Index

A
absolute references, Solver, 110
absolute values of errors, median regression, 221
additive smoothing, 86
adjacency matrix, 158
affinity matrix, 159
agglomerative clustering, 185
AI model
Bayes rule and, 83–86
dummy variables, 210–212
feature set, 207–208
versus optimization model, 101–102
overview, 206–207
pregnancy data, 378–385
pregnant customers (See RetailMart (pregnant customers))
training data, oversampling, 210
AIMMS, 118
algorithms, evolutionary, 115–116
alpha value calculation, 276–277
arrays, formulas, 19–20
autocorrelations, 306–313
Bayes rule, 82
AI model creation, 83–86
Big M, 133–137
binary tree, 193–197
BINOMDIST function, 116
blending model, 119
boosting, 251
model evaluation, 280–283
model training, 272–275
weighted errors, 272
reweighting, 277–278

C
CDF (cumulative distribution function), 146–148, 337
mean deviation, 147–148
standard deviation, 147–148
scenarios from, 148–150
cell formatting, 5–7
central limit theorem, 146
chain rule of probability, 81
charts. See also graphs
fan chart, 331–333
inserting, in spreadsheets, 8–9
classifiers, bagging, 254
cluster analysis, 29
cluster centers, solving for, 46–48
cluster labels, 193–197
clustering, 29–30
agglomerative, 185
cluster centroid, 31
community detection, 155–156
divisive, 185–192
hierarchical partitioning, 185
image clustering, 30
k-means, 30–35
  e-mail marketing, 35–66
k groups, 30–35
k-medians, 66–67
  cosine distance, 68–69
  Excel, 69–75
  Manhattan distance, 67–68
network graphs, 155, 156–157
  edges, 156
  nodes, 156
Solver, 34–35
  results, 49
coefficient, variables, 214
coefficient standard error, 226–227
coefficient tests, 226–230
community detection, clustering and,
  modularity maximization, 156
Concessions.xlsx file, 2
conditional formatting, 6–7
conditional probabilities, 80
  Bayes rule, 82
  naïve Bayes model, 94–98
  token counting, 92–93
constraints, 110–112
copying
  data, 4–5
  formulas, 4–5
correlogram, 310–313
cosine distance, k-medians clustering,
  68–69
cosine similarity matrix, 172–174
COUNTIF function, 116
COUNTIFS statement, 235
CPLEX, 118
CRAN (Comprehensive R Archive
  Network), 372–373
critical values, 310–311
cutoff values, 233

data
  copying, 4–5
  merging, VLOOKUP and, 12
Data Laboratory (Gephi), 168–170
data mining, exploratory, 29–30
data sources, k-means clustering, 37–38
data standardization, 40
dataframe, 368–370
decision stumps, 254–257, 260–263
  alpha value calculation, 276–277
  macros, 266
  number of, 257–258
dependent situations, probability theory,
  81–82
dependent variables, 208
design matrix (linear regression), 227
  SSCP, 227–228
distribution
  CDF (cumulative distribution function),
  146–148, 337
  mean deviation, 147–148
  standard deviation, 147–148
  central limit theorem, 146
  Monte Carlo simulation, 149
  probability distribution, 145–146
  standard normal distribution, 343–344
  uniform distribution, 146
divisive clustering, 185–192
DocGraph, 156
document classification, 77
double exponential smoothing,
  299–313
dummy variables, 210–212

E
edges, network graphs, 156, 158
  kNN (k nearest neighbors) graph, 176
  r-neighborhood graphs, 176
ensemble modeling, 251
Ensemble.xlsm, 252
error in calculation column, 217–218
Euclidean distance, 41–44, 345–347
evolutionary algorithms, 115–116
Excel
  constraints, 110–112
  GRG, 218
  k-medians clustering, 69–75
  silhouette, 57–60
  version differences, 1
exploratory data mining, 29–30
exponential smoothing, 288–290
double exponential smoothing,
  299–313
  forecast setup, 290–296
Holt’s Trend-Corrected Exponential
  Smoothing, 299–313
Index 403

