

Index

- 3-term DNF, 79
- F_1 -score, 207
- ℓ_1 norm, 149, 286, 315, 335
- accuracy, 18, 22
- activation function, 229
- AdaBoost, 101, **105**, 314
- all-pairs, 191, 353
- approximation error, 37, 40
- auto-encoders, 319
- backpropagation, 237
- backward elimination, 314
- bag-of-words, 173
- base hypothesis, **108**
- Bayes optimal, 24, 30, 221
- Bayes rule, 306
- Bayesian reasoning, 305
- Bennet’s inequality, 376
- Bernstein’s inequality, 376
- bias, 16, 37, 40
- bias-complexity tradeoff, 41
- Boolean conjunctions, 29, 54, 78
- boosting, 101
- boosting the confidence, 112
- boundedness, 133
- C4.5, 215
- CART, 216
- chaining, 338
- Chebyshev’s inequality, 373
- Chernoff bounds, 373
- class-sensitive feature mapping, 193
- classifier, 14
- clustering, 264
 - spectral, 271
- compressed sensing, 285
- compression bounds, 359
- compression scheme, 360
- computational complexity, 73
- confidence, 18, 22
- consistency, 66
- Consistent, 247
- contraction lemma, 331
- convex, 124
 - function, 125
 - set, 124
 - strongly convex, 140, 160
- convex-Lipschitz-bounded learning, 133
- convex-smooth-bounded learning, 133
- covering numbers, 337
- curse of dimensionality, 224
- decision stumps, 103, 104
- decision trees, 212
- dendrogram, 266, 267
- dictionary learning, 319
- differential set, 154
- dimensionality reduction, 278
- discretization trick, 34
- discriminative, 295
- distribution free, 295
- domain, 13
- domain of examples, 26
- doubly stochastic matrix, 205
- duality, 176
 - strong duality, 176
 - weak duality, 176
- Dudley classes, 56
- efficient computable, 73
- EM, 301
- Empirical Risk Minimization, *see* ERM
- empirical error, 15
- empirical risk, 15, **27**
- entropy, 298
 - relative entropy, 298
- epigraph, 125
- ERM, 15
- error decomposition, 40, 135

396 Index

- estimation error, 37, 40
- Expectation-Maximization, *see* EM
- face recognition, *see* Viola-Jones
- feasible, 73
- feature, 13
- feature learning, 319
- feature normalization, 316
- feature selection, 309, 310
- feature space, 179
- feature transformations, 318
- filters, 310
- forward greedy selection, 312
- frequentist, 305
- gain, 215
- GD, *see* gradient descent
- generalization error, 14
- generative models, 295
- Gini index, 215
- Glivenko-Cantelli, 35
- gradient, 126
- gradient descent, 151
- Gram matrix, 183
- growth function, 49
- halfspace, 90
 - homogenous, 90, 170
 - nonseparable, 90
 - separable, 90
- Halving, 247
- hidden layers, 230
- Hilbert space, 181
- Hoeffding's inequality, 33, 375
- holdout, 116
- hypothesis, 14
- hypothesis class, 16
- i.i.d., 18
- ID3, 214
- improper, *see* representation independent
- inductive bias, *see* bias
- information bottleneck, 273
- information gain, 215
- instance, 13
 - instance space, 13
- integral image, 113
- Johnson-Lindenstrauss lemma, 284
- k-means, 268, 270
 - soft k-means, 304
- k-median, 269
- k-medoids, 269
- Kendall tau, 201
- kernel PCA, 281
- kernels, 179
 - Gaussian kernel, 184
 - kernel trick, 181
 - polynomial kernel, 183
 - RBF kernel, 184
- label, 13
- Lasso, 316, 335
 - generalization bounds, 335
- latent variables, 301
- LDA, 300
- Ldim, 248, 249
- learning curves, 122
- least squares, 95
- likelihood ratio, 301
- linear discriminant analysis, *see* LDA
- linear predictor, 89
 - homogenous, 90
- linear programming, 91
- linear regression, 94
- linkage, 266
- Lipschitzness, 128, 142, 157
 - subgradient, 155
- Littlestone dimension, *see* Ldim
- local minimum, 126
- logistic regression, 97
- loss, 15
- loss function, **26**
 - 0-1 loss, 27, 134
 - absolute value loss, 95, 99, 133
 - convex loss, 131
 - generalized hinge loss, 195
 - hinge loss, 134
 - Lipschitz loss, 133
 - log-loss, 298
 - logistic loss, 98
 - ramp loss, 174
 - smooth loss, 133
 - square loss, 27
 - surrogate loss, 134, 259
- margin, 168
- Markov's inequality, 372
- Massart lemma, 330
- max linkage, 267
- maximum *a posteriori*, 307
- maximum likelihood, 295
- McDiarmid's inequality, 328
- MDL, 63, 65, 213
- measure concentration, 32, 372
- Minimum Description Length, *see* MDL
- mistake bound, 246
- mixture of Gaussians, 301
- model selection, 114, 117
- multiclass, 25, 190, 351
 - cost-sensitive, 194
 - linear predictors, 193, 354
 - multivector, 193, 355
 - Perceptron, 211
 - reductions, 190, 354
 - SGD, 198
 - SVM, 197
- multivariate performance measures, 206
- Naive Bayes, 299
- Natarajan dimension, 351
- NDCG, 202

