

Index

- C_p , 78, 205, 206, 210–213
- R^2 , 68–71, 79–80, 103, 212
- ℓ_2 norm, 216
- ℓ_1 norm, 219

- additive, 12, 86–90, 104
- additivity, 282, 283
- adjusted R^2 , 78, 205, 206, 210–213
- Advertising** data set, 15, 16, 20, 59, 61–63, 68, 69, 71–76, 79, 81, 82, 87, 88, 102–104
- agglomerative clustering, 390
- Akaike information criterion, 78, 205, 206, 210–213
- alternative hypothesis, 67
- analysis of variance, 290
- area under the curve, 147
- argument, 42
- AUC, 147
- Auto** data set, 14, 48, 49, 56, 90–93, 121, 122, 171, 176–178, 180, 182, 191, 193–195, 299, 371

- backfitting, 284, 300
- backward stepwise selection, 79, 208–209, 247
- bagging, 12, 26, 303, 316–319, 328–330
- baseline, 86
- basis function, 270, 273
- Bayes
 - classifier, 37–40, 139
 - decision boundary, 140
 - error, 37–40
- Bayes’ theorem, 138, 139, 226
- Bayesian, 226–227
- Bayesian information criterion, 78, 205, 206, 210–213
- best subset selection, 205, 221, 244–247
- bias, 33–36, 65, 82
- bias-variance
 - decomposition, 34
 - trade-off, 33–37, 42, 105, 149, 217, 230, 239, 243, 278, 307, 347, 357
- binary, 28, 130
- biplot, 377, 378

- Boolean, 159
- boosting, 12, 25, 26, 303, 316, 321–324, 330–331
- bootstrap, 12, 175, 187–190, 316
- Boston** data set, 14, 56, 110, 113, 126, 173, 201, 264, 299, 327, 328, 330, 333
- bottom-up clustering, 390
- boxplot, 50
- branch, 305

- Caravan** data set, 14, 164, 335
- Carseats** data set, 14, 117, 123, 324, 333
- categorical, 3, 28
- classification, 3, 12, 28–29, 37–42, 127–173, 337–353
 - error rate, 311
 - tree, 311–314, 324–327
- classifier, 127
- cluster analysis, 26–28
- clustering, 4, 26–28, 385–401
 - K*-means, 12, 386–389
 - agglomerative, 390
 - bottom-up, 390
 - hierarchical, 386, 390–401
- coefficient, 61
- College** data set, 14, 54, 263, 300
- collinearity, 99–103
- conditional probability, 37
- confidence interval, 66–67, 81, 82, 103, 268
- confounding, 136
- confusion matrix, 145, 158
- continuous, 3
- contour plot, 46
- contrast, 86
- correlation, 70, 74, 396
- Credit** data set, 83, 84, 86, 89, 90, 99–102
- cross-entropy, 311–312, 332
- cross-validation, 12, 33, 36, 175–186, 205, 227, 248–251
 - k*-fold, 181–184
 - leave-one-out, 178–181
- curse of dimensionality, 108, 168, 242–243

- data frame, 48
- Data sets
 - Advertising**, 15, 16, 20, 59, 61–63, 68, 69, 71–76, 79, 81, 82, 87, 88, 102–104
 - Auto**, 14, 48, 49, 56, 90–93, 121, 122, 171, 176–178, 180, 182, 191, 193–195, 299, 371
 - Boston**, 14, 56, 110, 113, 126, 173, 201, 264, 299, 327, 328, 330, 333
 - Caravan**, 14, 164, 335
 - Carseats**, 14, 117, 123, 324, 333
 - College**, 14, 54, 263, 300
 - Credit**, 83, 84, 86, 89, 90, 99–102
 - Default**, 14, 128–137, 144–148, 198, 199
 - Heart**, 312, 313, 317–320, 354, 355
 - Hitters**, 14, 244, 251, 255, 256, 304, 305, 310, 311, 334
 - Income**, 16–18, 22–24
 - Khan**, 14, 366
 - NCI60**, 4, 5, 14, 407, 409–412
 - OJ**, 14, 334, 371
 - Portfolio**, 14, 194
 - Smarket**, 3, 14, 154, 161, 162, 171
 - USArrests**, 14, 377, 378, 381–383

