

## Index

### Symbols

0-1 loss, 167, 343

$F$

- $5 \times 2$  cross-validation paired test, 213
- distribution, 213

$k$ -means, 280, 281

$p$ -value, 21, 40, 57, 141, 145, 153, 233, 386, 396, 403, 419, 422

$\chi^2$ , 42, 46, 59

- distribution, 141, 146, 195, 212, 213, 224, 225, 233, 234
- test, 41

### A

acceptance region, 19, 44

AdaBoost, 185, 333–338, 345, 365, 366, 371, 376

- AdaBoost.M1, 357, 371, 376
- AdaBoost.M2, 185, 358, 371
- pseudo-loss, 357

Anderson–Darling test, 50

arbiter tree, 126

artificial contrasts with ensembles (ACE), 412

asymptotic, 13

asymptotic mean integrated squared error (AMISE), 294

asymptotic significance level, 20

### B

backward elimination by change in margin (BECM), 407

bagging, 126, 127, 171, 361–363

- decision trees, 361, 364, 365, 382
- in bag, 361
- out of bag, 361, 411, 413

balanced box decomposition tree, 300

Bartlett test, 151, 225

batch learning, 124, 125, 258, 301

Bayes error, 167, 169, 338

Bayes neural network, 260

Bayesian statistics, 5, 10, 16

$BC_a$  confidence interval, 76, 77

bias, 7, 13

binned data, 39

binomial

- confidence interval (Clopper–Pearson), 175, 206
- distribution, 9, 21, 97, 131, 141, 174, 179, 181, 205, 208, 212, 215, 316, 318
- test, 212, 409, 413

binomial loss, 167

Bonferroni correction, 216

boosting, 126, 210, 366

- confidence-rated, 358
- corrective, 347
- decision trees, 197, 334, 338, 341, 344, 345, 349, 354, 360, 365, 382, 400
- learning rate, 340
- shrinkage, 340
- totally corrective, 347

bootstrap, 65, 100, 179, 180, 207, 361

- .632+ estimator, 180
- bias estimation, 67
- confidence intervals, 68–70
- leave-one-out, 179
- parametric, 70
- replica, 65
- smoothed, 70