Multiplicative Holt-Winters Smoothing, 313–333
trends, 296–299

F
F test, 223–225
factoring, R, 364–367
false positive rate, 236–237
fan chart, 331–333
features, independent variables, 208
filters, 13–16
Find and Replace, 9–10
floating-point underflow, 86
forecasting, 285
autocorrelations, 306–313
correlogram, 310–313
critical values, 310–311
future periods, 303–304
graphing, 296
one-step forecast column, 291–292
error optimization, 293–295
Holt’s Trend-Correct Exponential Smoothing, 304–306
prediction intervals, 285, 327–331
R, 385–389
smoothing
exponential, 288–299
SES (simple exponential smoothing), 288–290
time series data, 286–287
deseasonalizing, 318
seasonality, 314–315
Format Cells menu, 5–6
formatting
cells, 5–7
conditional, 6–7
formulas
arrays, 19–20
copying, 4–5
INDEX, 298
LINEST(), 220
SUMPRODUCT, 19–20
values, locating, 10–11
VLOOKUP, 12
Freeze Panes, 3
Freeze Top Row, 3
functions
BINOMDIST, 116
COUNTIF, 116
HLOOKUP, 116

G
Gephi, 158, 159
Data Laboratory, 168–170
graph layout, 162–164
installation, 160–162
modularity, 197–198
node degrees, 165–166
printing, 166–168
global outliers, 353
graphs. See also charts; network graphs
data preparation, 342–345
forecasting and, 296
kNN (k nearest neighbors), 347–348
modularity
penalties, 179–183
points, 179–183
outlier detection and, 345–347
indegree, 348–351
k-distance, 351–353
LOFs, 353–358
GRG, 218
Gurobi, 118

H
Hadjum versus Hadlum, 336–337
hierarchical partitioning, 185
high-level class probabilities, 84–85
HLOOKUP function, 116
Holt’s Trend-Corrected Exponential Smoothing, 299–313
idiot’s Bayes. See naïve Bayes

IF function, 114, 116
image clustering, 30
indegree (graphs), 166
outlier detection, 348–351
independent variables, 208
INDEX formula, 298
INDEX function, 116
integer programming, switches, 133
intercept of linear model, 214
IQR (Interquartile Range), 337–338

Joey Bag O’Donuts Wholesale Wine Emporium, 36
joint probability, 80–81
chain rule of probability, 81
JuiceLand, 120–121
Solver, 124–126

KDD (knowledge discovery in databases), 30
k-distance, graph outlier detection, 351–353
k-means clustering, 30–35
cluster centers, 46–48
data source, 37–38
distance, 44–46
matrix, 55–56
five clusters, 60–64
four clusters, 41
Joey Bag O’Donuts Wholesale Wine Emporium, 36
k groups, 30–35
PivotTables, 38–39
silhouette, 53–60
5-Means clustering, 64–66
spherical k-means, 372–373
k-medians clustering, 66–67
cosine distance, 68–69
Excel, 69–75
Manhattan distance, 67–68
kNN (k nearest neighbor), 336
outlier detection and, 347–348

L
LARGE function, 116
law of total probability, 80
layout, Gephi graph, 162–164
level sets, 105–106
lexical content, stop words and, 91
LibreOffice, 1
linear programming, 102, 103–104
Excel and, 108–117
fractional solutions, 113
level sets, 105–106
polytopes, 103–105
simplex method, 106–108
linear regression
coefficient, 214
cutoff values, 233
design matrix, 227
SSCP, 227–228
false positive rate, 236–237
intercept, 214
LINEST( ) formula, 220
logistic regression comparison, 245–248
metric trade-offs, 238–239
positive predictive value, 234–235
ROC (Receiver Operating Characteristic) curve, 238–239
simple model, 213–215
statistics, 221
coefficient standard error, 226–227
coefficient tests, 226–230
F test, 223–225
prediction standard error, 226
R-squared, 222–228
t distribution, 230
t test, 226–230
sum of squared error, 215
training the model, 218–220
true negative rate, 235–236
true positive rate/recall/sensitivity, 237
validation set, 231–233
LINEST( ) formula, 220
LINEST function, 297
link function, 240–241
link spam, 166
local outliers, 353
LOF (local outlier factors), 353–358
logistic regression, 239–240
linear regression comparison, 245–248
link function, 240–241
log-likelihood, 244–245
reoptimizing, 241–243
statistical tests, 245
lower inner fence (Tukey fences), 338

M
machine learning, 30
macros, recording, 266
MailChimp.com, 29
Mandrill.com, 77–79
Mandrill.com, 77–79
Mandrill.xlsx, 87
Manhattan distance, 67–68
MATCH function, 116
matrix inversion, 226
matrix multiplication, 226
MAX function, 116
mean deviation, CDF, 147–148
measurement, Euclidean distance, 41–44
MEDIAN function, 116
median regression, 221
merging, VLOOKUP and, 12
MIN function, 116
minimax formulation, 131–132
MINVERSE function, 226
missing values, 253–254
MMULT function, 226
modularity, Gephi, 197–198
modularity maximization, 156
penalties, 179–183
points, 179–183
Monte Carlo simulation, 149
Multiplicative Holt-Winters Smoothing, 313–333