- Nearest Neighbor, 219
 - k-NN, 220
- neural networks, 228
 - feedforward networks, 229
 - layered networks, 229
 - SGD, 236
- No-Free-Lunch, 37
- nonuniform learning, 59
- Normalized Discounted Cumulative Gain, *see* NDCG
- Occam’s razor, 65
- OMP, 312
- one-versus-all, 191, 353
- one-versus-rest, *see* one-versus-all
- online convex optimization, 257
- online gradient descent, 257
- online learning, 245
- optimization error, 135
- oracle inequality, 145
- orthogonal matching pursuit, *see* OMP
- overfitting, 15, 41, 121
- PAC, 22
 - agnostic PAC, 23, 25
 - agnostic PAC for general loss, **27**
- PAC-Bayes, 364
- parametric density estimation, 295
- PCA, 279
- Pearson’s correlation coefficient, 311
- Perceptron, 92
 - kernelized Perceptron, 188
 - multiclass, 211
 - online, 258
- permutation matrix, 205
- polynomial regression, 96
- precision, 206
- predictor, 14
- prefix free language, 64
- Principal Component Analysis, *see* PCA
- prior knowledge, 39
- Probably Approximately Correct, *see* PAC
- projection, 159
 - projection lemma, 159
- proper, 28
- pruning, 216
- Rademacher complexity, 325
- random forests, 217
- random projections, 283
- ranking, 201
 - bipartite, 206
- realizability, 17
- recall, 206
- regression, 26, 94, 138
- regularization, 137
 - Tikhonov, 138, 140
- regularized loss minimization, *see* RLM
- representation independent, 28, 80
- representative sample, 31, 325
- representer theorem, 182
- ridge regression, 138
 - kernel ridge regression, 188
- RIP, 286
- risk, 14, 24, **26**
- RLM, 137, 164
- sample complexity, 22
- Sauer’s lemma, 49
- self-boundedness, 130
- sensitivity, 206
- SGD, 156
- shattering, 45, 352
- single linkage, 267
- Singular Value Decomposition, *see* SVD
- Slud’s inequality, 378
- smoothness, 129, 143, 163
- SOA, 250
- sparsity-inducing norms, 315
- specificity, 206
- spectral clustering, 271
- SRM, 60, 115
- stability, 139
- Stochastic Gradient Descent, *see* SGD
- strong learning, 102
- Structural Risk Minimization, *see* SRM
- structured output prediction, 198
- subgradient, 154
- Support Vector Machines, *see* SVM
- SVD, 381
- SVM, 167, 333
 - duality, 175
 - generalization bounds, 172, 333
 - hard-SVM, 168, 169
 - homogenous, 170
 - kernel trick, 181
 - soft-SVM, 171
 - support vectors, 175
- target set, 26
- term frequency, 194
- TF-IDF, 194
- training error, 15
- training set, 13
- true error, 14, 24
- underfitting, 41, 121
- uniform convergence, 31, **32**
- union bound, 19
- unsupervised learning, 265
- validation, 114, 116
 - cross validation, 119
 - train-validation-test split, 120
- Vapnik-Chervonenkis dimension, *see* VC dimension
- VC dimension, 43, 46
- version space, 247
- Viola-Jones, 110
- weak learning, 101, **102**
- Weighted-Majority, 252