Index

SYMBOLS

0/1-loss, 297
F-distribution, 50
F-test, 105
G-test, 103
L_1-distance, 474
L_2-distance, 471
ℓ_1-loss minimization, 280
ℓ_1-regularization learning, 268
$\ell_1 + \ell_2$-constrained LS, 275
ℓ_2-constrained least squares, 259
ℓ_2-loss minimization, 245
ℓ_2-regularization learning, 261
ℓ_p-norm, 273
$\ell_{1,2}$-constrained LS, 277
f-divergences, 470
k-means clustering, 447
k-nearest neighbor classifier, 182
p-value, 99
t-distribution, 49
t-test, 101
z-test, 101

A

absolute loss, 280
acceptance, 99, 213
activation function, 243
adaboost, 348
additive law, 13
additive model, 239
Akaike information criterion, 150
Ali-Silvey-Csiszár divergences, 470
almost sure convergence, 77
alternating direction method of multipliers, 268
alternative hypothesis, 99
anomaly detection, 6
arithmetic mean, 76
asymptotic efficiency, 142
asymptotic normality, 79, 143
asymptotic theory, 151
asymptotic unbiasedness, 141
augmented Lagrange function, 270
autoencoder, 436
autoregressive model, 368

B

backward elimination, 272
bagging, 344
bandwidth, 93, 174
base distribution, 228
basis function, 237
batch learning, 355
Bayes decision rule, 119
Bayes risk, 120
Bayes' theorem, 53, 93, 186
Bayesian credible interval, 97
Bayesian inference, 92, 185, 244
Bayesian information criterion, 202
Bayesian optimization, 373
Bayesian predictive distribution, 186
Bennett's inequality, 90
Bernoulli distribution, 27
Bernoulli trial, 26
Bernstein's inequality, 89
beta distribution, 44
beta function, 44, 66
between-class scatter matrix, 413
big data, 3
bin, 169
binomial coefficient, 26
binomial distribution, 26
binomial theorem, 27, 35
blocked Gibbs sampling, 220
boosting, 346
bootstrap, 97, 344
burn-in, 220

C

Cantelli's inequality, 83
category, 94, 113
Cauchy distribution, 47
centering matrix, 432
central limit theorem, 78, 143
centralization, 406
chain rule, 438
change detection, 6, 469
characteristic function, 22
characteristic kernel, 476
Chebyshev's inequality, 83
Chernoff's inequality, 83
chi-square test, 103
chi-squared distribution, 44
class imbalance, 302
class-balance change, 385
class-balance weighted LS, 386
class-conditional probability density, 115

class-posterior probability, 117
class-prior probability, 115
classification, 5, 94, 236
clustering, 5, 167, 447
collapsed Gibbs sampling, 220
column marginal total, 53
complementary event, 13
complementary slackness, 310
completing the square, 38
composite event, 11
concavity, 164
concentration parameter, 70, 228
conditional expectation, 52
conditional probability density function, 53
conditional probability distribution, 52
conditional probability mass function, 52
conditional random field, 332
conditional risk, 120
conditional variance, 52
confidence interval, 95
confidence level, 95
confusion matrix, 136
conjugate prior, 188
consistency, 139
constrained least squares, 257
constraint, 159
contingency table, 53
continuous random variable, 14
continuous Sylvester equation, 398
continuous uniform distribution, 37
contrastive divergence algorithm, 443
convergence in distribution, 79
convergence in law, 79
convergence in probability, 77, 139
convex function, 84
convolution, 73
correlation coefficient, 55
countable set, 14
covariance, 55
covariate, 113
covariate shift, 378
Cramér-Rao inequality, 141
critical region, 100
critical value, 100
cross entropy loss, 437
cross validation, 94, 154
cumulative distribution function, 15, 208
curse of dimensionality, 7, 239, 406