N
naive Bayes, 77
bag of words model, 79
conditional probability tables, 94–98
rare words, 85–86
navigation, Control button, 2–3
network graphs, 155
adjacency matrix, 158
affinity matrix, 159
binary tree, 193–197
cosine similarity matrix, 172–174
DocGraph, 156
edges, 156, 158
kNN (k nearest neighbors) graph, 176
r-neighborhood graphs, 176
Gephi, 158
layout, 162–164
node degrees, 165–166
printing, 166–168
indegree, 166
link spam, 166
nodes, 156, 158
NodeXL, 158
outdegree, 166
outlier detection, 166
r-Neighborhood graph, 174–185
symmetry, 158
undirected, 158
visualizing, 157–158
WineNetwork.xlsx, 170–172
NLP (natural language processing), 87
lexical content, 91
stop words, 91
node impurity, 255–256
nodes, network graphs, 156, 158
NodeXL, 158
non-linear functions, 116
NORMDIST function, 116, 337
null hypothesis, 224

O
OFFSET function, 116
one-step forecast column, 291–292
error optimization, 293–295
Holt's Trend-Correct Exponential Smoothing, 304–306
OpenSolver, 26–27, 118
variables, multiplying, 137–144
optimization, need for, 102–103
Optimization Model tab, 127
optimization models, 20–26, 121–124, 127–128
versus artificial intelligence model, 101–102
OrangeJuiceBlending.xlsx, 118
outdegree (graphs), 166
outlier detection, 166, 335–336
global outliers, 353
graphing, 345–347
data preparation, 342–345
indegree, 348–351
k-distance, 351–353
IQR (Interquartile Range), 337–338
kNN (k nearest neighbor), 347–348
local outliers, 353
LOF (local outlier factors), 353–358
R, 389–394
Tukey fences, 337–338
  limitations, 340–341
  spreadsheets, 338–340
unsupervised machine learning, 336
outliers
  bagging and, 271
  overview, 335
oversampling, 210

P-Q
p( ), 79–80
partitioning, hierarchical, 185
Paste Special, 7–8
PERCENTILE function, 337–338
PivotTables, 16–19
  k-means clustering, 38–39
  PivotTable Builder, 16–17
polytopes, 103–105
  simplex method, 106–108
positive predictive value, 234–235
prediction intervals, 327–331
prediction standard error, 226
Pregnancy Duration.xlsm, 336
pregnancy length, 336–337
printing in Gephi, 166–168
probability distribution, 145–146
probability theory, 79–80
  Bayes rule, 82
  chain rule of probability, 81
  conditional probabilities, 80
  Bayes rule, 82
  token counting, 92–93
dependent situations, 81–82
floating-point underflow, 86
high-level class probabilities, 84–85
independent events, 81
joint probability, 80–81
law of total probability, 80
multiplication rule of probability, 81