Brownian motion, 352

bump hunting, 121, 417–423

### C

canonical correlation analysis (CCA), 238

categorical variable, 129–132, 324

- dummy variables, 327
- nominal, 129, 320
- ordinal, 129, 320

- Cauchy distribution, 10, 66
    - median, 66
  - centering, 145, 146, 228, 229, 237, 238, 269, 270
  - Chernoff–Lehmann theorem, 44
  - class label
    - known, 165, 252
    - predicted, 165
    - true, 165
  - class posterior odds, 132, 222, 223, 231, 233, 234
  - class posterior probability, 140, 166–168, 183, 191, 202, 222, 224, 233, 241, 247, 260, 265, 288, 294, 309, 310, 312, 324, 339, 340, 343, 357, 361, 362, 364, 376, 377, 387, 395, 396
  - class prior probability, 166, 182, 183, 184, 188, 191, 199, 222, 223, 286, 294, 312, 362, 400
  - classification, 165
    - bias, 170, 361
    - binary, 165, 269
    - coefficient of pairwise correlation, 359
    - confidence, 360
    - hard label, 166, 167, 376, 382
    - irreducible noise, 170, 354
    - multiclass, 126, 131, 165, 166, 185, 223, 225, 230, 234, 235, 236, 293, 296, 357, 358, 371–378, 413
    - ordinal, 131
    - prediction strength, 359
    - soft score, 166, 196, 200–202, 205, 224, 265, 283, 284, 334, 336, 343, 357, 358, 376, 382, 396, 422
    - statistical, 340, 343
    - variance, 170, 361
  - classification cost, 190
  - classification edge, 343, 345, 349, 358
  - classification error, 167, 171, 173–177, 178, 181, 182, 195, 196, 198, 309, 313, 316, 341, 343, 349, 354–356
  - classification margin, 283, 343, 349, 351, 352, 358, 359, 366, 396, 407, 433
  - classifier diversity, 358–365, 373
  - clustering, 121, 184, 280
    - *k*-means, 280
  - condition number, 150, 151, 282
  - conditional likelihood, 21
  - conditional Monte Carlo, 64
  - confidence interval, 10, 14, 17
    - confidence set, 10
    - hypothesis test, 45
    - pivotal quantity, 45
  - confidence level, 11
  - confusion matrix, 106
  - consistent estimator, 42, 97, 99
  - correlation, 391
  - covariance matrix, 66, 147–158, 160, 221, 222, 223–231, 233, 238, 269, 274
  - Cramér–von Mises test, 50
  - critical region, 11, 28
  - cross-entropy, 257, 261, 310, 338
  - cross-validation, 58, 78–82, 101, 126, 177–179, 315
    - *K*-fold, 81
    - folds, 177
    - leave-one-out, 79, 140, 178
    - repeated, 178, 214
  - curse of dimensionality, 102, 302, 303, 327
- D**
- datasets
    - BaBar PID data, 325, 364, 365, 378, 394
    - ionosphere data, 152, 154–158, 197, 198
    - MAGIC telescope data, 239–246, 275–278, 280–282, 295, 297, 299, 359, 396–399, 409, 413
    - two-norm data, 331, 341
  - decision tree, 127, 130, 138, 139, 171, 173, 183, 365, 371, 374, 381
    - branch node, 309
    - C4.5, 138, 139, 321
    - CART, 307, 321
    - CHAID, 307
    - impurity gain, 138, 309, 312, 319, 321, 323, 325
    - leaf node, 166, 171, 173, 176, 187, 307, 309, 312, 313, 316, 334, 360, 361
    - node impurity, 307, 309, 310, 314, 319
    - optimal pruning level, 315
    - optimal pruning sequence, 314
    - predictive association, 139, 322, 325, 326, 411, 412
    - probabilistic splits, 139
    - pruning, 313–318
    - risk, 313
    - surrogate splits, 139, 321–323, 324, 390, 411, 412
    - terminal node, 309
  - deconvolution, 112
  - deflation, 161, 237
  - degrees of freedom (DOF), 59, 214, 224, 233
  - density estimation, 89, 120, 121, 294, 295, 296, 299, 301
    - empirical (epdf), 90
    - error, 95
    - histogram, 90
    - kernel, 56

- Monte Carlo, 111
- nonparametric, 93
- optimal, 94
- orthogonal series, 108
- parametric, 89, 93
- deviance, 310
- dimensionality reduction, 146, 147, 230, 236, 279
- discriminant analysis, 183, 221–231, 374
  - linear, 166, 206, 210, 222, 232, 236, 364, 400
  - pseudolinear, 197, 202
  - quadratic, 222
- E**
- efficient estimator, 7, 8, 13
- eigenvalue decomposition (EVD), 148–150, 230, 266
  - generalized, 230
- elbow plot, 152
- entropy, 118, 160
  - cross-entropy, 257
  - differential, 159, 392
  - Shannon, 159, 392
- error correcting output code (ECOC), 126
  - complete design, 372, 374
  - exhaustive design, 372
- error function, 252, 257
- expectation-maximization algorithm, 135
- expected prediction error (EPE), 78
- exponential family, 9
- exponential loss, 167
- F**
- false positive rate (FPR), 196, 227
- feature irrelevance, 387, 412
- feature ranking, 389–401
  - embedded algorithm, 389
  - filter, 389
  - wrapper, 389
- feature redundancy, 388, 412
- feature selection, 386
  - embedded algorithm, 389
  - filter, 389
  - wrapper, 389
- feature strong relevance, 387, 411
- feature weak relevance, 387
- feature-based sensitivity of posterior probabilities (FSPP), 395
- Feldman–Cousins (FC), 36
- Fisher discriminant, 221, 229
- Fisher information, 6, 47
  - matrix, 6
- Fokker–Planck equation, 353
- frequentist statistics, 5, 10, 14
  - coverage, 22–25
- fuzzy rules, 301
  - membership function, 301
- G**
- Gauss–Markov theorem, 37
- Gauss–Seidel method, 291
- generalization error, 169, 178, 180, 312, 338, 340, 344, 351, 361
- genetic algorithms, 262
- genetic crossover, 262
- GentleBoost, 187, 338, 339, 341, 349, 356, 357, 376
- Gini diversity index, 127, 310, 334, 338, 389, 411
- goodness of fit (GOF), 39
- Gram matrix, 266, 268, 270, 272, 273, 279, 288, 296, 299, 375
- Gram–Schmidt orthogonalization, 161, 237
- Green’s function, 272
- H**
- Hamming loss, 372
- Hessian, 6, 232, 233, 261, 377
- heterogeneous value difference metric (HVDM), 302
- heteroscedastic, 75, 85
- hinge loss, 167
- histogram, 22, 39, 90, 91
  - binning, 97
- Hosmer–Lemeshow test, 234, 244
- Hotelling transform, 147
- Householder reflections, 269
- hypothesis, 11
  - composite, 11, 19, 47
  - simple, 59
- hypothesis test, 11
  - $p$ -value, 211
  - alternative hypothesis, 11, 211, 403, 407
  - confidence interval, 45
  - critical region, 11
  - decision rule, 44
  - level, 403
  - likelihood ratio, 46
  - multiple, 216
  - null hypothesis, 11, 211, 212, 217, 233, 386, 396, 403, 405–407, 410, 412
  - one-sample, 39
  - power, 12, 28, 40, 211, 403
  - replicability, 214
  - $\alpha$ -level test, 211
  - $\alpha$ -size test, 211
  - score, 46

