition, 381
- Objective function, 79
- Offset, 157
- One-shot learning, 480
- Orthodox statistics, *see* frequentist statistics
- Orthogonal matrix, 37
- Orthogonality, 37
- Overfitting, 363

- Parallel distributed processing, 15
- Parameter initialization, 262
- Parameter sharing, 277
- Parameter tying , Parameter sharing**221**
- Parametric model, **boldindex107**
- Partial derivative, 82
- Partition function, 419, 515, 569
- PCA, *see* principal components analysis
- PCD, *see* stochastic maximum likelihood
- Perceptron, 13, 24
- Perplexity, 125

- Persistent contrastive divergence, *see* stochastic maximum likelihood
- Point Estimator, 114
- Pooling, 272, 576
- Positive definite, 84
- Positive phase, 517, 519
- Pre-training, 469
- Precision (of a normal distribution), 62
- Predictive sparse decomposition, 296, 444, 456, 458
- Preprocessing, 381
- Primary visual cortex, 297
- Principal components analysis, 43, 386, 450, 538
- Principle components analysis, 138–140, 152
- Prior probability distribution, **boldindex126**
- Probabilistic max pooling, 576
- 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, 524

- Quadrature pair, 301

- Radial basis function, 161
- Random search, 366
- Random variable, 51
- Ratio matching, 528
- RBF, 161
- RBM, *see* restricted Boltzmann machine
- Receptive field, 277
- Rectified linear unit, 161
- Rectifier, 161
- Recurrent network, 24
- Recurrent neural network, 308
- Recursive neural networks, 306
- Regression, 94
- Regularization, **boldindex111, 111, 194, 364**
- Reinforcement learning, 186
- ReLU, 161

- Representation learning, 3
- Restricted Boltzmann machine, 433, 538, 550, 552, 572, 573, 575, 576
- Ridge regression, *see* weight decay198
- Risk, 235

- Sample mean, 116
- Scalar, iv, v, 28
- Score matching, 527
- Second derivative, 83
- Second derivative test, 84
- Self-information, 57
- Semi-supervised learning, 483
- Separable convolution, 295
- Separation (probabilistic modeling), 422
- Set, v
- SGD, *see* stochastic gradient descent
- Shannon entropy, vi, 57, 546
- Sigmoid, vii, *see* logistic sigmoid, 161
- Sigmoid belief network, 24
- Simple cell, 298
- Simulated annealing, 269
- Singular value, *see* singular value decomposition
- Singular value decomposition, 40, 139, 140
- Singular vector, *see* singular value decomposition
- SML, *see* stochastic maximum likelihood
- Softmax, 161, 165
- Softplus, vii, 65, 162
- Spam detection, 2
- Sparse coding, 444, 453, 538
- Sparse initialization, 265
- Sparse representations, 222, 455
- Spearmint, 369
- spectral radius, 332
- Speech recognition, 388
- Sphering, *see* Whitening, 384
- Spike and slab restricted Boltzmann machine, 575
- Square matrix, 34
- ssRBM, *see* spike and slab restricted Boltzmann machine
- Standard deviation, 55
- Statistic, 114
- Statistical learning theory, 103
- Steepest descent, *see* gradient descent
- Stochastic gradient descent, 13, 237, boldindex249, 572
- Stochastic maximum likelihood, 521, 569, 572
- Stochastic pooling, 230
- Structure learning, 430
- Structured output, 94
- Structured probabilistic model, 69, 409
- Student-t, 445
- Sum rule of probability, 53
- Sum-product network, 493
- Supervised learning, boldindex98
- Support vector machine, 133
- Surrogate loss function, 235
- SVD, *see* singular value decomposition
- Symbolic differentiation, 182
- Symmetric matrix, 37, 40

- t-SNE, 473
- Tangent Distance, 511
- Tangent plane, 502
- Tangent-Prop, 512
- Tanh, 161
- TDNN, *see* time-delay neural network
- Teacher forcing, 310
- Tensor, iv, v, 30
- Test set, 103
- Tikhonov regularization, *see* weight decay
- Tiled convolution, 290
- Time-delay neural network, 300, 306
- Time-delay neural networks, 333
- Toeplitz matrix, 276
- Trace operator, 42
- Training error, 102
- Transcription, 94
- Transfer learning, 476
- Translation, 94
- Transpose, v, 30
- Triangle inequality, 35
- Triangulated graph, *see* chordal graph

- Unbiased, 115
- Undirected graphical model, 69
- Undirected model, 416

- Uniform distribution, 52
- Unit norm, 37
- Unit vector, 37
- Universal approximation theorem, 186
- Universal approximator, 492
- Unnormalized probability distribution, 417
- Unsupervised learning,
 - boldindex98**, 137
- Unsupervised pre-training, 469

- V-structure, *see* explaining away
- V1, 297
- Variance, vi, 55
- Variational autoencoder, 186
- Variational derivatives, *see* functional derivatives
- Variational free energy, *see* evidence lower bound
- Vector, iv, v, 29
- Visible layer, 6
- Viterbi algorithm, 347
- Viterbi decoding, 350
- Volumetric data, 293

- Weight decay, 110,
 - boldindex198**, 365
- Weights, 13, 100
- Whitening, 384, 386

- ZCA, *see* zero-phase components analysis
- Zero-data learning, 480
- Zero-phase components analysis, 386
- Zero-shot learning, 480