# Index

activation function, 218  
adjusted coefficient of determination, 349  
adjusted $R^2$, 368  
agglomerative hierarchical clustering, 253  
akaike information criterion, 451  
Anderson–Darling test, 55  
ANOVA, 662  
application score, 556  
apriori algorithm, 289  
arcing, 525  
Arc-x4, 526  
area under the ROC curve, 546  
atificial ants, 513  
association rules, 287  
  lift of, 288  
asymptotic test, 652  
attitudinal data, 96  
attribution score, 556  
augmentation method, 539  
average linkage, 257  

backward selection, 85, 396  
bagging, 518–521  
Bartlett test, 58  
Basel II ratio, 557  
Bayesian classification rule, 342  
Bayesian information criterion, 451  
Bayesian networks, 497–499  
Bayesian statistics, 492  
Bayes’ theorem, 492  
behavioural mega-databases, 97  
best linear unbiased estimator, 358  
between-class variance, 662  
between-cluster sum of squares, 244  
bias–variance dilemma, 310  
binarization, 80  
binary logistic regression, 437  
biomimetic algorithms, 513  
BIRCH algorithm, 262  
block level, 100  
Bonferroni method, 326  
boosting, 521–528  
bootstrap, 516  
Box–Cox transformation, 74  
Box plot, 649  
BRIDGE algorithm, 262  
Burt table, 201  
canonical correlation analysis, 658  
CART tree, 321  
categories of the variable, 27  
Cattell’s scree test, 190  
ecauchit, 482  
censored data, 664  
central tendency characteristics, 648  
centroid method, 257  
CHAID tree, 325–327  
chain effect, 254  
Chuprov’s coefficient, 662  
circle of correlation, 181  
CLARA algorithm, 249  
CLARANS algorithm, 249  
clustering, 235  
cluster sampling, 91  
coefficient of determination, 365  
coefficient of multiple correlation, 365  
combined stepwise selection, 85  
common log format, 638  
complete linkage, 254  
cordance tests, 458  
conditional likelihood, 446  
condition indices, 88  
Condorcet criterion, 274  
Condorcet’s paradox, 274  
confidence index, 288  
confidence interval for mean, 652  
confidence interval of frequency, 654  
cosnfusion matrix, 541
conjugate gradient descent, 224  
conservative test, 651  
contingency table, 45  
contribution of an individual, 186  
of a variable, 183  
cookies, 641  
Cook’s distance, 375  
correspondence analysis, 194  
covariance, 177  
covariate, 434  
Cox–Snell $R^2$, 455  
Cox’s proportional hazards regression model, 665  
Cramér’s coefficient, 661  
credit bureau, 565  
credit scoring history, 615–616  
cross-validation, 306  
$\chi^2$ distance, 195  
$\chi^2$ test, 659  
C4.5 tree, 324  
C5.0 tree, 324  
cubic clustering criterion, 245  
customer relationship management (CRM), 2  

table, 44  

data frame, 370  
DBSCAN algorithm, 262  
decision tree, 313  
dendrogram, 253  
density estimation methods, 258  
dependent variable, 301  
deviance, 449  
discrete AdaBoost, 524  
discretizing, 32  
discriminant factor analysis, 333–338  
discriminant linear function, 339  
dispersion characteristics, 648  
DISQUAL method, 353  
divisive hierarchical clustering, 238  
Durbin–Watson statistic, 374  
dynamic clouds, 248  

Efron-Tibshirani formula, 517  
elastic net, 364  
entropy, 318  
equal (parallel) slopes model, 482  
error function, 225  
error rate, 304  
error surface, 226  
exact test, 652  

extended log format, 638  
FactoMineR package, 131  
factor axis, 178  
factorial plane, 178  
factors, 434  
in success of project, 23  
FICO scores, 565  
first name scoring method, 95  
Fisher algorithm, 448  
Fisher–Snedecor test, 663  
Fisher test, 58  
exact test, 659  
fitness function, 512  
fixed effects, 434  
forward selection, 85, 396  
frequency table, 44  
fuzzy clustering, 238  

GEE method, 480  
generalized additive model, 491–492  
generalized linear model, 484–487  
generalized logit, 482  
general linear model, 434  
general regression neural networks, 224  
genetic algorithms, 510–514  
geocoding, 105  
geographical information software, 102  
geomarketing, 99  
geometric predictive discriminant analysis, 338–341  
Gini index, 551  
gradient back-propagation, 224  
graphic forms, 630  
grouped lasso, 365  
heteroscedasticity, 58  
hidden layer, 218  
HOMALS, 160  
homoscedasticity, 58  
Hosmer–Lemeshow test, 455  
Huygens’ formula, 244  
hybrid clustering methods, 261  
hypergeometric distribution, 659  

illustrative  
individual, 180  
variable, 187  
independence of two variables, 659  
indicator matrix, 201  
inductive techniques, 302
information extraction, 629
information retrieval, 629
input layer, 217
interactions, 82
iterative reclassification, 540

Jarque–Bera test, 56
Jonckheere–Terpstra test, 70

Kaiser criterion, 190
Kaplan–Meier model, 664
Kendall’s tau, 658
k-means, 248
k-medoids, 249
k-modes, 249
k-nearest-neighbour method, 302
Kohonen map, 231
Kolmogorov–Smirnov test, 54
k-prototypes, 249
Kruskal–Wallis test, 69
kurtosis, 649

