

## Numerics

---

1D convolutions 237  
2D grid, random walk on 24  
10 base estimators, bagging ensembles with 153

## A

---

active learning 138  
    Cramer-Rao lower bound 141–150  
    query strategies 139–141  
        query by committee 140–141  
        uncertainty sampling 139–140  
    variance reduction 141  
*Active Learning Literature Survey* (Settles) 139  
AdaBoost (adaptive boosting) 154  
AdaGrad 243  
Adam optimizer 245, 248  
ADASYN (adaptive synthetic sampling) 137  
adjusted Rand index (ARI) 166  
AGI (artificial general intelligence) 12  
algorithmic paradigms 61  
algorithms, definition and purpose 3  
amortized VI (variational inference) 68, 259–265  
    applications in MDNs 280  
    mixture density networks 259–265  
    overview of 259  
approximating distribution 10, 47  
ARI (adjusted Rand index) 166  
arrays, searching and sorting operations on 59  
artificial general intelligence (AGI) 12  
arXiv papers 285  
attention matrix heatmap 267  
attention mechanisms, transformer  
    architecture and 265–272

autoencoders 251–257  
    overview of 251–252  
    variational autoencoder anomaly detection  
        in time series 253–257

## B

---

backpropagation 6, 9, 221, 230  
backward LSTM model 233  
backward pass 130  
bagging (bootstrap aggregation) 151–154  
batch mode classification 273  
Bayes algorithms  
    naive 93–98  
    variational 188–195  
Bayesian inference  
    overview of 8–9  
    types of 9–11  
Bayesian nonparametric models 166, 213  
Bayesian optimization 148–150  
Bayesian regression  
    hierarchical 111–114  
    linear 107–111, 117  
Bellman equations 66  
Bernoulli Naive Bayes algorithm 93, 96  
Bernoulli RV (random variable) 16–17  
bias correction 245  
bidirectional LSTM architecture 233  
binary classification 74  
binary heap 60  
binary logistic regression 139  
binary random variable, mean of 50  
binary search 64  
binomial coefficients 65

binomial tree model 21–24  
 books 283–284  
 boosting 154–157  
 bootstrapping 138

## C

California house pricing dataset 111  
 Caltech 101 dataset 230  
 CART (classification and regression trees)  
   algorithm 64, 104  
 CDF (inverse cumulative density function)  
   method 10  
 Chow-Liu algorithm 201–202  
 CIFAR-100 dataset 242  
 cifar100 248  
 classification algorithms 73–105  
   decision trees 98–104  
   logistic regression 86–93  
   naive Bayes algorithm 93–98  
   overview of 74  
   perceptron algorithm 74–80  
   support vector machine 80–86  
 cluster centroids 134  
 clusters, DP (Dirichlet Process) K-means and 166  
 CNNs (convolutional neural networks)  
   9, 242, 279  
   LeNet architecture on MNIST dataset 226–229  
   ResNet image search 229–232  
 coin flips, posterior distribution of 16–18  
 competitive programming 283  
 complete search 61–62  
 computational biology, density estimation in 195  
 compute\_posterior function 118  
 computer vision (CV) 7, 279  
 confusion matrix 75  
 conjugate priors 9  
 context vector 265  
 Cramer-Rao lower bound 141–150  
 CV (computer vision) 7, 279  
 CVXOPT software package 82  
 cyclic permutation property 184

