

Index

adaptive boosting, 309–320	Gaussian process regression, 183–191	layer normalization, 534
Bayes method, 298, 315, 348, 352, 356, 361, 371	Hessian matrix, 36–45, 47, 48, 54, 55, 60, 61, 103, 159, 172, 173, 178, 193, 196, 202	natural language processing (NLP), 523
Bayes’ theorem, 114, 191, 297, 298	hierarchical classification/clustering, 387–400	positional encoding, 529, 534, 548
bias-variance tradeoff, 117–120	binary classification, 389–394	reinforcement learning with human feedback (RLHF), 535, 536
boosting, 123, 309, 310	binary clustering, 394–400	scaling laws, 537–538
clustering analysis, 363–381	bottom-up vs. top-down, 387–389	self-attention, 527, 528, 529, 530, 531, 538, 541, 543
Bernoulli mixture model, 381		token, 526–532, 534–536, 538, 541–544, 548
Gaussian mixture model, 370	ill-conditioned, 5, 122, 135, 136, 137, 138	transfer learning, 535, 536
hierarchical clustering, 394	independent component analysis (ICA), 209, 276–287	transformer architecture, 525
k-means clustering, 363	FastICA, 279	latent function, 191, 195, 196, 200, 201, 204
confidence, 148, 150, 153, 166, 186, 187, 196, 204, 294, 310, 549	negentropy, 280, 281	latent variable, 250, 252, 253, 255, 257–259, 371, 372, 374, 378, 382, 383
confusion matrix, 181, 205, 206, 303–305, 320, 369, 370, 385	non-Gaussianity, 276	least-squares method, 2, 148, 149, 163, 252, 277, 417, 434
cross validation, 121	whitening, 278	likelihood, 117, 149, 164, 165, 167, 170, 171, 176, 191–193, 196, 200, 201, 219, 254, 255, 257, 278, 284, 286, 297, 298, 372–374, 376, 383, 434
decision boundary, 86, 120, 166, 169, 171, 291, 293, 294, 298, 300, 331, 343, 344	independently and identically distributed (i.i.d.), 112, 183, 291, 372, 488, 492, 512	linear regression, 124–148, 159, 173, 321, 405, 416, 435, 504, 509
discriminant function, 174, 299, 301, 390	Jacobian matrix, 12, 17, 18, 25–30, 36, 40, 54, 102, 103, 158–160, 173, 178, 202, 281, 474	Bayesian regression, 115, 148–155, 186
distance and separability, 211	Karhunen-Loeve transformation (KLT), 220–223, 248, 306, 369, 388, 435	least squares estimate, 124
ensemble learning, 122	kernel method/mapping, 152, 243–245, 247, 328, 330, 348–350, 423	ridge regression, 122, 135–138, 163, 171, 428
entropy, 116, 269, 280, 297, 434	Kullback-Leibler (KL) divergence, 164, 270, 271, 284, 438, 551	local minimum, 43, 72–73, 415
cross-entropy, 116, 117, 434		logistic regression, 164–174
negentropy, 280, 281	large language models (LLMs), 523–550	log-likelihood, 116, 149, 170, 176, 254, 255, 257, 284, 373, 374, 383, 552
expectation maximization (EM), 250–255, 371	autoencoder, 551	
factor analysis, 250	cross attention, 526, 531, 532	Mahalanobis distance, 212, 213, 269, 295, 301
feature extraction, 209, 211–215, 389, 435	embedding, 526–532, 534, 538, 542	maximum a posteriori (MAP), 117, 148, 150, 164, 170, 171, 257, 298, 301
feature space, 154, 166, 169, 175, 192, 209, 211, 218, 243, 291–295, 312, 321, 344, 363, 451, 458	emergence, 539–541, 544, 545, 547	
Gaussian process classification, 191–199	latent space, 550–552	
binary, 191		
multiclass, 199		