- Wage**, 1, 2, 9, 10, 14, 267, 269, 271, 272, 274–277, 280, 281, 283, 284, 286, 287, 299
- Weekly**, 14, 171, 200
- decision tree, 12, 303–316
- Default** data set, 14, 128–137, 144–148, 198, 199
- degrees of freedom, 32, 241, 271, 272, 278
- dendrogram, 386, 390–396
- density function, 138
- dependent variable, 15
- derivative, 272, 278
- deviance, 206
- dimension reduction, 204, 228–238
- discriminant function, 141
- dissimilarity, 396–398
- distance
 - correlation-based, 396–398, 416
 - Euclidean, 379, 387, 388, 394, 396–398
- double-exponential distribution, 227
- dummy variable, 82–86, 130, 134, 269
- effective degrees of freedom, 278
- elbow, 409
- error
 - irreducible, 18, 32
 - rate, 37
 - reducible, 18
 - term, 16
- Euclidean distance, 379, 387, 388, 394, 396–398, 416
- expected value, 19
- exploratory data analysis, 374
- F-statistic, 75
- factor, 84
- false discovery proportion, 147
- false negative, 147
- false positive, 147
- false positive rate, 147, 149, 354
- feature, 15
- feature selection, 204
- Fisher’s linear discriminant, 141
- fit, 21
- fitted value, 93
- flexible, 22
- for loop, 193
- forward stepwise selection, 78, 207–208, 247
- function, 42
- Gaussian (normal) distribution, 138, 139, 142–143
- generalized additive model, 6, 26, 265, 266, 282–287, 294
- generalized linear model, 6, 156, 192
- Gini index, 311–312, 319, 332
- Heart** data set, 312, 313, 317–320, 354, 355
- heatmap, 47
- heteroscedasticity, 95–96
- hierarchical clustering, 390–396
 - dendrogram, 390–394
 - inversion, 395
 - linkage, 394–396
- hierarchical principle, 89
- high-dimensional, 78, 208, 239
- hinge loss, 357
- histogram, 50
- Hitters** data set, 14, 244, 251, 255, 256, 304, 305, 310, 311, 334
- hold-out set, 176
- hyperplane, 338–343
- hypothesis test, 67–68, 75, 95
- Income** data set, 16–18, 22–24
- independent variable, 15
- indicator function, 268
- inference, 17, 19

- inner product, 351
- input variable, 15
- integral, 278
- interaction, 60, 81, 87–90, 104, 286
- intercept, 61, 63
- interpretability, 203
- inversion, 395
- irreducible error, 18, 39, 82, 103
- K-means clustering, 12, 386–389
- K-nearest neighbors
 - classifier, 12, 38–40, 127
 - regression, 104–109
- kernel, 350–353, 356, 367
 - linear, 352
 - non-linear, 349–353
 - polynomial, 352, 354
 - radial, 352–354, 363
- kernel trick, 351
- Khan** data set, 14, 366
- knot, 266, 271, 273–275
- Laplace distribution, 227
- lasso, 12, 25, 219–227, 241–242, 309, 357
- leaf, 305, 391
- least squares, 6, 21, 61–63, 133, 203
 - line, 63
 - weighted, 96
- level, 84
- leverage, 97–99
- likelihood function, 133
- linear, 2, 86
- linear combination, 121, 204, 229, 375
- linear discriminant analysis, 6, 12, 127, 130, 138–147, 348, 354
- linear kernel, 352
- linear model, 20, 21, 59
- linear regression, 6, 12
 - multiple, 71–82
 - simple, 61–71
- linkage, 394–396, 410
 - average, 394–396
 - centroid, 394–396
 - complete, 391, 394–396
 - single, 394–396
- local regression, 266, 294
- logistic
 - function, 132
- logistic regression, 6, 12, 26, 127, 131–137, 286–287, 349, 356–357
 - multiple, 135–137
- logit, 132, 286, 291
- loss function, 277, 357
- low-dimensional, 238
- main effects, 88, 89
- majority vote, 317
- Mallow's C_p , 78, 205, 206, 210–213
- margin, 341, 357
- matrix multiplication, 12
- maximal margin
 - classifier, 337–343
 - hyperplane, 341
- maximum likelihood, 132–133, 135
- mean squared error, 29
- misclassification error, 37
- missing data, 49
- mixed selection, 79
- model assessment, 175
- model selection, 175
- multicollinearity, 243
- multivariate Gaussian, 142–143
- multivariate normal, 142–143
- natural spline, 274, 278, 293
- NCI60** data set, 4, 5, 14, 407, 409–412
- negative predictive value, 147, 149
- node
 - internal, 305
 - purity, 311–312
 - terminal, 305