## D

data structures 58–61  
   linear 59  
   nonlinear 60–61  
   overview of 12  
   probabilistic 61  
 decision boundary 80  
 decision trees 98–104, 153  
 deep latent variable models (DLVM) 68

deep learning algorithms 219–281  
   amortized variational inference 259–265  
     mixture density networks 259–265  
     overview of 259  
   attention mechanisms and transformer  
     architecture 265–272  
   autoencoders 251–257  
     overview of 251–252  
     variational autoencoder anomaly detection  
       in time series 253–257  
   convolutional neural networks 225–242  
     LeNet architecture on MNIST dataset  
       226–229  
     ResNet image search 229–232  
   graph neural networks 273–279  
   machine learning algorithms research 279–280  
   modern algorithms 11–12  
   multilayer perceptron 220–224  
   neural network optimizers 242–248  
   recurrent neural nets 232–242  
     long short-term memory sequence  
       classification 233–236  
     multi-input model 237–242  
 deep learning architecture, evolution of 11  
 deep learning models 9  
 deep neural networks (DNNs) 9  
 dense layers, for computing probability of  
   duplicate questions 237  
 density estimation 195–200  
   kernel density estimator 195–198  
   tangent portfolio optimization 198–200  
 dimensionality reduction 179–184  
   principal component analysis 179–181  
   t-SNE manifold learning on images 182–184  
 Dirichlet distribution 184  
 divide and conquer 64–65  
 DLVM (deep latent variable models) 68  
 DNNs (deep neural networks) 9  
 DP (Dirichlet process) K-means 166–170  
 DP (dynamic programming) 65–67  
 DP-means algorithm 170  
 dropout 222, 233  
 duplicate questions, identifying 237  
 dynamically resizable array 59

## E

EI (expected improvement) 148  
 eigenvalue decomposition 180  
 eigenvectors 179  
 ELBO (Evidence Lower Bound) 43, 49, 51  
 EM (expectation-maximization) algorithm  
   171–179  
 empirical covariance matrix 179  
 Empirical variance 19

ENN (edited nearest neighbors) 138  
 ensemble methods 150–159  
   bagging 151–154  
   boosting 154–157  
   definition 150  
   stacking 157–159  
 entropy 57, 99, 140  
 estep function 174  
 Evidence Lower Bound (ELBO) 43, 49, 51  
 expected improvement (EI) 148  
 extremely randomized trees 154

## F

---

first-order Markov chain 18  
 Fisher information matrix 141  
 fit function 83, 109  
 fixed-size array 59  
 forward algorithm 129  
 forward KL 45  
 forward LSTM model 233  
 forward\_backward function 131  
 Forward-backward HMM algorithm 132  
 fully conditional distributions 28  
 fully factored approximation 47

## G

---

GAs (genetic algorithms) 210–213  
 gauss\_conditional function 29  
 Gaussian mixture models 16, 171–179  
   expectation-maximization algorithm 171–179  
   overview of 171  
 Gaussian processes (GPs) regression 117–121  
 GCNs (graph convolutional networks) 273  
 generative models 7, 279  
 Gibbs sampling 10, 28–31  
 gibbs\_gauss class 29  
 Gini index 99–100  
 GMM class 174  
 gmm\_em function 174  
 GMMs (Gaussian mixture models) 165, 171  
 GNNs (graph neural networks) 9, 273–279  
 GPs (Gaussian processes) regression 117–121  
 gradient descent 87, 108  
 gradient smoothing 245  
 graph lasso algorithm 206  
 greedy algorithms 62–64  
 grid search strategy 147

## H

---

handwritten digit classification 226  
 hash table 60

hierarchical Bayesian regression 111–114  
 hierarchical regression models 107  
 high-dimensional parameter spaces, sampling  
   from 14  
 hinge loss function 79  
 HMMs (hidden Markov models) 128–134  
 homogeneous ensembles 150  
 hyperparameter tuning 147–150

## I

---

image\_denoising class 52  
 images  
   image denoising in Ising models 49–56  
   ResNet image search 229–232  
   t-SNE manifold learning on 182–184  
 imbalanced learning 134–138  
   oversampling strategies 136–138  
   undersampling strategies 134–136  
 IMDb movie reviews, sentiment analysis of 233  
 importance sampling 15, 19, 39  
 importance weights 36  
 importance\_sampler class 36  
 information projection 45  
 inverse covariance estimation 202–206  
 IS (importance sampling) 35–41  
 Ising models, image denoising in 49–56