- simple, 28
  - two-sample, 39
  - Type I error, 11, 211, 216, 403
  - Type II error, 12, 211, 403
  - uniformly most powerful (UMP), 12, 21, 29, 59, 215
  - Wald, 46
- I**
- IB3 algorithm, 301
  - ideogram, 93
    - Gaussian, 93
  - independent component analysis (ICA), 146, 149, 158–162, 236
  - independent identically distributed (i.i.d.), XV
  - indicator function, 55, 91
  - influence function, 83
  - integrated squared error (ISE), 89, 96
  - interrater agreement, 360
  - irrelevant feature, 387
  - iterative single data algorithm (ISDA), 291, 292, 302, 384
- J**
- jackknife, 70–77, 101
    - delete- $d$ , 74
    - generalized, 74
- K**
- kappa statistic, 360
  - Karhunen–Loeve transform, 147
  - kd-tree, 300
  - kernel density estimation, 56, 92, 201, 202, 205, 207
    - adaptive, 103
    - fuzzy rules, 301
    - multivariate, 92, 106
    - standard deconvolution, 116
  - kernel dot product, 267, 268, 286
  - kernel function, 92, 266
  - kernel regression, 269, 270
  - kernel ridge regression, 274–278, 279, 295, 296, 302, 383
  - kernel trick, 268, 274, 286, 288
  - Kolmogorov–Smirnov (KS) test, 49
  - Kullback–Leibler divergence, 159
  - kurtosis, 162
    - excess, 162
    - sample, 225
- L**
- label noise, 340, 354
  - labeled data, 165
  - Laplace approximation, 261
  - learning curve, 81
  - learning rate, 187, 258, 270, 340, 341
  - least squares estimation, 13, 22, 60, 201, 232, 236, 269, 377
  - likelihood
    - equation, 12
    - equation roots, 12
    - extended, 22
    - function, 5
    - maximum likelihood estimator (MLE), 12
    - ratio, 14, 46
    - with missing data, 134, 135
  - linear correlation, 7
    - Kendall tau, 391
    - Pearson, 389, 391
    - Spearman rank, 391
  - linear discriminant analysis (LDA), 304, 381
  - linear regression, 235, 236, 274
    - intercept, 232
    - multiple, 237
    - multivariate, 236
    - regression of an indicator matrix, 236
    - ridge, 274
  - linear statistic, 72, 75
  - linearly separable, 252, 283
  - local density tests, 55
  - location parameter, 17
  - log odds, 254
  - logistic regression, 231–235, 338, 381
  - logit function, 231, 254
  - LogitBoost, 339, 340, 357
  - loss
    - binomial, 167, 339
    - exponential, 167, 341, 376
    - Hamming, 130, 374–376
    - hinge, 167, 283
    - quadratic, 167, 372, 376
  - loss function, 165, 257, 260
  - LPBoost, 346, 348, 349, 358
- M**
- machine learning, 121
  - Mahalanobis distance, 41, 140, 225, 243
  - maintained hypothesis, 43, 47
  - Mann–Whitney  $U$  test, 200
  - marginal likelihood, 16
  - Markov blanket, 387
  - Markov boundary, 387–389, 401
  - maximum likelihood, 12
  - McNemar’s test, 211, 215
  - mean integrated squared error (MISE), 97, 202
  - mean squared error (MSE), 6, 95, 127, 168, 236
  - median, 73

