

Index

- accuracy, 421
- activation function, 404
- activations, 404
- additive, 12, 87–91, 104–105
- additivity, 307
- adjusted p -values, 584
- adjusted R^2 , 78, 227, 228, 232–235
- Advertising** data set, 15, 16, 20, 59, 61–63, 68, 69, 71–76, 79, 81, 82, 87–89, 103–105
- agglomerative clustering, 521
- Akaike information criterion, 78, 227, 228, 232–235
- alternative hypothesis, 67, 555
- analysis of variance, 314
- ANOVA, 585
- area under the curve, 151, 480–481
- argument, 43
- AUC, 151
- Auto** data set, 14, 48, 50, 56, 91–94, 123, 194, 198–200, 202, 204, 213, 215–217, 324, 401
- auto-correlation, 427
- autoregression, 430
- backfitting, 309, 325
- backpropagation, 436
- backward stepwise selection, 79, 231, 270
- bag-of- n -grams, 421
- bag-of-words, 419
- bagging, 12, 26, 327, 340–343, 351, 357–359
- BART, 340, 348, 351, 360–361
- baseline, 86, 140, 158
- basis function, 294, 297
- Bayes
 - classifier, 37–41, 142, 143
 - decision boundary, 143, 144
 - error, 37–41
- Bayes' theorem, 141, 142, 248
- Bayesian, 248–249, 350
- Bayesian additive regression trees, 327, 340, 348, 351, 360–361

- Bayesian information criterion, 78, 227, 228, 232–235
- Benjamini-Hochberg procedure, 573–575
- Bernoulli distribution, 170
- best subset selection, 227, 243, 267–270
- bias, 33–36, 65, 82, 155, 409
- bias-variance
 - decomposition, 34
 - trade-off, 33–37, 42, 105–106, 153, 155, 160, 161, 239, 252, 262, 266, 302, 331, 377, 386
- bidirectional, 431
- Bikeshare** data set, 14, 164–170, 185, 188
- binary, 28, 132
- biplot, 501, 502
- Bonferroni method, 573–575, 584
- Boolean, 175
- boosting, 12, 25, 26, 327, 340, 345–348, 351, 359–360
- bootstrap, 12, 197, 209–212, 340
- Boston** data set, 14, 57, 111, 115, 128, 195, 223, 287, 324, 356, 357, 359, 360, 363, 552
- bottom-up clustering, 521
- boxplot, 50
- BrainCancer** data set, 14, 464, 466–468, 475, 483
- branch, 329
- burn-in, 350
- C-index, 481
- Caravan** data set, 14, 182, 364
- Carseats** data set, 14, 119, 124, 353, 363
- categorical, 3, 28
- censored data, 461–495
- censoring
 - independent, 463
 - interval, 464
 - left, 463
 - mechanism, 463
 - non-informative, 463
 - right, 463
 - time, 462
- chain rule, 436
- channel, 411
- CIFAR100** data set, 411, 414–417, 448
- classification, 3, 12, 28–29, 37–42, 129–195, 367–383
 - error rate, 335
 - tree, 335–338, 353–356
- classifier, 129
- cluster analysis, 26–28
- clustering, 4, 26–28, 516–532
 - agglomerative, 521
 - bottom-up, 521
 - hierarchical, 517, 521–532
 - K*-means, 12, 517–520
- Cochran-Mantel-Haenszel test, 467
- coefficient, 61
- College** data set, 14, 54, 286, 324
- collinearity, 99–104
- conditional probability, 37
- confidence interval, 66–67, 81, 82, 104, 292
- confounding, 139
- confusion matrix, 148, 174
- continuous, 3
- contour, 244
- contour plot, 46
- contrast, 86
- convolution filter, 412
- convolution layer, 412
- convolutional neural network, 411–419
- correlation, 70, 74–75, 527
- count data, 164, 167
- Cox's proportional hazards model, 473, 476, 478–480
- C_p , 78, 227, 228, 232–235
- Credit** data set, 14, 83, 84, 86, 89, 90, 99, 100, 102
- cross-entropy, 410