LARS procedure, 364
lasso, 363
leaps and bounds algorithm, 86
learning process
  consistency, 306
  learning rate, 226
  lemmatization, 631
leptokurtic distribution, 649
Levenberg–Marquardt algorithm, 224
Levene test, 58
lexicometry, 627
lifetime value (LTV), 106
lift, 550
lift curve, 548
likelihood function, 446
Lilliefors test, 54
linear correlation coefficient (Pearson’s r), 656
LOESS regression, 430
logistic regression, 437
  with correlated data, 479–481
  on individuals with different weights, 479
logit model, 440
log-log model, 441
longitudinal data, 479
LOWESS regression, 430
Mahalanobis distance, 340
Mallows $C_p$, 415
Manhattan distance, 237
MANOVA, 664
margin (of a SVM), 502
market basket analysis, 287
Markov chain, 633
MARS algorithm, 330
maximum likelihood method, 446
MAXR selection method, 397
mean square error (MSE), 358
median test, 70
medoid, 249
memory-based reasoning, 302
Mesokurtic distribution, 649
mixed-effects, 435
moment, 226
Monte Carlo simulation methods, 516
MOSAIC, 101
moving centres, 247
multicollinearity, 88
multilayer perceptron, 225
multinomial logistic regression, 482–483
multiple correspondence analysis, 201
multiple imputation, 47
multiple linear regression, 359
multi-type data mining, 636
Nagelkerke’s $R^2$, 455
naive Bayesian classifier, 492–497
neural clustering, 272
neural networks, 217
Newton–Raphson algorithm, 448
nomenclature of territorial units for statistics (NUTS), 99
non-parametric test, 652
normalization, 73
normalized deviance, 456
normalized PCA, 177
normal law, 52
Normit model, 441
nosology, 236
oblique PCA, 192
odds ratio, 443–445
offset variable, 486
one-to-one marketing, 5
open requests, 633
optimal hyperplane, 502
ordered Twoing criterion, 317
ordinal logistic regression, 482
ordinary least squares, 357
orthogonal PCA, 192
output delivery system (ODS), 70
output layer, 217
overdispersion, 457
overfitting, 310
overtraining, 310

PAM algorithm, 249
paragon, 249
parametric test, 652
passenger name record (PNR), 22
percentage of total inertia explained, 190
platykurtic distribution, 649
PLS logistic regression, 483–484
PLS regression, 397
PMML, 15
Poisson regression, 487–491
powerful test, 651
P-P (probability–probability) plot, 55
predefined requests, 633
prediction, 301
principal component, 179
principal component analysis, 176
probabilistic discriminant analysis, 342–345
probabilistic neural networks, 224
probit model, 441
promax PCA, 193
propensity score, 556
pseudo F, 245
pseudo $R^2$, 246
psychographic data, 97
$p$-value, 651

quality of representation, 181
quartimax PCA, 192
QUEST tree, 322

radial basis function network, 227
random effects, 434
random forests, 521
real AdaBoost, 526
recency, frequency, monetary value (RFM) analysis, 93
recovery score, 556
regression, 301
regressor, 366
regularization methods, 363
reject inference, 539–540
relational analysis, 273
relational data, 96
repeated measurements analysis, 480
residual plot, 372
residuals, 357

response curve, 551
ridge regression, 362
risk (behaviour) score, 556
ROC curve, 542–548

safe harbor privacy principles, 22
sample standard deviation, 653
saturated model, 448
Schwarz criterion, 451
scorecard, 565
scorecard, calculation, 599
segmentation, 236
self-organizing map (SOM), 231
semi-partial $R^2$, 246
semisupervised learning, 230
sensitivity, 542
shape characteristics, 648
Shapiro–Wilk test, 55
sigmoid function, 219
significance level, 651
similarity aggregation, 273
simple imputation, 47
simple linear regression, 356
simple random sampling, 90
single linkage, 254
size effect, 184
skewness, 648
smoothing parameter, 259
Somers’ $D$ statistic, 551
Spearman’s rank correlation coefficient, 657
specificity, 542
standardized residuals, 371
statdisc, 80
stepwise selection, 397
stratified sampling, 90
strong forms, 251
structural risk minimization, 309
studentized residuals, 372
student’s test, 653
stump, 521
stylometry, 627
‘subprime’ loans, 567
supervised learning, 217
supplementary
individual, 180
variable, 187
support index, 288
support vector machines, 501–510
support vector regression, 505
survival analysis, 664
systematic sampling, 90
taxonomy
  numerical, 236
  of products, 291
temporal associations, 292
total sum of squares (or inertia), 243
transductive techniques, 302
transition matrix, 40
tree diagram, 253
Twoing criterion, 317
type 1 analysis, 489
type 3 analysis, 490

univariate statistics, 648
unsupervised learning, 217
unsupervised pattern recognition, 236

Vapnik-Chervonenkis dimension, 306
Vapnik’s learning theory, 308
VARCLUS method, 279

variance inflation factor, 88
varimax PCA, 193

Wald statistic, 448
Ward method, 257
web mining, 637
weighted least squares, 430
White test, 373
Wilcoxon–Mann–Whitney test, 66
Wilcoxon’s rank-sum statistic, 67
Wilks’ lambda, 342–349
Winsorization, 52
within-class variance, 662
within-cluster sum of squares, 244
Wong hybrid method, 258

Zipf’s law, 628
Z-score, 615