## J

---

Jensen's inequality 46  
 Joint distribution 87  
 joint posterior distribution 43  
 joint probability density 128

## K

---

K-means algorithm 174  
 k-means++ 184  
 Kaggle competition 237  
 KD tree 183  
 KDE (kernel density estimator) 195–198  
 Keras/TensorFlow  
   creating CNN Architecture with 226  
   image search using pretrained ResNet50  
     with 230  
   implementation 251  
   practical implementations using 253  
 kernel trick 120  
 kernel\_func function 118  
 kernels 82, 214  
 KL (Kullback-Leibler) divergence, variational  
   inference and 44–47  
 knapsack problem 62

KNN (K nearest neighbors) bagging  
     ensemble 153  
 KNN classifiers 153  
 KNN regression 115–117

## L

l2 regularization 233  
 Lagrangian objective function 81  
 latent states 128  
 LDA (latent Dirichlet allocation) 187–195  
     overview of 187–188  
     variational Bayes 188–195  
 Least confident definition 140  
 LeCun, Yann 226  
 LeNet architecture 226–229  
 leveraging model structure 68  
 likelihood function 117  
 linear classifier 74  
 linear data structures 59  
 linear regression 107  
 linked lists 59  
 log likelihood 86–87, 178  
 log partition function 45  
 loss functions 88, 156, 260  
 low-dimensional spaces, sampling from 10  
 LR (logistic regression) 86–93, 160  
 LSTM (long short-term memory) sequence  
     classification 233–236

## M

M-projection (moment projection) 46  
 machine learning algorithms. *See* ML algorithms  
 MAP (maximum a posteriori) estimates 171  
 maps 60  
 marginal likelihood 45  
 marginal probability 130  
 Markov chain for page rank 18–19  
 Markov models 124–134  
     hidden Markov models 128–134  
     page rank algorithm 125–128  
 masked attention 269  
 Max margin strategy 140  
 max-pooling operations 226, 237  
 maximum likelihood (ML) 171  
 MCMC (Markov chain Monte Carlo) 14–42  
     binomial tree model 21–24  
     comparison with variational inference 67  
     estimating  $\pi$  19–21  
     Gibbs sampling 28–31  
     importance sampling 35–41  
     Markov chain for page rank 18–19  
     Metropolis-Hastings sampling 32–35

    overview of 9–11, 15–19  
     posterior distribution of coin flips 16–18  
     self-avoiding random walk 24–28  
 MDNs (mixture density networks) 259–265  
 mean of binary random variable 50  
 mean-field approximation 47–48  
 mean-field factorization 11  
 mean-variance analysis 198  
 meta-classifier 157  
 meta-regressor 157  
 MH (Metropolis-Hastings) sampling 10, 32–35  
 mh\_gauss class 33  
 MI (mutual information) maximization 56–57  
 mixture density networks 12  
 ML (machine learning) algorithms 3–13  
     Bayesian inference  
       overview of 8–9  
       types of 9–11  
     deep learning 11–12  
     implementing 12–13  
       data structures 12  
       problem-solving paradigms 12–13  
     mathematical concepts 7–8  
     reasons to learn from scratch 7  
     research  
       deep learning 279–280  
       sampling methods and variational  
         inference 67–68  
       supervised learning algorithms 160–161  
       unsupervised learning 214  
     research conferences 285  
     types of 4–7  
     using divide and conquer paradigm 64  
 ML (maximum likelihood) 171  
 MLP (multilayer perceptron) 220–224  
 MNIST dataset, LeNet architecture on 226–229  
 model selection  
     Bayesian optimization 148–150  
     hyperparameter tuning 147–150  
 modern deep learning algorithms 11  
 momentum parameter 243  
 MRF (Markov random field) 49  
 MSE (mean squared error) 107  
 mstep function 174  
 multi-input model 237–242  
 multiclass classification 74  
 multidimensional distribution 28  
 multihead attention 268  
 multivariate Gaussian distribution 28, 41