- cross-validation, 12, 33, 36, 197–208, 227, 250, 271–274
 - k*-fold, 203–206
 - leave-one-out, 200–203
- curse of dimensionality, 107, 190, 265–266
- data augmentation, 417
- data frame, 48
- Data sets
 - Advertising**, 15, 16, 20, 59, 61–63, 68, 69, 71–76, 79, 81, 82, 87–89, 103–105
 - Auto**, 14, 48, 50, 56, 91–94, 123, 194, 198–200, 202, 204, 213, 215–217, 324, 401
 - Bikeshare**, 14, 164–170, 185, 188
 - Boston**, 14, 57, 111, 115, 128, 195, 223, 287, 324, 356, 357, 359, 360, 363, 552
 - BrainCancer**, 14, 464, 466–468, 475, 483
 - Caravan**, 14, 182, 364
 - Carseats**, 14, 119, 124, 353, 363
 - CIFAR100**, 411, 414–417, 448
 - College**, 14, 54, 286, 324
 - Credit**, 14, 83, 84, 86, 89, 90, 99, 100, 102
 - Default**, 14, 130, 131, 133, 134, 136–139, 147, 148, 150–152, 156, 157, 220, 221, 459
 - Fund**, 14, 564–567, 569, 573, 574, 584, 586, 587
 - Heart**, 336, 337, 341–345, 350, 352, 383–385
 - Hitters**, 14, 267, 274, 278, 279, 328, 329, 333–335, 364, 432, 433
 - IMDb**, 419–421, 423, 424, 426, 452, 454, 460
 - Income**, 16–18, 22–24
 - Khan**, 14, 396, 577–581, 588, 591
 - MNIST**, 407, 409, 410, 437, 439, 445, 448
 - NCI60**, 4, 5, 14, 542–544, 546, 547
 - NYSE**, 14, 429, 430, 460
 - OJ**, 14, 363, 401
 - Portfolio**, 14, 216
 - Publication**, 14, 475–481, 483, 486
 - Smarket**, 3, 14, 171, 177, 179, 180, 182, 193
 - USArrests**, 14, 501–503, 505–508, 510, 512–514, 535
 - Wage**, 1–3, 9, 10, 14, 291, 293, 295, 296, 298–301, 304, 306–308, 310, 311, 323, 324
 - Weekly**, 14, 193, 222
- decision tree, 12, 327–340
- deep learning, 403–458
- Default** data set, 14, 130, 131, 133, 134, 136–139, 147, 148, 150–152, 156, 157, 220, 221, 459
- degrees of freedom, 31, 265, 295, 296, 302
- dendrogram, 517, 521–527
- density function, 142
- dependent variable, 15
- derivative, 296, 302
- detector layer, 415
- deviance, 228
- dimension reduction, 226, 251–261
- discriminant function, 145
- discriminant method, 141–158
- dissimilarity, 527–530
- distance
 - correlation-based, 527–530, 550
 - Euclidean, 503, 504, 518, 519, 525, 527–530
- double descent, 439–443
- double-exponential distribution, 249
- dropout, 411, 438

- dummy variable, 82–86, 132, 137, 293
- early stopping, 438
- effective degrees of freedom, 302
- eigen decomposition, 500, 511
- elbow, 544
- embedding, 424
- embedding layer, 425
- ensemble, 340–352
- entropy, 335–336, 353, 361
- epochs, 437
- error
 - irreducible, 18, 32
 - rate, 37
 - reducible, 18
 - term, 16
- Euclidean distance, 503, 504, 518, 519, 525, 527–530, 550
- event time, 462
- expected value, 19
- exploratory data analysis, 498
- exponential distribution, 170
- exponential family, 170
- F-statistic, 75
- factor, 84
- factorial, 167
- failure time, 462
- false
 - discovery proportion, 151, 571
 - discovery rate, 554, 571–575, 578–580, 582
 - negative, 151, 559
 - positive, 151, 559
 - positive rate, 151, 152, 384
- family-wise error rate, 561–571, 575
- feature, 15
- feature map, 411
- feature selection, 226
- featurize, 419
- feed-forward neural network, 404
- fit, 21
- fitted value, 94
- flattening, 430
- flexible, 22
- for loop, 215
- forward stepwise selection, 78, 229–230, 270–271
- function, 43
- Fund** data set, 14, 564–567, 569, 573, 574, 584, 586, 587
- Gamma distribution, 170
- Gamma regression, 170
- Gaussian (normal) distribution, 141, 143, 145–146, 170, 557
- generalized additive model, 6, 25, 159, 289, 290, 306–311, 318
- generalized linear model, 6, 129, 164–170, 172, 214
- generative model, 141–158
- Gini index, 335–336, 343, 361
- global minimum, 434
- gradient, 435
- gradient descent, 434
- Harrell’s concordance index, 481
- hazard function, 469–471
 - baseline, 471
- hazard rate, 469
- Heart** data set, 336, 337, 341–345, 350, 352, 383–385
- heatmap, 47
- heteroscedasticity, 96–97, 166
- hidden layer, 405
- hidden units, 404
- hierarchical clustering, 521–527
 - dendrogram, 521–525
 - inversion, 526
 - linkage, 525–527
- hierarchical principle, 89
- high-dimensional, 78, 230, 262
- hinge loss, 387
- histogram, 51
- Hitters** data set, 14, 267, 274, 278, 279, 328, 329, 333–335, 364, 432, 433
- hold-out set, 198