- Mercer's Theorem, 286  
 meta learning, 126  
 minimal covariance determinant, 140  
 minimal volume ellipsoid, 140  
 missing data, 132–139, 323, 324
  - augmentation, 137
  - casewise deletion, 135
  - hot deck imputation, 138
  - ignorable missingness, 133
  - imputation, 137
  - imputation by regression, 138
  - lazy evaluation, 136
  - missing at random (MAR), 132
  - missing completely at random (MCAR), 132, 136, 138
  - missing not at random (MNAR), 132
  - reduced model, 136
  - testing MCAR, 133
- Monte Carlo
  - bootstrap, 66
  - density estimation, 111
  - permutation test, 64
- multiclass learning
  - error correcting output code (ECOC), 372
  - one versus all (OVA), 372
  - one versus one (OVO), 372
- multinomial distribution, 22, 48, 141, 185, 234, 363
- multiple hypothesis test, 216, 404
  - Bonferroni correction, 216, 405, 412
  - complete null, 405
  - Hochberg procedure, 406, 407, 409, 412
  - Holm procedure, 405, 407
  - Sidak correction, 216, 405
  - strong control, 405
  - weak control, 386, 405
- multivariate regression, 236
- mutual information, 159, 347, 377, 389, 391
  - symmetric uncertainty, 392
- N**
- Nadaraya–Watson estimator, 294, 295, 299
- naive Bayes, 105, 210, 364, 381
- nearest neighbor rules, 184, 186, 268, 364, 365, 383
  - approximate neighbor, 300
  - IB3 algorithm, 301
- neural network, 280, 374
  - feed-forward, 254–260
  - prior distribution, 260
- Neyman modified chi-square, 42
- Neyman smooth test, 51
- nominal variable, 129, 165, 302, 320
- nonsmooth statistics, 73
- normal distribution, 14
- nuisance parameter, 19, 43
- O**
- one versus all (OVA), 236, 357, 372, 375–377
- one versus one (OVO), 372, 374–377
- one-sided sampling, 184
- online learning, 124, 125, 258, 301
- ordinal variable, 129, 320
- outliers, 139–141
  - masking, 140
- overfitting, 366
- overtraining, 173, 366
- P**
- pairwise similarity, 265
- parallel learning, 125–127, 362, 399
- partial least squares, 232, 236–239
  - loadings, 237
  - scores, 237
  - weights, 237
- patient rule induction method (PRIM), 421
- Pearson chi-square, 41
- perceptron,
  - multi-layer, 254
  - perceptron criterion, 252
- permutation
  - sampling, 63–65
  - test, 64
- permutation sampling, 152, 180, 395, 410–413
- pivotal quantity, 17, 45
- point spread function, 112
- Poisson distribution, 21, 22
- pooled data, 57
- posterior distribution, 260
- power, 12
- power divergence family, 48
- principal component analysis (PCA), 146, 147–158, 230, 237
  - correlation PCA, 149
  - covariance PCA, 149
  - loadings, 149
  - nontrivial component, 152
  - principal components, 149
  - scores, 149
- prior distribution, 260
- profile likelihood, 20
- proportional odds model, 132, 233
- pseudo-inverse, 228, 238, 268
- Q**
- QR decomposition, 228, 269, 377
- quadratic loss, 167