maximum likelihood, 1, 115, 148, 149, 164, 167, 170, 219, 278, 286, 298, 373, 434	overfit, 6, 112–114, 118–122, 135–137, 149, 166, 169–171 184, 191, 192, 199, 206, 294, 331, 352, 387, 428, 435, 449, 532, 538	softmax function, 174–177, 179–181, 191, 199, 200, 204, 373, 378, 513, 528, 531
natural gradient descent, 164	overdetermination, 4, 125, 159	softmax regression, 174–180, 191, 199, 200, 204
neural networks	posterior probability, 114, 117, 149, 170, 297, 298, 373, 378, 383	solving equations, 3–24
autoencoder, 435–444	principal component analysis (PCA), 243–266	bisection search, 6–8
backpropagation, 425–435, 449, 511, 522, 524, 530–535	kernel PCA, 247–250	fixed-point method, 9–20, 21, 34, 474
competitive learning, 364, 451–457, 458, 459, 463	probabilistic PCA, 257–261	secant method, 8–9
deep learning, 431, 444–449, 521, 525, 532, 541, 550	prior probability, 114, 148, 149, 170, 171, 183, 187, 200, 211, 297, 371, 382	squared error, 1–3, 35, 115, 125, 137, 149, 157, 164, 260, 277, 427, 428, 436, 487, 503, 504, 511, 551
Hebbian learning, 409–411, 412	pseudo-inverse, 5, 26, 27, 40, 126, 140, 155, 159, 253, 259, 260	statistic classification
Hopfield network, 411–415	regularization, 5, 122, 137, 177, 428, 437, 438, 439	kernel Bayes classifier, 348–360
learning rate, 410, 417, 418, 428, 430, 452, 453, 459	reinforcement learning, 118, 471–518, 536, 537	k-nearest neighbors, 294–297
perceptron, 416–425, 444	agent, 469, 470, 472, 475, 476, 484, 512, 536	minimum distance classifier, 295, 299, 302
self-organizing map (SOM), 458–467	Bellman equation, 9, 474, 477, 479–481, 487, 490, 511	naïve Bayes classifier, 206, 297–309, 371, 372, 375, 432
Newton-Raphson method, 20–30	deep Q-learning, 509–512	support vector machine (SVM), 321–361
multivariate, 24–30	Markov chain, 471, 472	dual problem, 325, 327, 328, 333, 345, 347, 348
order of convergence, 32–34	Markov decision process (MDP), 469, 471, 475, 493	kernel SVM, 330
univariate, 20–24	Markov reward processing, 472	margin, 323, 324, 328, 341, 343–345
nonlinear least squares regression, 157–164	model-based planning, 478–483, 484, 485, 493	multiclass classification, 343–34
normal equation, 72, 126	model-free control, 483–502, 507, 509	primal problem, 325–328, 331, 333, 345
optimization (constrained), 75–101	policy, 469, 470, 472, 475–479, 481–492, 500, 506–509, 512–519, 537	sequential minimal optimization, 333–340
dual problem, 75, 84–89, 94, 110	TD (lambda), 497–502, 505, 506, 508, 509	soft margin, 331, 334
duality and KKT conditions, 84–87	temporal difference (TD), 490	support vectors, 321–325, 327, 329, 331, 333, 338, 340, 341, 344
equality constraints, 76–79	sigmoid function, 142, 144, 146, 147, 167–169, 171, 173, 175	t-distributed stochastic neighbor embedding (tSNE), 266–274
inequality constraints, 79–84	slack variable, 98, 331	uncertainty, 114, 116, 187, 190, 256, 293, 294, 297, 478
interior point method, 101–110		underdetermination, 4, 277
linear programming, 87–94		underfit, 6, 112–114, 118–122, 137, 169–171, 184, 190–192, 331
primal problem, 75, 84–86, 88, 89, 93		
quadratic programming, 99–101		
simplex algorithm, 94–99		
optimization (unconstrained), 35–72		
conjugate gradient method, 60–72		
gradient descent method, 43–46		
Newton’s method, 37–43		
quasi-Newton methods, 54–60		