D

data mining, 3
data visualization, 116
De Morgan's law, 13
decision boundary, 114
decision region, 114
decision stump, 343
density ratio, 465
dependence, 56
dependent variable, 113
design matrix, 246
determinant, 40
digamma function, 225
dimensionality reduction, 7, 405, 429
Dirac's delta function, 204
Dirichlet distribution, 63
Dirichlet process, 228
Dirichlet process mixture, 228
discrete random variable, 14
discrete uniform distribution, 25
discrimination function, 114
discriminative approach, 121, 236
disjoint events, 13
distance function, 148
distributional change detection, 469
distributive law, 13
domain adaptation, 7
double exponential distribution, 48
dynamic programming, 334

E

effective number of parameters, 366
efficiency, 141
eigenvalue, 65
eigenvalue decomposition, 65
eigenvector, 65
elastic-net regression, 276
elementary event, 11
ellipse, 63
embedding matrix, 406
empirical Bayes, 95, 193
empty event, 11
energy distance, 388
energy function, 441
ensemble learning, 7, 343
entropy, 149
error back-propagation, 254
estimate, 91
estimator, 91
Euclidean norm, 93
Euler number, 32
event, 11
evidence, 193
evidence maximization, 193
exogenous variable, 113

expectation, 16
expectation-maximization algorithm, 162
expected absolute error, 18
expected squared error, 18, 140
explanatory variable, 113
exponential distribution, 44
exponential loss, 349
extrapolation, 379

F

factorial, 26
false negative, 101
false positive, 101
feature selection, 272
feature vector, 113
Fisher discriminant analysis, 413
Fisher information, 141
Fisher information matrix, 141
Fisher score, 141
Fisher's exact test, 109
Fisher's linear discriminant analysis, 131, 297
forward selection, 272
Fourier transform, 22
free energy, 193
frequentist inference, 93, 187
Frobenius norm, 398
function approximation, 236
fused lasso, 273

G

gamma distribution, 41
gamma function, 41, 70
Gaussian approximation, 198
Gaussian bandwidth, 240
Gaussian center, 240
Gaussian distribution, 37
Gaussian function, 243
Gaussian kernel, 175, 240, 312
Gaussian Markov network, 478
Gaussian mixture model, 157
Gaussian model, 125
generalization ability, 3, 115
generalized eigenvalue, 65
generalized eigenvector, 65
generalized inverse, 247
generalized KL divergence, 466
generalized mean, 76
generative approach, 121
geometric distribution, 35
geometric mean, 76
Gibbs sampling, 218
global optimal solution, 162
goodness-of-fit test, 102
gradient method, 161
grand total, 53
graphical lasso, 479
group-lasso regression, 277

H

Hölder's inequality, 85
hard margin support vector machine, 305
harmonic mean, 76
Hessian matrix, 143, 198
hierarchical model, 242
hinge loss, 315
histogram method, 169
Hoeffding's inequality, 89
Huber loss, 282
hypercube, 174
hyperellipsoid, 129
hypergeometric distribution, 28
hypergeometric series, 31
hyperparameter, 193
hyperplane, 131
hypersphere, 178
hypothesis testing, 99

I

imaginary unit, 22
importance, 207
importance sampling, 207, 381
importance weighted cross validation, 382
importance weighted LS, 379
importance weighting, 379
independence, 56
independence test, 102
independent and identically distributed, 75, 115
independent variable, 113
inlier-based outlier detection, 464
input variable, 113
inscribed hypersphere, 405
integration by parts, 42
intersection of events, 12
inverse transform sampling, 208
invertible matrix, 63
irrational number, 39
iterative algorithm, 160
iterative retargeted LS, 317
iteratively reweighted LS, 285

J

Jacobian, 22, 40
Jacobian matrix, 40
Jensen's inequality, 85, 164
joint probability, 186
joint probability density function, 52
joint probability distribution, 51
joint probability mass function, 51