- quadratic programming (QP), 289, 347, 377
- quantile
  - confidence bounds, 207
  - QQ plot, 225
- queue, 292
- R**
- radial basis functions (RBF), 273, 302
  - RBF network, 267, 280
- random forest, 361–363, 362, 364, 366, 400, 411
  - balanced, 185
  - weighted, 362
- random subspace, 126, 127, 136, 363, 364, 366, 400
- random variable elimination (RVE), 402
- Rao–Cramér–Frechet (RCF) bound, 7–10, 35
- receiver operating characteristic (ROC), 179, 196–210, 227, 382
  - threshold averaging, 205
  - vertical averaging, 205
- reduced model, 363
- redundant feature, 388
- reflection, 145
- regression, 127, 128, 138, 251, 294, 413
  - bias, 169
  - irreducible noise, 169
  - linear, 235
  - locally weighted, 295–297, 383
  - multiple, 127, 131, 168, 265, 269, 296, 339
  - multivariate, 127, 296
  - stepwise, 338
  - variance, 169
- regularization, 114, 260, 270–278
  - Tikhonov, 117
- ReliefF, 390, 400
- Representer Theorem, 272
- resubstitution error, 173, 313, 344
- ridge regression, 274
- RobustBoost, 354, 355, 358
- Rosenblatt’s theorem, 96
- runs, 51
- RUSBoost, 185, 187
- S**
- sample size, 5
- scale
  - interval, 129, 390
  - nominal, 129
  - ordinal, 129, 390
- scaling, 145, 146, 160, 237, 238, 269, 270, 298
- score, 6, 48
  - score statistic, 48
  - test, 46
- scree plot, 152
- semi-supervised learning, 122
- sequential backward elimination (SBE), 388, 389, 395, 413
  - backward elimination by change in margin (BECM), 407
  - remove  $r$  add  $n$ , 402
- sequential forward selection (SFS), 395
  - add  $n$  remove  $r$ , 402
- sequential minimal optimization (SMO), 290, 291, 302, 375, 384
- Sidak correction, 216
- sigmoid function, 253, 278, 352, 395
- signed-rank test, 396, 407
- significance level, 11, 12, 40, 44
  - asymptotic, 20
- signum, 252
- simple test, 47
- singular value decomposition (SVD), 150, 228, 238
  - thin, 150
- smoothed bootstrap, 70
- stacked generalization, 126
- stagewise modeling, 338
- standardization, 145
- stationary sequence, 75
- stepwise modeling, 338
- stratified sampling, 181, 182
- strongly relevant feature, 387
- Student  $t$ , 19
  - 10-fold cross-validation test, 214
  - $5 \times 2$  cross-validation paired test, 213
  - $10 \times 10$  cross-validation test with calibrated degrees of freedom, 214
  - distribution, 64, 212, 213
  - test, 412
- subsampling, 75
- substitution method, 8
- sufficient statistic, 13, 18, 21
- supervised learning, 122
- support vector machines (SVM), 166, 232, 265, 295, 302, 371, 375, 400
  - bias term, 284
  - box constraint, 267, 286
  - dual problem, 284, 287
  - linear, 283–285, 381
  - nonlinear, 285, 286, 383
  - primal problem, 284, 287
  - support vectors, 285
  - working set algorithm, 289
- synthetic minority oversampling technique (SMOTE), 165, 186
- systematic error, 112, 210, 383, 385

**T**

tables, 39  
tall data, 123, 247, 382  
target coding, 252  
test power, 211  
test replicability, 214  
Tikhonov regularization, 117, 260  
Tomek links, 184, 186  
TotalBoost, 347–349, 358  
training error, 173  
transductive learning, 122  
true label, 252  
true positive rate (TPR), 196, 227  
twoing criterion, 319, 323  
Type I error, 211  
Type II error, 211

**U**

unbiased estimator, 7  
unbinned data, 39  
unfolding, 112–120  
uniformly most powerful (UMP) test, 12  
unlabeled data, 165  
unsupervised learning, 121

**V**

value difference metric (VDM), 302  
Vapnik–Chervonenkis dimension, 344  
variable ranking, 386, 389–401  
– embedded algorithm, 389  
– filter, 389  
– wrapper, 389  
variable selection, 386  
– embedded algorithm, 389  
– filter, 389  
– wrapper, 389

**W**

Wald statistic, 47  
Wald test, 46  
Watson test, 51  
weakly relevant feature, 387  
weight decay, 260  
whitening, 161  
wide data, 123, 247, 382  
Wilcoxon rank sum test, 200, 412