## N

naive Bayes algorithm 93–98  
 narrow AI 12  
 natural language processing (NLP) 7, 280, 285

nearest neighbor portfolio weights 198  
 nearest neighbors approximation methods 116  
 Nesterov Momentum 243, 248  
 neural model capacity, increasing 248  
 neural network optimizers 242–248  
 NLL (negative log likelihood) 87  
 NLP (natural language processing) 7, 280, 285  
 NMI (normalized mutual information) 166  
 nonlinear data structures 60–61  
 nonnegative matrix factorization 214  
 normalizing constant 9

## O

objective function 6  
 Occam's razor principle 147  
 optimal latent state sequence 131  
 optimization algorithms 220  
 optimum approximating distribution 50  
 OSS (one-sided selection) method 135  
 output constraints 260  
 overfitting 222, 248  
 oversampling strategies 136–138

## P

page rank algorithm  
   Markov chain for 18–19  
   supervised learning 125–128  
 parallel ensemble methods 150  
 parallel processing 237  
 parallelizing Monte Carlo algorithms 67  
 PCA (principal component analysis)  
   138, 179–181, 230  
 perceptron algorithm 74–80, 160  
 Perceptron binary classifier confusion matrix 79  
 perceptron classifier 74  
 perceptron update rule 80  
 perplexity hyperparameter 184  
 PGM (probabilistic graphical models) 9  
 pi, estimating 19–21  
 pool-based sampling 139  
 pooled graphical models 111  
 positional encodings 268  
 posterior distribution 8–9, 16–18, 117  
 power method algorithm 126–127  
 precision-recall plot 76  
 predict function 83, 109  
 principal component analysis (PCA)  
   138, 179–181, 230  
 probabilistic data structures 61  
 probabilistic graphical models (PGM) 9, 68  
 problem-solving paradigms 61–67  
   complete search 61–62

  divide and conquer 64–65  
   dynamic programming 65–67  
   greedy 62–64  
   overview of 12–13

## Q

QBC (query by committee) 140–141  
 quadratic dual optimization program 81–82  
 query strategies 139–141  
   query by committee 140–141  
   uncertainty sampling 139–140  
 query\_by\_committee function 143  
 queues 59

## R

random forests 153  
 random oversampling 136  
 random search strategy 147  
 random undersampling 134  
 random walk Metropolis algorithm 35  
 random walk square distance 27  
 random walk, self-avoiding 24–28  
 RBF (radial basis function) kernel 82, 118, 121  
 reconstruction error 179  
 regression algorithms 106–122  
   Bayesian linear regression 107–111  
   Gaussian processes regression 117–121  
   hierarchical Bayesian regression 111–114  
   K nearest neighbors regression 115–117  
   overview of 107  
 regression models 107  
 regularized loss function 80  
 research conferences 284–285  
   computer vision 285  
   machine learning 285  
   natural language processing 285  
   theoretical computer science 285  
 ResNet image search 229–232  
 RFC (random forest classifier) 148  
 ridge\_reg class 109  
 RL (reinforcement learning) 66  
 RMSProp 244–245  
 RNA-seq data 196  
 RNA-Seq density estimate 198  
 RNNs (recurrent neural networks) 9, 232–242  
   long short-term memory sequence  
     classification 233–236  
   multi-input model 237–242  
 Robbins-Monro conditions 87  
 ROC (receiver operating characteristic) plot 76  
 run-time complexity 3

