

---

# Index

- $\ell_p$ -norm, 589
- $\nu$ -property, 80, 206, 238, 262
  - optimal  $\nu$ , 82
- $\sigma$ -algebra, 576
- $k$ -means, 520
- $p$ -convex hulls, 445
- $p$ -norm, 589
- adaptive loss, 80
- AdaTron, 316
- alignment, 57
- almost everywhere, 578
- annealed entropy, 140
- automatic relevance determination, 477
- Banach space, 586
- barrier method, 178
- basis, 581
  - canonical, 582
  - expansion, 584
  - Hilbert space, 587
  - orthonormal, 584
- Bayes
  - classifier, 5
  - point, 226
  - rule, 474
- Bernoulli trial, 129
- best element of a set, 179
- bias-variance dilemma, 126
- bilinear form, 583, 584
- bound
  - Chernoff, 129
  - concentration, 360
  - covering number, 396
  - entropy number, 397
  - generalization error, 10, 138, 394
- Hoeffding, 129
- leave-one-out, 198, 367, 369
- margin, 194
- McDiarmid, 360
- PAC-Bayesian, 381
- risk, 138
- span, 370
- bracket cover, 533
- cache, 286
- capacity, 9, 139, 433
- case, 1
- Cauchy sequence, 586
- Cauchy-Schwarz inequality, 584
- characteristic function, 626
- chunking, 300
- classification, 191
  - C-SV, 205
  - $\nu$ -SV, 206
  - binary, 2
  - Gaussian process, 486
  - LP, 214
  - multi-class, 211
- compact, 393
- complexity, 128
- compression, 518, 546
- concentration of measure, 360
- condition, 159
- conditional probability, 474
- conjugate gradient descent, 160, 162
- consistency, 131
- constraint, 13
- continuous, 275
  - Lipschitz, 275
  - uniformly, 275
- contrast function, 445
- convergence

- Cauchy, 586
- in probability, 130
- uniform, 133
- convex
  - combination, 581
  - set, 150
    - intersection of, 151
- convexity constraint, 444
- convolution, 46
- coordinate descent, 526
- covariance, 580
  - function, 30
  - matrix, 428, 580
    - centered, 442
- Cover's Theorem, 200
- covering number, 136, 393
- cross-validation, 217
- data
  - iid, 8, 252
  - snooping, 128
  - test, 8
  - training, 1
  - validation, 217
- dataset
  - Abalone, 293, 497, 572
  - Boston housing, 272
  - DNA, 416
  - MNIST, 569
  - oil flow, 537
  - Santa Fe, 273
  - separable, 191
  - small MNIST, 571
  - TIS recognition, 416
  - USPS, 244, 294, 440, 463, 556, 569
- decision function, 4, 15, 190
- decomposition algorithm, 22
- deflation method, 161
- denoising, 546
- density, 578
  - class-conditional, 465
  - estimation, 68, 549
- dimension, 581
- dimensionality reduction, 434
- Dirac  $\delta$ , 626
- direct sum, 411
- discriminant
  - Fisher, 457
  - kernel Fisher, 457
  - QP, 459
- distribution, 576
  - function, 579
- domain, 1, 29, 575
- dot product, 2, 584
  - canonical, 2
  - Frobenius, 57
  - space, 584
- dual
  - problem, 169
  - representation, 6
  - Wolfe, 171
- eigenvalue, 585
- eigenvector, 585
- empirical risk minimization, 127
- entropy number, 393
- equivalence relation, 57
- ERM, *see* empirical risk minimization
- error
  - false negative, 63, 565
  - false positive, 63
  - margin, 194, 206
  - misclassification, 62
  - punt, 211
  - reject, 211
  - test, 8
  - training, 8
- essential supremum, 590
- estimator, 72
  - almost unbiased, 199
  - M, 67
  - minimax, 148
  - quantile, 265
  - trimmed mean, 264
- event, 575
- evidence, 474
- example, 578
- expectation, 580

- extreme point, 153, 447
- fat shattering dimension, 392
- feasible point, 166
- feature, 428
  - extraction, 431, 443
  - map, 32, 93
    - continuity, 41
    - empirical kernel, 42
    - GNS, 425
    - kernel PCA, 43
    - Mercer, 36
    - normalized, 413
    - pairwise, 44
    - reproducing kernel, 32
  - monomial, 26
  - polynomial, 26
  - product, 26
  - space, 3, 39, 429
    - equivalence, 39
    - infinite-dimensional, 47
  - RKHS, 35
- Fisher
  - information, 72
  - matrix, 419
  - score, 419
- Fletcher-Reeves method, 163
- Gamma
  - distribution, 506
  - function, 107
- Gaussian
  - approximation, 478
  - kernel, 45
  - process, 481
- Generalized Portrait, 11
- generative model, 529
- generative topographic map, 530
- Gibbs classifier, 381
- global minimum, 152
- gradient descent, 157, 315
  - conjugate, 160
- Gram-Schmidt orthonormalization, 588
- graphical model, 418
- greedy selection, 291
- growth function, 9, 140, 393
- Heaviside function, 303
- hidden Markov model, 418
- Hilbert space, 35, 586
  - reproducing kernel, 36
  - separable, 586
- hit rate, 286
- Hough transform, 225
- Huber's loss, 76
- hyperparameter, 475
- hyperplane, 4
  - canonical, 12
  - optimal, 11
  - separating, 11, 168, 189
  - soft margin, 16
  - supporting, 237
- iid sample, 577
- implementation, 279
- induction principle, 127
- input, 1
- instance, 1
- integral operator, 30
- interior point, 295
  - method, 175
- interval cutting, 155
- invariance
  - translation, 46
  - unitary, 46
- Karush-Kuhn-Tucker conditions, 13, 166
- differentiable, 170
- necessary, 169
- kernel, 2, 30
  - B-spline, 46
  - R-convolution, 411
  - admissible, 30
  - alignment, 57
  - ANOVA, 272, 411
  - Bayes, 57
  - codon-improved, 417

