

## Index

- K*-means clustering, 280
- K*-medoids clustering, 282
- K*-nearest neighbors classifier, 243
- $\alpha$ -divergence, 422, 438
- $\alpha$ -log-loss, 72
- f*-divergences, 437
  
- activation function, 211
- active learning, 116, 123, 464, 496
- adversarial learning, 223
- adversarial training, 568
- adversarial variational Bayesian learning, 480
- aleatoric uncertainty, 115
- algorithm deficit, 11
- amortized variational inference, 369, 404
- ancestral data sampling, 225
- ancestral sampling, 237
- approximate Bayesian computation, 475
- Armijo's rule, 163
- artificial intelligence, 3
- assumed density filtering, 538
- attention, 105, 214, 513
- automatic differentiation, forward mode, 175
  
- backpropagation, 176
- bag-of-words model, 198, 279
- bagging, 240
- batch normalization, 216
- Bayes' theorem, 42
- Bayes-by-backprop, 477
- Bayesian active learning by disagreement, 496
- Bayesian continual learning, 536
- Bayesian inference, 55
- Bayesian learning, 455
- Bayesian meta-learning, 538
- Bayesian networks, 566
- Bayesian non-parametrics, 485
- Bayesian transfer learning, 535
- Bernoulli random variable, 15
  
- beta-Bernoulli model, 373
- bits-back coding, 497
- black-swan problem, 227
- Blahut-Arimoto algorithm, 432
- Boltzmann distribution, 347
- boosting, 240
- Bregman divergence, 360
  
- calibration, 453
- categorical random variable, 15
- Cauchy-Schwartz inequality, 24, 33, 128
- causality, 570
- chain rule of probability, 41
- classification, 95
- Claude Shannon, 4
- computational graph, 173
- concept shift, 507
- condition number, 188
- conformal prediction, 494
- conjugate exponential family, 371
- continual learning, 515
- contrastive density learning, 255
- contrastive density ratio learning, 258
- contrastive divergence, 276, 302
- contrastive representation learning, 269
- convex function, 144, 151
- convolutional neural networks, 242
- coreset, 517
- cosine similarity, 23
- covariance, 33
- covariance coefficient, 33
- Cramer-Rao bound, 351
- credible interval, 466
- cross entropy, 69
- cross-entropy loss, 71, 122
- curse of dimensionality, 243
  
- data fragmentation, 227
- data processing inequality, 81
- deep ensembling, 240
- deep kernels, 489
- deep learning, 5, 210
- deep sets, 242
  
- delta rule, 201
- density estimation, 251, 426
- detection-error loss, 57
- determinant, 29
- differential entropy, 75
- differential privacy, 551
- directional derivative, 148
- Dirichlet process, 489
- discriminative probabilistic models, 121
- domain knowledge, 5
- Donsker-Varadhan variational representation, 434
- double descent, 117
- doubly stochastic gradient, 405, 409
  
- eigendecomposition, 27
- elastic weight consolidation, 518
- ELBO, 288
- embedding, 279
- empirical Bayes, 482
- empirical Fisher information matrix, 468
- empirical risk minimization, 103
- energy-based models, 276
- entropy, 74
- epistemic uncertainty, 115
- error-correcting codes, 231
- expectation maximization, 285
- expectation propagation, 538
- expectation step, 289
- experience replay, 517
- explainable machine learning, 181
- exponential family, 333
- exponential family principal component analysis, 356
- exponential family, mean parameters, 338
- exponential family, minimal parametrization, 339
- exponential family, sufficient statistics, 333
- exponentiated gradient, 367

- fast gradient sign method, 569
- feature engineering, 196
- feature vector, 101
- federated averaging, 546
- federated learning, 541
- federated variational inference, 557
- fine-tuning, 526
- first-order optimality condition, 140
- Fisher information matrix, 349
- Fisher's identity, 274, 301
- free energy, 82, 369
- free energy principle, 297
  
- Gaussian radial basis function kernel, 128
- general AI, 5
- generalization error, 321
- generalization loss, 102
- generalized Bayesian federated learning, 555
- Generalized Bayesian learning, 491
- generalized conditional entropy, 91
- generalized entropy, 91
- generalized free energy, 424
- generalized Gauss–Newton approximation, 469
- generalized linear models, 209, 354
- generalized mutual information, 91
- generalized posterior distribution, 425, 491
- generalized variational expectation maximization, 424
- generalized variational inference, 423
- generative adversarial networks, 440
- generative probabilistic models, 121
- Gershgorin theorem, 27
- Gibbs posterior distribution, 425
- Gibbs sampling, 302
- Gibbs' inequality, 67
- global minima, 137
- good old fashioned AI, 4
- gradient descent, 153
- gradient vector, 146
- graph neural networks, 242
  
- hard predictor, 55
- Hebbian learning, 276
- hinge loss, 200
- hinge-at-zero loss, 200
- histogram, 252
- Hoeffding's inequality, 321
  
- Hopfield networks, 272
- hyperparameter, 101
  
- I-projection, 421
- importance sampling, 476, 508
- importance weighting, 304
- independent and identically distributed random variables, 41
- independent random variables, 40
- inductive bias, 7, 97
- inference, 53
- InfoMax, 428
- information bottleneck, 430
- information risk minimization, 328
- inner product, 22
- inner product gate, 180
- integral probability metrics, 443
- Internet of Things, 5
- invariance, 241
- inverse multi-quadric kernel, 128
- Ising model, 381
- iteration complexity, 524
  