- noise, 22, 228
- non-linear, 2, 12, 265–301
 - decision boundary, 349–353
 - kernel, 349–353
- non-parametric, 21, 23–24, 104–109, 168
- normal (Gaussian) distribution, 138, 139, 142–143
- null, 145
 - hypothesis, 67
 - model, 78, 205, 220
- odds, 132, 170
- OJ** data set, 14, 334, 371
- one-standard-error rule, 214
- one-versus-all, 356
- one-versus-one, 355
- optimal separating hyperplane, 341
- optimism of training error, 32
- ordered categorical variable, 292
- orthogonal, 233, 377
 - basis, 288
- out-of-bag, 317–318
- outlier, 96–97
- output variable, 15
- overfitting, 22, 24, 26, 32, 80, 144, 207, 341
- p-value, 67–68, 73
- parameter, 61
- parametric, 21–23, 104–109
- partial least squares, 12, 230, 237–238, 258, 259
- path algorithm, 224
- perpendicular, 233
- polynomial
 - kernel, 352, 354
 - regression, 90–92, 265–268, 271
- population regression line, 63
- Portfolio** data set, 14, 194
- positive predictive value, 147, 149
- posterior
 - distribution, 226
 - mode, 226
 - probability, 139
- power, 101, 147
- precision, 147
- prediction, 17
 - interval, 82, 103
- predictor, 15
- principal components, 375
 - analysis, 12, 230–236, 374–385
 - loading vector, 375, 376
 - proportion of variance explained, 382–384, 408
 - regression, 12, 230–236, 256–257, 374–375, 385
 - score vector, 376
 - scree plot, 383–384
- prior
 - distribution, 226
 - probability, 138
- projection, 204
- pruning, 307–309
 - cost complexity, 307–309
 - weakest link, 307–309
- quadratic, 91
- quadratic discriminant analysis, 4, 149–150
- qualitative, 3, 28, 127, 176
 - variable, 82–86
- quantitative, 3, 28, 127, 176
- R functions
 - x²**, 125
 - abline()**, 112, 122, 301, 412
 - anova()**, 116, 290, 291
 - apply()**, 250, 401
 - as.dist()**, 407
 - as.factor()**, 50
 - attach()**, 50
 - biplot()**, 403
 - boot()**, 194–196, 199

`bs()`, 293, 300
`c()`, 43
`cbind()`, 164, 289
`coef()`, 111, 157, 247, 251
`confint()`, 111
`contour()`, 46
`contrasts()`, 118, 157
`cor()`, 44, 122, 155, 416
`cumsum()`, 404
`cut()`, 292
`cutree()`, 406
`cv.glm()`, 192, 193, 199
`cv.glmnet()`, 254
`cv.tree()`, 326, 328, 334
`data.frame()`, 171, 201, 262, 324
`dev.off()`, 46
`dim()`, 48, 49
`dist()`, 406, 416
`fix()`, 48, 54
`for()`, 193
`gam()`, 284, 294, 296
`gbm()`, 330
`glm()`, 156, 161, 192, 199, 291
`glmnet()`, 251, 253–255
`hatvalues()`, 113
`hclust()`, 406, 407
`hist()`, 50, 55
`I()`, 115, 289, 291, 296
`identify()`, 50
`ifelse()`, 324
`image()`, 46
`importance()`, 330, 333, 334
`is.na()`, 244
`jitter()`, 292
`jpeg()`, 46
`kmeans()`, 404, 405
`knn()`, 163, 164
`lda()`, 161, 162
`legend()`, 125
`length()`, 43
`library()`, 109, 110
`lines()`, 112
`lm()`, 110, 112, 113, 115, 116, 121, 122, 156, 161, 191, 192, 254, 256, 288, 294, 324
`lo()`, 296
`loadhistory()`, 51
`loess()`, 294
`ls()`, 43
`matrix()`, 44
`mean()`, 45, 158, 191, 401
`median()`, 171
`model.matrix()`, 251
`na.omit()`, 49, 244
`names()`, 49, 111
`ns()`, 293
`pairs()`, 50, 55
`par()`, 112, 289
`pcr()`, 256, 258
`pdf()`, 46
`persp()`, 47
`plot()`, 45, 46, 49, 55, 112, 122, 246, 295, 325, 360, 371, 406, 408
`plot.gam()`, 295
`plot.svm()`, 360
`plsr()`, 258
`points()`, 246
`poly()`, 116, 191, 288–290, 299
`prcomp()`, 402, 403, 416
`predict()`, 111, 157, 160, 162, 163, 191, 249, 250, 252, 253, 289, 291, 292, 296, 325, 327, 361, 364, 365
`print()`, 172
`prune.misclass()`, 327
`prune.tree()`, 328
`q()`, 51
`qda()`, 162
`quantile()`, 201
`rainbow()`, 408
`randomForest()`, 329
`range()`, 56
`read.csv()`, 49, 54, 418