## S

---

SA (simulated annealing) 206–210  
 sampling methods, variational inference and 67–68  
 scalable machine learning 160  
 self-attention 268  
 self-avoiding random walk 24–28  
 self-supervised learning 6  
 semi-supervised technique in self-training 146  
 sensor placement 63  
 Seq2Seq 232, 265  
 Seq2Vec 232  
 sequence classification 219  
 sequence similarity 220  
 sequential data, encoding and decoding 232  
 sequential ensemble methods 150  
 Sequential Monte Carlo (SMC) 67  
 Settles, Burr 139  
 SGD (stochastic gradient descent) 87, 92, 242  
 sigmoid activation function 234, 237  
 similarity between objects, measuring 82  
 simulated annealing implementation 207  
 SMC (Sequential Monte Carlo) 67  
 SMOTE (synthetic minority oversampling technique) 136–138  
 softmax function 220  
 software implementation 58–69  
   data structures 58–61  
     linear 59  
     nonlinear 60–61  
     probabilistic 61  
   machine learning algorithms research 67–68  
   problem-solving paradigms 61–67  
     complete search 61–62  
     divide and conquer 64–65  
     dynamic programming 65–67  
     greedy 62–64  
 sparse precision matrix 203  
 spectral GCN 273  
 Spektral Keras/Tensorflow library 274  
 stacking 157–159  
 stacks 59  
 state transition matrix 18  
 stationary distribution 9, 125–126  
 stochastic gradient descent (SGD) 87, 92, 242  
 stochastic matrices 18  
 Stochastic Monte Carlo methods 67  
 stock clusters 203, 206  
 stream-based sampling 139  
 streaming Monte Carlo 67  
 structure learning 201–206  
   Chow-Liu algorithm 201–202  
   inverse covariance estimation 202–206  
   simulated annealing 206–210

supervised learning algorithms 123–162  
   active learning 138–147  
     Cramer-Rao lower bound 141–150  
     query strategies 139–141  
     variance reduction 141  
   ensemble methods 150–159  
     bagging 151–154  
     boosting 154–157  
     stacking 157–159  
   imbalanced learning 134–138  
     oversampling strategies 136–138  
     undersampling strategies 134–136  
   machine learning algorithms research 160–161  
 Markov models 124–134  
   hidden Markov models 128–134  
   page rank algorithm 125–128  
 model selection  
   Bayesian optimization 148–150  
   hyperparameter tuning 147–150  
 SVM (support vector machine) 80–86, 148, 160  
 SVM quadratic program 83  
 synthetic minority oversampling technique (SMOTE) 136–138

## T

---

t-SNE (t-distributed stochastic neighbor embedding) manifold learning 182–184  
 tangent portfolio optimization 198–200  
 Tomek links algorithm 134, 136  
 trace identity 174  
 tractable distribution 10  
 transformer architecture, attention mechanisms and 265–272  
 transition probability 124  
 tree-based models 98  
 true posterior 47  
 two-state Markov model 124

## U

---

UCB (upper confidence bound) 148  
 uncertainty sampling 139–140  
 underfitting 222  
 undersampling strategies 134–136  
 unsupervised learning algorithms 165–215  
   density estimation 195–200  
     kernel density estimator 195–198  
     tangent portfolio optimization 198–200  
   dimensionality reduction 179–184  
     principal component analysis 179–181  
     t-SNE manifold learning on images 182–184  
 Dirichlet process K-means 166–170

unsupervised learning algorithms (*continued*)

Gaussian mixture models 171–179

expectation-maximization algorithm 171–179

overview of 171

genetic algorithms 210–213

latent Dirichlet allocation 187–188

overview of 187–188

variational Bayes 188–195

machine learning algorithms research 214

structure learning 201–206

Chow-Liu algorithm 201–202

inverse covariance estimation 202–206

simulated annealing 206–210

**V**

VAE (variational autoencoder) anomaly detection

in time series 253–257

variance reduction 141

variational Bayes algorithm 188–195

variational free energy 45

Vec2Seq 232

VI (variational inference) 43–57

amortized 259–265

mixture density networks 259–265

overview of 259

image denoising in Ising models 49–56

KL divergence and 44–47

mean-field approximation 47–48

mutual information maximization 56–57

overview of 9–11

sampling methods and 67–68

Viterbi algorithm 131

viterbi function 131

**W**

weak learners, converting to strong learners 154

weight decay 222

wireless communications scenario 56