- Jensen's inequality, 89
- Jensen–Shannon divergence, 70
- jointly Gaussian random vector, 309
  
- kernel density estimation, 254
- kernel density estimator, 402
- kernel function, 128, 254
- Kullback–Liebler divergence, 65
  
- Langevin Monte Carlo, 475
- Laplace approximation, 377, 467
- LASSO, 120
- law of iterated expectations, 45
- law of large numbers, 103, 124, 317
- law of total variance, 80
- learning rate, 155
- least squares problem, 104
- linear independent vectors, 25
- linear system, 28
- Lloyd–Max quantization, 253
- locality, 201
- log-density ratio, 66
- log-distribution ratio, 66
- log-likelihood, 122
- log-likelihood ratio, 66
- log-loss, 71, 122
- log-partition function, 274, 334
- logistic loss, 200
- logistic regression, 204
  
- M-projection, 421
- machine unlearning, 573
- majorization minimization, 290
- manifold, 280
- manifold assumption, 280
- MAP learning, 125
- map-reduce, 174
- marginal likelihood, 456
- marginal training log-loss, 459
- Markov chain Monte Carlo, 472
- Markov random fields, 382, 567
- matrix inversion lemma, 129
- maximum likelihood learning, 122
- maximum mean discrepancy, 443
- mean-field variational inference, 380
- mean matching, 345
- Mercer's theorem, 128
- message passing, 242, 384
- meta-overfitting, 534
- Metropolis adjusted Langevin algorithm, 475
- Metropolis–Hastings algorithm, 472
- mini-batch, 165
- minimum description length, 126, 496
- mirror descent, 366
- mixture models, 238
- model agnostic meta-learning, 530
- model class, 7
- model deficit, 11
- model-driven design, 5
- multi-head model, 512
- multi-layer feedforward neural networks, 210
- multi-sample ELBO, 305
- multi-task learning, 512
- Munro–Robbins, 170
- mutual information, 78
- mutual information neural estimator, 436
  
- nats, 67
- natural gradient descent, 353, 365
- neural networks, 210
- neurons, 213
- Newton's method, 164
- no-free-lunch theorem, 97
- noise contrastive estimation, 279
- noise outsourcing, 393
- normalizing flows, 258, 298
  
- Occam's razor, 482
- odds, 15

## 578 Index

- one-hot representation, 16
- online learning, 463
- optimal Bayesian predictor, 56
- out-of-distribution detection, 460
- outer product, 26
- overfitting, 108
  
- PAC Bayes, 327, 462
- polynomial kernel, 128
- pooling, 241
- positive semi-definite kernels, 128
- pre-conditioning, 164
- principal component analysis, 261
- principal component analysis
  - probabilistic, 294
- probabilistic graphical models, 259
- probably approximately correct
  - learning, 315
- probit regression, 356
- pruning, 213
  
- quadratic discriminant analysis, 228, 437
- quadratic functions, 151
- quantum machine learning, 571
  
- random walk Metropolis–Hastings, 473
- rate-distortion theory, 431
- recommendation system, 24, 271, 513, 522
- recurrent neural networks, 242
- regression, 95
- regularization, 118
- regularized ERM, 119
- REINFORCE gradient, 386
- rejection sampling, 470
- reproducing kernel Hilbert space, 127
- restricted Boltzmann machines, 272
- reverse KL divergence, 83
- ridge regression, 119
- right to erasure, 573
  
- saddle point, 140, 150
- sample complexity, 106, 317, 524
- sampling bias, 506
- score matching, 278
- score vector, 386
- self-supervised learning, 260
- semi-supervised learning, 225
- sequential Monte Carlo, 476
- sigmoid function, 204
- Simpson’s paradox, 570
- smooth function, 158
- soft predictor, 55
- softmax function, 232
- sparse dictionary learning, 266
- sparse kernel methods, 130
- sparsity, 267
- spike-and-slab prior distribution, 479
- stationary point, 140
- Stein variational gradient descent, 403
- stochastic average gradient, 190
- stochastic gradient Hamiltonian Monte Carlo, 475
- Stochastic gradient Langevin dynamics, 474
- stochastic gradient Markov chain Monte Carlo, 278, 474
- stochastic variance reduced gradient, 190
- strictly convex function, 146, 151
- subspace, 261
- successive convex approximation, 155
- sufficient statistics, 81
- support vector machines, 200, 209
- surprise, 72
  
- symmetric matrix, 27
- synapses, 213
  
- t-SNE, 299
- total variation distance, 438
- trace, 28
- training, 7
- training data, 9, 95
- transductive learning, 98
- transfer learning, 504
- two-sample estimators, 433
- type-II maximum likelihood, 482
  
- unadjusted Langevin algorithm, 475
- underfitting, 108
- underspecification, 570
- unnormalized distribution, 85
  
- validation, 109
- Vapnik–Chervonenkis dimension, 323
- variational autoencoders, 411
- variational continual learning, 537
- variational distribution, 296
- variational dropout, 479
- variational expectation maximization, 296, 370, 407
- variational inference, 369, 379
- variational InfoMax, 430
- variational information bottleneck problem, 431
- variational posterior, 369
  
- wall-clock time, 166
- Wasserstein distance, 443
- Weibull distribution, 394
  
- zero-shot learning, 522