- Holm's method, 565, 574, 584
- hypergeometric distribution, 493
- hyperplane, 368–373
- hypothesis test, 67–68, 75, 96, 554–582
- IMDb** data set, 419–421, 423, 424, 426, 452, 454, 460
- imputation, 510
- Income** data set, 16–18, 22–24
- independent variable, 15
- indicator function, 292
- inference, 17, 19
- inner product, 380, 381
- input layer, 404
- input variable, 15
- integral, 302
- interaction, 60, 81, 87–91, 104–105, 310
- intercept, 61, 63
- interpolate, 439
- interpretability, 225
- inversion, 526
- irreducible error, 18, 39, 82, 104
- joint distribution, 155
- K-means clustering, 12, 517–520
- K-nearest neighbors, 129, 160–164
 - classifier, 12, 38–41
 - regression, 105–110
- Kaplan-Meier survival curve, 464–466, 476
- kernel, 380–383, 386, 397
 - linear, 382
 - non-linear, 379–383
 - polynomial, 382, 383
 - radial, 382–384, 393
- kernel density estimator, 156
- Khan** data set, 14, 396, 577–581, 588, 591
- knot, 290, 295, 297–299
- ℓ_1 norm, 241
- ℓ_2 norm, 238
- lag, 427, 428
- Laplace distribution, 249
- lasso, 12, 25, 241–249, 264–265, 333, 387, 478
- leaf, 329, 522
- learning rate, 436
- least squares, 6, 21, 61–63, 135, 225
 - line, 63
 - weighted, 97
- level, 84
- leverage, 98–99
- likelihood function, 135
- linear, 2, 59–110
- linear combination, 123, 226, 251, 499
- linear discriminant analysis, 6, 12, 129, 132, 142–151, 161–164, 377, 383
- linear kernel, 382
- linear model, 20, 21, 59–110
- linear regression, 6, 12, 59–110, 170
 - multiple, 71–82
 - simple, 60–71
- link function, 170
- linkage, 525–527, 545
 - average, 525–527
 - centroid, 525–527
 - complete, 522, 525–527
 - single, 525–527
- local minimum, 434
- local regression, 290, 318
- log odds, 140
- log-rank test, 466–469, 476
- logistic function, 134
- logistic regression, 6, 12, 26, 129, 133–139, 161–164, 170, 310–311, 378, 386–387
 - multinomial, 140, 160
 - multiple, 137–139
- logit, 135, 315
- loss function, 302, 386
- low-dimensional, 261
- LSTM RNN, 426

- main effects, 88, 89
- majority vote, 341
- Mallow's C_p , 78, 227, 228, 232–235
- Mantel-Haenszel test, 467
- margin, 371, 387
- marginal distribution, 155
- Markov chain Monte Carlo, 350
- matrix completion, 510
- matrix multiplication, 12
- maximal margin
 - classifier, 367–373
 - hyperplane, 371
- maximum likelihood, 134–137, 168
- mean squared error, 29
- minibatch, 436
- misclassification error, 37
- missing at random, 511
- missing data, 50, 510–516
- mixed selection, 79
- MNIST** data set, 407, 409, 410, 437, 439, 445, 448
- model assessment, 197
- model selection, 197
- multicollinearity, 102, 266
- multinomial logistic regression, 140, 160
- multiple testing, 553–582
- multi-task learning, 408
- multivariate Gaussian, 145–146
- multivariate normal, 145–146