- conditionally positive definite, 49, 53, 118
- conformal transformation, 408
- direct sum, 411
- DNA, 417
- examples, 45
- feature analysis, 443
- Fisher, 418
- for structured objects, 411
- Gaussian, 45, 46, 402, 411
- Hilbert space representation, 29
- infinitely divisible, 53
- inhomogeneous polynomial, 46
- jittered, 354
- Laplacian, 402
- locality-improved, 417
- locally linear embedding, 436
- map, *see* feature map
- Mercer, 30
- monomial, 27
- natural, 418, 419
- nonnegative definite, 30
- on a  $\sigma$ -algebra, 47
- on a group, 424
- optimal, 423
- PCA, 41, 92, 588
- pd, 31
- polynomial, 27, 45, 56
- positive definite, 30, 34
- properties, 45
- RBF, 46
- regularization interpretation, 94
- reproducing, 30, 33
- scaling, 216
- set, 47
- sigmoid, 45
- sparse vector, 412
- spline, 98
- strictly positive definite, 31
- symmetric, 2
- tanh, 45
- tensor product, 410
- trick, 15, 34, 195, 201, 429
- KKT, *see* Karush-Kuhn-Tucker conditions
- KKT gap, 170, 282
- Kronecker delta, 582
- Kuhn-Tucker conditions, *see* Karush-Kuhn-Tucker conditions
- label, 1
- Lagrange multiplier, 13
- Lagrangian, 13, 166
- SVM, 318
- pseudocode, 320
- Laplace approximation, 488
- Laplacian Prior, 501
- Laplacian process, 499
- learning
  - from examples, 1
  - machine, 8
  - online, 320
- leave-one-out, 250
  - machine, 370
  - mean field approximation, 377
- Lie group, 337
- likelihood, 69
- linear
  - combination, 581
  - independence, 581
  - map, 582
- Lipschitz continuous, 534
- LLE, *see* locally linear embedding
- locally linear embedding, 436
- log-likelihood, 419
- logistic regression, 63, 471
- loss function, 18, 19, 62, 394
  - $\varepsilon$ -insensitive, 18, 251
  - $\varepsilon$ -tube, 251
  - hinge, 324
  - zero-one, 8
- LP machine, 120, 214
- luckiness, 384
- MAP, *see* maximum a posteriori estimate
- map

- injective, 181
- surjective, 577
- margin, 142, 192
  - $\varepsilon$ -, 253
  - and flatness, 253
  - coding interpretation, 194
  - computational considerations, 204
  - of canonical hyperplane, 192
  - VC bound, 194
  - vs. training with noise, 192
- matrix, 582
  - adjoint, 46
  - condition of a, 68, 159
  - conditionally positive definite, 49
  - decoding, 213
  - diagonalization, 585
  - Gram, 30, 430
  - inverse, 583
  - inversion lemma, *see* Sherman-Woodbury-Morrison formula
  - kernel, 30
  - natural, 419
  - nonsingular, 583
  - orthogonal, 585
  - positive definite, 30, 585
  - product, 582
  - pseudo-inverse, 583
  - strictly positive definite, 31
  - symmetric, 585
  - tangent covariance, 346, 350
  - transposed, 583
  - unitary, 585
- Maurey's Theorem, 398
- maximum a posteriori estimate, 476
- maximum likelihood, 69
- measurable
  - function, 579
  - set, 577
- measure, 577
  - empirical, 579
- metric, 583
  - semi, 583
- Minimum Description Length, 194
- multidimensional scaling, 436
- network
  - neural, 202
  - RBF, 203
- Newton's method, 156
- noise
  - heteroscedastic, 269, 276
  - input, 192
  - parameter, 194
  - pattern, 192
- norm, 583
  - operator, 590
  - semi, 583
- notation table, 625
- objective function, 13
- observation, 1
- OCR, *see* optical character recognition
- offset, *see* threshold
- online learning, 320
- operator, 582
  - bounded, 590
  - compact, 393
  - norm, 590
- optical character recognition, 22, 211, 440
- optimization
  - constrained, 165
  - constraint qualification, 167
  - optimality conditions, 166
  - problem
    - dual, 199
    - infeasible, 173
    - sequential minimal, 234
- orthonormal set, 584
- outlier, 236
- output, 1
- overfitting, 127
- parameter optimization, 217
- Parzen windows, 6, 233
- pattern, 1