K

Kantorovich's inequality, 87
Karush-Kuhn-Tucker conditions, 310
kernel k-means clustering, 448
kernel density estimation, 93, 175
kernel function, 93, 175
kernel matrix, 249
kernel method, 242
kernel model, 240
kernel trick, 311
Kolmogorov, 13
Kronecker product, 71
Kullback-Leibler divergence, 149
kurtosis, 19

L

Lagrange dual problem, 260
Lagrange function, 260
Lagrange multiplier, 260, 270
Laplace approximation, 197
Laplace distribution, 48
Laplacian eigenmap, 433
Laplacian regularization, 377
lasso regression, 268
latent Dirichlet allocation, 229
latent variable, 202
least absolute deviations, 280
least squares, 94, 245
least squares quadratic mutual information estimation, 453
leave-one-out cross validation, 266
left singular vector, 248
left-tail probability, 16
likelihood, 92, 123
likelihood equation, 124
likelihood-ratio test, 102
linear combination, 158
linear separability, 304
linear-in-input model, 237
linear-in-parameter model, 237
local Fisher discriminant analysis, 415
local optimal solution, 161
local outlier factor, 457
local RD, 458
locality preserving projection, 410
log-likelihood, 125
logistic loss, 325
logistic regression, 321
logitboost, 354
loss, 120
lower-tail probability, 16
LS relative density ratio estimation, 385

M

machine learning, 3
madaboost, 353
Mahalanobis distance, 129
manifold, 376
Mann-Whitney U-test, 107
margin, 298, 304
margin maximization principle, 303
marginal likelihood, 95, 193
marginal probability density function, 52
marginal probability distribution, 51
marginal probability mass function, 51
marginalization, 52
Markov chain Monte Carlo, 214
Markov's inequality, 82, 140
matrix imputation, 425
matrix inversion lemma, 266
matrix learning, 7
maximum a posteriori probability estimation, 93, 189
maximum a posteriori probability rule, 117
maximum likelihood estimation, 92, 123
maximum mean discrepancy, 476
median, 17
metric learning, 7
Metropolis-Hastings sampling, 215
minimum description length, 202
minimum misclassification rate rule, 118
Minkowski's inequality, 86
missing entry, 425
mode, 18
model, 237
model misspecification, 187
model selection, 94, 148, 264
moment, 19
moment-generating function, 20
Monte Carlo method, 108, 205
Monte Carlo test, 108
Moore-Penrose pseudoinverse, 247
multiclass support vector classification, 312

multilabel classification, 395
multimodality, 157, 414
multinomial distribution, 61
multinomial theorem, 62
multiplicative model, 238
multitask learning, 391
multivariate normal distribution, 63
mutual information, 419, 452

N

nearest neighbor classifier, 180
nearest neighbor density estimation, 93, 178
necessary condition, 124
negative binomial coefficient, 34
negative binomial distribution, 33
negative correlation, 56
neural network, 243
Neyman-Pearson lemma, 102
nonlinear model, 242
nonparametric method, 92, 169
normal distribution, 37
novelty detection, 6
nuclear norm, 278
null hypothesis, 99

O

objective function, 161
Occam's razor, 151
one-sided Chebyshev's inequality, 83
one-sided probability, 16
one-sided test, 100
one-versus-one, 300
one-versus-rest, 300
online learning, 7, 355
ordinal regression, 5
orthonormality, 65, 248
outlier, 17, 180, 318
outlier detection, 6, 457
output variable, 113
overfitting, 190, 257