- naive Bayes, 129, 153–158, 161–164
- natural spline, 298, 302, 317
- NCI60** data set, 4, 5, 14, 542–544, 546, 547
- negative binomial distribution, 170
- negative binomial regression, 170
- negative predictive value, 151, 152
- neural network, 6, 403–458
- node
 - internal, 329
 - purity, 335–336
 - terminal, 329
- noise, 22, 250
- non-linear, 2, 12, 289–326
 - decision boundary, 379–383
 - kernel, 379–383
- non-parametric, 21, 23–24, 105–110, 190
- normal (Gaussian) distribution, 141, 143, 145–146, 170, 468, 557
- null, 148
 - distribution, 557, 576
 - hypothesis, 67, 555
 - model, 78, 227, 242
- NYSE** data set, 14, 429, 430, 460

- Occam's razor, 432
- odds, 134, 140, 192
- OJ** data set, 14, 363, 401
- one-hot encoding, 84, 407, 446, 449, 452
- one-standard-error rule, 236
- one-versus-all, 385
- one-versus-one, 385
- optimal separating hyperplane, 371
- optimism of training error, 32
- ordered categorical variable, 316
- orthogonal, 256, 501
 - basis, 312
- out-of-bag, 342–343
- outlier, 97–98
- output variable, 15
- over-parametrized, 459
- overdispersion, 169
- overfitting, 22, 24, 26, 32, 80, 148, 229, 371

- p-value, 67–68, 74, 556–558, 576–578
 - adjusted, 584, 585
- parameter, 61
- parametric, 21–23, 105–110
- partial least squares, 252, 259–261, 281, 282
- partial likelihood, 473
- path algorithm, 247

- permutation, 576
- permutation approach, 575–582
- perpendicular, 256
- pipe, 444
- Poisson distribution, 167, 170
- Poisson regression, 129, 164–170
- polynomial
 - kernel, 382, 383
 - regression, 91–92, 289–292, 295
- pooling, 415
- population regression line, 63
- Portfolio** data set, 14, 216
- positive predictive value, 151, 152
- posterior
 - distribution, 248
 - mode, 249
 - probability, 142
- power, 101, 151, 559
- precision, 151
- prediction, 17
 - interval, 82, 104
- predictor, 15
- principal components, 499
 - analysis, 12, 252–259, 498–510
 - loading vector, 500
 - missing values, 510–516
 - proportion of variance explained, 505–510, 543
 - regression, 12, 252–259, 279–281, 498–499, 510
 - score vector, 500
 - scree plot, 509–510
- prior
 - distribution, 248
 - probability, 142
- probability density function, 469, 470
- projection, 226
- proportional hazards assumption, 471
- pruning, 331–333
 - cost complexity, 331–333
 - weakest link, 331–333
- Publication** data set, 14, 475–481, 483, 486
- q-value, 586
- quadratic, 92
- quadratic discriminant analysis, 4, 129, 152–153, 161–164
- qualitative, 3, 28, 82, 129, 164, 198
 - variable, 82–86
- quantitative, 3, 28, 82, 129, 164, 198
- R functions
 - %>%,** 444
 - abline(),** 113, 123, 326, 545
 - all.equal(),** 187
 - anova(),** 117, 314, 315
 - apply(),** 273, 532, 533
 - as.dist(),** 541
 - as.factor(),** 50
 - as.raster(),** 449
 - attach(),** 50
 - BART,** 360
 - biplot(),** 534
 - boot(),** 216–218, 221
 - bs(),** 317, 324
 - c(),** 43
 - cbind(),** 182, 313
 - coef(),** 112, 173, 270, 275
 - compile(),** 445
 - confint(),** 112
 - contour(),** 46, 47
 - contrasts(),** 120, 174
 - cor(),** 45, 124, 171, 550
 - coxph(),** 484
 - cumsum(),** 535
 - cut(),** 316
 - cutree(),** 541
 - cv.glm(),** 214, 215, 222
 - cv.glmnet(),** 277
 - cv.tree(),** 355, 357, 364
 - data.frame(),** 194, 223, 286, 353
 - dataset_mnist(),** 445
 - dev.off(),** 46
 - dim(),** 48, 50
 - dist(),** 540, 550