- pattern recognition, 2
- PCA, *see* principal component analysis
- Peano curve, 531
- perceptron, 193
- Polak-Ribiere method, 163
- pre-image
  - approximate, 546
  - exact, 544
- precompact, 393
- prediction, 288
- predictor corrector method, 163
- principal component analysis, 427, 442
  - kernel, 20, 431
  - linear, 428
  - nonlinear, 429, 434
  - oriented, 348
- principal curves, 435, 522
  - length constraint, 524
- principal manifold, 517
- prior, 472
  - data dependent, 500
  - improper, 477
- probability, 575
  - conditional, 464, 474
  - distribution, 576
  - measure, 577
  - posterior, 464
  - space, 575
- programming problem
  - dual, 14, 15, 19, 171
  - linear, 172
  - primal, 12, 16, 18
  - quadratic, 172
- projection pursuit, 445
  - kernel, 445
- proof, *see* pudding
- pudding, *see* proof
- quadratic form, 30
- Quantifier Reversal Lemma, 388
- quantile, 80, 450
  - multidimensional, 229
- quantization error, 518
  - empirical, 519
  - expected, 518
- random
  - evaluation, 181
  - quantity, 576
  - subset, 179
  - subset selection, 290
  - variable, 576
  - vector, 576
- rank-1 update, 290
- Rayleigh Coefficient, 457
- reduced KKT system, 176, 296
- reduced set, 258, 544
  - Burges method, 561
  - construction, 561
  - expansion, 553
  - method, 22
  - selection, 554
- regression, 19
  - C-SV, 253
  - $\nu$ -LP, 268
  - $\nu$ -SV, 260
- regularization, 87, 363, 433
  - operator, 93
  - Fisher, 420
  - for polynomial kernels, 110
  - for translation invariant kernels, 96
  - natural, 420
- regularized principal manifold, 517
  - algorithm, 526
- regularized quantization functional, 522
- Relevance Vector Machine, 258, 506
- replacing the metric, 160
- Representer Theorem, 89
- restart, 285
- risk, 127
  - actual, 8
  - Bayes, 9
  - bound, 138
  - empirical, 8, 127

- frequentist, 9
- functional, 89, 139
- regularized, 479
- robust estimator, 76
- RPM, *see* regularized principal manifolds
- RS, *see* reduced set
- RVM, *see* Relevance Vector Machine
- sample, 7, 577
  - complexity, 395
  - iid, 577
  - mean, 519
- score
  - function, 72
  - map, 419
- semi-norm, 583
- semiparametric models, 115
- Sequential Minimal Optimization,
  - 305
    - classification, 307
    - regression, 308
    - selection rule, 311
- SGMA, *see* sparse greedy matrix approximation
- shattering, 9, 136
  - coefficient, 136
- Sherman-Woodbury-Morrison, 299, 489
- significant figures, 178
- similarity measure, 2
- slack variable, 16, 205
- SMO, *see* Sequential Minimal Optimization
- snooping, 128
- soft margin loss, 63
- space
  - Banach, 586
  - dot product, 584
  - Hilbert, 586
  - linear, 581
  - normed, 584
  - probability, 575
  - vector, 581
- version, 225
- span, 581
  - of an SVM solution, 371
- sparse coding, 500
- sparse decomposition, 443
- sparse greedy
  - algorithm, 183
  - approximation, 461, 493
  - matrix approximation, 288
- sparsity, 120, 460
- SRM, *see* structural risk minimization
- stability, 361
- statistical manifold, 418
- stochastic process, 580
- stopping criterion, 282
- structural risk minimization, 138
- subset selection, 302
- Support Vector, 6, 14, 197, 202
  - bound, 210
  - essential, 210
  - expansion, 14, 255
  - in-bound, 210
  - mechanical interpretation, 14
  - novelty detection, 227
  - pattern recognition, 15
  - pattern recognition
    - primal reformulation, 559
  - quantile estimation, 227
  - regression, 17
  - regression using  $\nu$ , 19
  - set, 21
  - single-class-classification, 227
  - virtual, 22, 337
- SV, *see* Support Vector
- symbol list, 625
- symmetrization, 135
- target, 1
- Taylor series expansion, 155
- tensor product, 410
- test error, 66
- text categorization, 221
- threshold, 17, 203, 205, 209, 298, 310
- topological space, 41

training  
example, 8  
Gaussian Processes, 488  
kernel PCA, 429  
KFD, 460  
sparse KFA, 447  
SVM, 279  
transduction, 66

union bound, 135, 535  
unit ball, 590  
USPS, *see* dataset

variable  
dual, 13  
primal, 13  
variance, 580  
VC  
dimension, 9, 141, 391  
real-valued functions, 141  
entropy, 9, 139  
vector quantization, 243, 520  
version space, 225  
virtual example, 335  
virtual SV, 337

whitening, 347  
working set, 301