P

paired t-test, 106
paired z-test, 106
parametric method, 92
parametric model, 91, 123
partial derivative, 124
Parzen window function, 174
Parzen window method, 175
Pascal distribution, 35
passive-aggressive learning, 356
pattern, 113
pattern recognition, 5
pattern space, 113
Pearson divergence, 102
penalized maximum likelihood estimation, 190
permutation test, 109
point estimation, 92
Poisson distribution, 32
Poisson's law of small numbers, 32
polynomial kernel, 312
positive correlation, 56
positive definite matrix, 65
positive semidefinite matrix, 65, 142
posterior probability, 53, 185
power, 101
precision, 188
precision matrix, 478
predictor variable, 113
pretraining, 244
principal axis, 63, 130
principal component analysis, 116, 407
principal component regression, 410
principal value, 47
principle of parsimony, 151
prior probability, 53, 185
probabilistic classification, 321
probability, 13
probability and statistics, 10
probability density function, 15
probability distribution, 14
probability mass function, 14
projection, 250
proof by contradiction, 100
proposal distribution, 212
proposal point, 212
proximal gradient method, 401
proximal operator, 401
proxy distribution, 207

Q

quadratic hypersurface, 130
quadratic mutual information, 420, 453
quadratic programming, 307
quantile, 17

R

random variable, 14, 114
random walk, 216
ranking, 5

reachability distance, 457
realization, 14
recommender systems, 425
rectified linear function, 243
recursive LS, 363
regression, 4, 93, 236
regressor, 113
regularization, 94, 190
regularization matrix, 261
regularization parameter, 261
reinforcement learning, 4
rejection, 99, 213
rejection sampling, 212
relative importance, 382
reliability, 321
reproducing kernel Hilbert space, 476
reproductive property, 74
residual, 245
responsibility, 160
restricted Boltzmann machine, 441
ridge regression, 261
right singular vector, 248
right-tail probability, 16
robust support vector machine, 320
robustness, 279
root mean square, 76
row marginal total, 53

S

sample mean, 127
sample point, 11
sample space, 11
sample variance-covariance matrix, 127
sampling with replacement, 28
sampling without replacement, 28
scatter matrix, 70
Schwarz's inequality, 86
semisupervised learning, 6, 375
semisupervised local Fisher discriminant analysis, 417
sigmoidal function, 243
sign function, 210
significance level, 99
Silverman's bandwidth selector, 177
similarity, 410
simulated annealing, 162
singular value, 248
singular value decomposition, 248
skewness, 19
slack variable, 305
soft margin support vector machine, 305
sparse learning, 267
spectral clustering, 449
squared error, 140
squared hinge loss, 316, 357
standard deviation, 19
standard normal distribution, 40
standardization, 22
statistical estimation, 91
statistical pattern recognition, 114
stochastic gradient, 161
stochastic process, 214
strong law of large numbers, 77
structural change detection, 478
structured support vector machine, 339
Studentization, 96
subspace-constrained least squares, 257
sufficient condition, 124
sufficient dimensionality reduction, 419
supervised learning, 3, 236
support vector, 309
support vector data description, 458
support vector machine, 303
surrogate loss, 298
symmetric Dirichlet distribution, 70

T

Takeuchi information criterion, 153
target variable, 113
Taylor series expansion, 21, 143, 198
tensor, 134
tensor learning, 7
test statistic, 100
time-series prediction, 368
topic model, 229
total risk, 120
total scatter matrix, 408
total variation denoising, 273
trace norm, 427
training sample, 115
transfer learning, 7
trial distribution, 202
Tukey loss, 291
two-sided probability, 16
two-sided test, 100
type-I error, 101
type-II error, 101
type-II maximum likelihood estimation, 95, 193

U

unbiasedness, 140
uncorrelatedness, 56

union bound, 81
union of events, 11
unpaired t-test, 104
unpaired z-test, 104
unsupervised learning, 4
upper-tail probability, 16

V

variance, 18
variance–covariance matrix, 55
variational approximation, 202
variational Bayesian expectation-maximization algorithm, 203
variational free energy, 203
vectorization operator, 71
volume, 171
Voronoi diagram, 180

W

weak law of large numbers, 75
weighted least squares, 247
Welch's t-test, 105
whole event, 11
Wilcoxon rank-sum test, 107
Wilcoxon signed-rank test, 108
Wishart distribution, 70
within-class scatter matrix, 413
within-cluster scatter, 447