`for()`, 215
`gam()`, 309, 318, 320
`gbart()`, 360
`gbm()`, 359
`glm()`, 172, 177, 214, 221, 315
`glmnet()`, 274, 275, 277, 278, 453
`hatvalues()`, 114
`hclust()`, 540, 541
`head()`, 49
`hist()`, 51, 56
`I()`, 116, 313, 315, 321
`identify()`, 51
`ifelse()`, 353
`image()`, 47
`importance()`, 358, 363
`is.na()`, 267
`jitter()`, 316
`jpeg`, 449
`jpeg()`, 46
`keras_model_sequential()`, 444
`kmeans()`, 538–540
`knn()`, 181, 182
`layer_conv_2D()`, 450
`layer_dense()`, 446
`layer_dropout()`, 446
`lbart()`, 360
`lda()`, 177, 179, 180
`legend()`, 126
`length()`, 43
`library()`, 110, 111
`lines()`, 113
`lm()`, 111, 113, 114, 116, 118, 123, 124, 172, 177, 213, 214, 277, 279, 312, 318, 353, 444, 456
`lo()`, 320
`loadhistory()`, 52
`loess()`, 318
`ls()`, 43
`matrix()`, 44
`mean()`, 45, 174, 213, 532
`median()`, 194
`mfrow()`, 113
`model.matrix()`, 272, 274, 444
`na.omit()`, 50, 267
`naiveBayes()`, 180
`names()`, 50, 112
`ns()`, 317
`p.adjust()`, 584
`pairs()`, 51, 55
`par()`, 113, 313
`pbart()`, 360
`pcr()`, 279–281
`pdf()`, 46
`persp()`, 47
`plot()`, 45, 46, 50, 56, 113, 123, 269, 319, 354, 389, 390, 401, 448, 540, 543
`plot.Gam()`, 319
`plsr()`, 281
`points()`, 269
`poly()`, 118, 213, 312–314, 324
`prcomp()`, 533, 534, 550
`predict()`, 113, 173, 177–179, 181, 213, 272, 273, 276, 277, 313, 315, 316, 320, 354, 355, 391, 394, 395
`prune.misclass()`, 355
`prune.tree()`, 357
`q()`, 52
`qda()`, 179, 180
`quantile()`, 223
`rainbow()`, 543
`randomForest()`, 357, 358
`range()`, 56
`read.csv()`, 49, 55, 552
`read.table()`, 48, 49
`regsubsets()`, 268, 270–272, 286
`residuals()`, 114
`return()`, 195
`rm()`, 43
`rnorm()`, 45, 126, 285, 551
`rstudent()`, 114
`runif()`, 551
`s()`, 318
`sample()`, 213, 216, 549
`savehistory()`, 52

- `scale()`, 183, 444, 541, 552
- `sd()`, 45
- `seq()`, 46
- `set.seed()`, 45, 213, 540
- `sim.survdata()`, 487
- `smooth.spline()`, 317, 318
- `softImpute`, 538
- `sqrt()`, 44, 45
- `sum()`, 267
- `summary()`, 51, 55, 56, 114, 123, 124, 173, 218, 221, 268, 269, 279, 280, 319, 353, 356, 359, 363, 390, 391, 393, 402, 543
- `survdifff()`, 484
- `survfit()`, 483
- `svd()`, 535
- `svm()`, 389, 390, 392, 393, 395, 396
- `t.test()`, 582
- `table()`, 174, 551
- `text()`, 354
- `title()`, 313
- `tree()`, 328, 353
- `TukeyHSD()`, 585
- `tune()`, 390, 391, 394, 402
- `update()`, 116
- `validationplot()`, 280
- `var()`, 45
- `varImpPlot()`, 359
- `View()`, 48, 55
- `vif()`, 115
- `which.max()`, 114, 269
- `which.min()`, 269
- `with()`, 443
- `write.table()`, 48
- radial kernel, 382–384, 393
- random forest, 12, 327, 340, 343–345, 351, 357–359
- re-sampling, 575–582
- recall, 151
- receiver operating characteristic (ROC), 150, 383–384
- recommender systems, 511
- rectified linear unit, 405
- recurrent neural network, 421–433
- recursive binary splitting, 330, 333, 335
- reducible error, 18, 81
- regression, 3, 12, 28–29
 - local, 289, 290, 304–306
 - piecewise polynomial, 295
 - polynomial, 289–292, 300
 - spline, 290, 294, 317
 - tree, 328–334, 356–357
- regularization, 226, 237, 411, 478–480
- ReLU, 405
- resampling, 197–212
- residual, 61, 72
 - plot, 93
 - standard error, 66, 68–69, 79–80, 103
 - studentized, 98
 - sum of squares, 62, 70, 72
- residuals, 262, 346
- response, 15
- ridge regression, 12, 237–241, 387, 478
- risk set, 465
- robust, 374, 377, 532
- ROC curve, 150, 383–384, 480–481
- R^2 , 68–71, 79–80, 103, 234
- rug plot, 316
- scale equivariant, 239
- scatterplot, 50
- Scheffé’s method, 569
- scree plot, 506, 509–510, 544
 - elbow, 509
- seed, 213
- semi-supervised learning, 28
- sensitivity, 149, 151
- separating hyperplane, 368–373
- Seq2Seq, 431
- shrinkage, 226, 237, 478–480
 - penalty, 237
- sigmoid, 405
- signal, 250