- `read.table()`, 48, 49
- `regsubsets()`, 244–249, 262
- `residuals()`, 112
- `return()`, 172
- `rm()`, 43
- `rnorm()`, 44, 45, 124, 262, 417
- `rstudent()`, 112
- `runif()`, 417
- `s()`, 294
- `sample()`, 191, 194, 414
- `savehistory()`, 51
- `scale()`, 165, 406, 417
- `sd()`, 45
- `seq()`, 46
- `set.seed()`, 45, 191, 405
- `smooth.spline()`, 293, 294
- `sqrt()`, 44, 45
- `sum()`, 244
- `summary()`, 51, 55, 113, 121, 122, 157, 196, 199, 244, 245, 256, 257, 295, 324, 325, 328, 330, 334, 360, 361, 363, 372, 408
- `svm()`, 359–363, 365, 366
- `table()`, 158, 417
- `text()`, 325
- `title()`, 289
- `tree()`, 304, 324
- `tune()`, 361, 364, 372
- `update()`, 114
- `var()`, 45
- `varImpPlot()`, 330
- `vif()`, 114
- `which.max()`, 113, 246
- `which.min()`, 246
- `write.table()`, 48
- radial kernel, 352–354, 363
- random forest, 12, 303, 316, 320–321, 328–330
- recall, 147
- receiver operating characteristic (ROC), 147, 354–355
- recursive binary splitting, 306, 309, 311
- reducible error, 18, 81
- regression, 3, 12, 28–29
 - local, 265, 266, 280–282
 - piecewise polynomial, 271
 - polynomial, 265–268, 276–277
 - spline, 266, 270, 293
 - tree, 304–311, 327–328
- regularization, 204, 215
- replacement, 189
- resampling, 175–190
- residual, 62, 72
 - plot, 92
 - standard error, 66, 68–69, 79–80, 102
 - studentized, 97
 - sum of squares, 62, 70, 72
- residuals, 239, 322
- response, 15
- ridge regression, 12, 215–219, 357
- robust, 345, 348, 400
- ROC curve, 147, 354–355
- rug plot, 292
- scale invariant, 217
- scatterplot, 49
- scatterplot matrix, 50
- scree plot, 383–384, 409
 - elbow, 384
- seed, 191
- semi-supervised learning, 28
- sensitivity, 145, 147
- separating hyperplane, 338–343
- shrinkage, 204, 215
 - penalty, 215
- signal, 228
- slack variable, 346
- slope, 61, 63
- Smarket** data set, 3, 14, 154, 161, 162, 171
- smoother, 286
- smoothing spline, 266, 277–280, 293
- soft margin classifier, 343–345

- soft-thresholding, 225
- sparse, 219, 228
- sparsity, 219
- specificity, 145, 147, 148
- spline, 265, 271–280
 - cubic, 273
 - linear, 273
 - natural, 274, 278
 - regression, 266, 271–277
 - smoothing, 31, 266, 277–280
 - thin-plate, 23
- standard error, 65, 93
- standardize, 165
- statistical model, 1
- step function, 105, 265, 268–270
- stepwise model selection, 12, 205, 207
- stump, 323
- subset selection, 204–214
- subtree, 308
- supervised learning, 26–28, 237
- support vector, 342, 347, 357
 - classifier, 337, 343–349
 - machine, 12, 26, 349–359
 - regression, 358
- synergy, 60, 81, 87–90, 104
- systematic, 16
- t-distribution, 67, 153
- t-statistic, 67
- test
 - error, 37, 40, 158
 - MSE, 29–34
 - observations, 30
 - set, 32
- time series, 94
- total sum of squares, 70
- tracking, 94
- train, 21
- training
 - data, 21
 - error, 37, 40, 158
 - MSE, 29–33
- tree, 303–316
- tree-based method, 303
- true negative, 147
- true positive, 147
- true positive rate, 147, 149, 354
- truncated power basis, 273
- tuning parameter, 215
- Type I error, 147
- Type II error, 147
- unsupervised learning, 26–28, 230, 237, 373–413
- USArrests** data set, 14, 377, 378, 381–383
- validation set, 176
 - approach, 176–178
- variable, 15
 - dependent, 15
 - dummy, 82–86, 89–90
 - importance, 319, 330
 - independent, 15
 - indicator, 37
 - input, 15
 - output, 15
 - qualitative, 82–86, 89–90
 - selection, 78, 204, 219
- variance, 19, 33–36
 - inflation factor, 101–103, 114
- varying coefficient model, 282
- vector, 43
- Wage** data set, 1, 2, 9, 10, 14, 267, 269, 271, 272, 274–277, 280, 281, 283, 284, 286, 287, 299
- weakest link pruning, 308
- Weekly** data set, 14, 171, 200
- weighted least squares, 96, 282
- within class covariance, 143
- workspace, 51
- wrapper, 289