- singular value decomposition, 535
- slack variable, 376
- slope, 61, 63
- Smarket** data set, 3, 14, 171, 177, 179, 180, 182, 193
- smoother, 310
- smoothing spline, 290, 301–304, 317
- soft margin classifier, 373–375
- soft-thresholding, 248
- softmax, 141, 410
- sparse, 242, 250
- sparse matrix format, 420
- sparsity, 242
- specificity, 149, 151
- spline, 289, 295–304
 - cubic, 297
 - linear, 297
 - natural, 298, 302
 - regression, 290, 295–300
 - smoothing, 31, 290, 301–304
 - thin-plate, 23
- standard error, 65, 94
- standardize, 183
- statistical model, 1
- step function, 105, 289, 292–294
- stepwise model selection, 12, 227, 229
- stochastic gradient descent, 436
- stump, 347
- subset selection, 226–236
- subtree, 331
- supervised learning, 26–28, 260
- support vector, 371, 377, 387
 - classifier, 367, 373–378
 - machine, 6, 12, 26, 379–388
 - regression, 388
- survival
 - analysis, 461–495
 - curve, 464, 477
 - function, 464
 - time, 462
- synergy, 60, 81, 87–91, 104–105
- systematic, 16
- t-distribution, 67, 162
- t-statistic, 67
- t-test
 - one-sample, 582, 586
 - paired, 585
 - two-sample, 556, 568, 576–580, 582, 588
- test
 - error, 37, 41, 175
 - MSE, 29–34
 - observations, 30
 - set, 32
 - statistic, 556
- theoretical null distribution, 575
- time series, 94
- total sum of squares, 70
- tracking, 95
- train, 21
- training
 - data, 21
 - error, 37, 41, 175
 - MSE, 29–33
- tree, 327–340
- tree-based method, 327
- true negative, 151
- true positive, 151
- true positive rate, 151, 152, 384
- truncated power basis, 297
- Tukey's method, 568, 584, 585
- tuning parameter, 237, 478
- two-sample *t*-test, 466
- Type I error, 151, 559–561
- Type I error rate, 559
- Type II error, 151, 559, 565, 582
- unsupervised learning, 26–28, 253, 259, 497–547
- USArrests** data set, 14, 501–503, 505–508, 510, 512–514, 535
- validation set, 198
 - approach, 198–200
- variable, 15
 - dependent, 15
 - dummy, 82–86, 89–91

- importance, 343, 358
- independent, 15
- indicator, 37
- input, 15
- output, 15
- qualitative, 82–86, 89–91
- selection, 78, 226, 242
- variance, 19, 33–36, 155
 - inflation factor, 102–104, 115
- varying coefficient model, 306
- vector, 43
- Wage** data set, 1–3, 9, 10, 14, 291, 293, 295, 296, 298–301, 304, 306–308, 310, 311, 323, 324
- weak learner, 340
- weakest link pruning, 332
- Weekly** data set, 14, 193, 222
- weight freezing, 419, 425
- weight sharing, 423
- weighted least squares, 97, 306
- weights, 408
- with replacement, 211
- within class covariance, 146
- workspace, 52
- wrapper, 313