R
R (programming language)
  aggregate( ) function, 374
  boxplot( ) function, 390–392
c( ) function, 364–365
cbind( ) function, 368, 375–376
crAN (Comprehensive R Archive Network), 372–373
data input, 363
dataframe, 368–370
data.frame( ) function, 368–369
downloading, 362
factor( ) function, 369–370
factoring, 364–367
forecast( ) function, 387–388
forecasting, 385–389
functions, built-in, 363
glm( ) function, 378
installation, 362
IQR( ) function, 390
Length( ) function, 365
library( ) function, 372–373
lofactor( ) function, 393
matrices, 367–368
matrix function, 367
order( ) function, 376–377
outlier detection, 389–394
packages, 363
performance( ) function, 383
plot( ) function, 384
predict( ) function, 382
print function, 362
randomForest( ) function, 378
rbind( ) function, 368
read.csv( ) function, 374
reading data into, 370–371
row.names( ) function, 374
scale( ) function, 392–393
setwd( ) command, 370
skmeans( ) function, 373
skmeans package, 372
spherical k-means, 372–373
str( ) function, 373–374, 378
summary( ) function, 378
summary function, 370
t function, 367
ts( ) function, 386
varImpPlot( ) function, 381
vector math, 364–367
which() function, 366, 374, 390
working directory, 370–371
write.csv() function, 374
random forest model, 251
replacement and, 271
randomForest package, 271
rare words, naïve Bayes and, 85–86
references, absolute, Solver, 110
regression
linear
coefficient, 214
compared to logistic, 245–248
cutoff values, 233
design matrix, 227–228
false positive rate, 236–237
intercept, 214
LINEST() formula, 220
metric trade-offs, 238–239
positive predictive value,
234–235
ROC (Receiver Operating
Characteristic) curve, 238–239
simple model, 213–215
statistics, 221–230
sum of squared error, 215
training the model, 218–220
true negative rate, 235–236
true positive rate/recall/sensitivity,
237
validation set, 231–233
logistic, 239–240
compared to linear, 245–248
link function, 240–241
log-likelihood, 244–245
reoptimizing, 241–243
statistical tests, 245
median, 221
residual sum of squares, 222
RetailMart (pregnant customers)
data, 215–217
dummy variables, 210–212
error in calculation column,
217–218
feature set, 207–208
folic acid stump, 254–257
linear regression, 213–239
logistic regression, 239–248
training data, 209–210
reweighting weighted errors,
277–278
risk, 144–145
distribution
CDF (cumulative distribution
function), 146–148
central limit theorem, 146
probability distribution, 145–146
r-neighborhood graph, 174–185
ROC (Receiver Operating Characteristic)
curve, 238–239, 252
rows, freezing, 3
R-squared, linear regression, 222–223
S
scenarios, standard deviation, 148–150
constraints, 151–153
school dance analogy for clustering,
31–35
seasonality (forecasting), 314–315
SES (simple exponential smoothing),
288–290
silhouette
5-Means clustering, 64–66
Excel and, 57–60
k-means clustering, 53–60
simplex method, 106–108
smoothing, exponential, 288–290
double exponential smoothing, 299–313
forecast setup, 290–296
Holt’s Trend-Corrected Exponential
Smoothing, 299–313
Multiplicative Holt-Winters Smoothing,
313–333
trends, 296–299
Solver, 20–26
absolute references, 110
clustering, 34–35
results, 49
JuiceLand problem, 124–126
linear regression, training the model,
218–220
OpenSolver, 26–27
sorting, 13–16
spaces, 88–91
spherical k-means, 372–373
spreadsheets
arrays, formulas, 19–20
charts, inserting, 8–9
copying
data, 4–5
formulas, 4–5
filters, 13–16
formatting
cells, 5–7
conditional, 6–7
Freeze Panes, 3
Freeze Top Row, 3
Holt’s Trend-Corrected Exponential Smoothing, 300–306
navigating, Control button, 2–3
Paste Special option, 7–8
PivotTables, 16–19
sorting, 13–16
Tukey fences, 338–340
limitations, 340–341
SSCP (sum of squares and cross products) matrix, 227–228
standard deviation
CDF, 147–148
scenarios from, 148–150
constraints, 151–153
standard normal distribution, 343–344
standardizing data, 40
statistics, 221
coefficient standard error, 226–227
coefficient tests, 226–230
F test, 223–225
logistic regression, 245
matrix inversion, 226
matrix multiplication, 226
prediction standard error, 226
residual sum of squares, 222
R-squared, 222–223
t distribution, 230
t test, 226–230
total sum of squares, 222
stop words, 91
SUBSTITUTE command, 87–88
sum of squared error, 215
SUMIF function, 116
SUMPRODUCT formula, 19–20
SUMPRODUCT function, 109
supervised machine learning, 30
switches, 133
SwordForecasting.xlsm, 286
symmetry in network graphs, 158

t distribution, 230
t test, 226–230
TDIST function, 297
time series data, forecasting and, 286–287
tokens, conditional probability, 92–93
total sum of squares, 222
training data
decision stumps, 260–263
oversampling, 210
random sample, 258–260
trends, forecasting, exponential smoothing, 296–299
triple exponential smoothing, 313–333
true negative rate, 235–236
true positive rate/recall/sensitivity, 237
Tukey fences, 337–338
limitations, 340–341
lower inner fence, 338
spreadsheets, 338–340
upper inner fence, 338

undirected network graphs, 158
uniform distribution, 146
unsupervised machine learning, 30, 336
upper inner fence (Tukey fences), 338

validation set, 231–233
values
locating, with formulas, 10–11
missing, 253–254
variables
coefficient, 214
dependent, 208
dummy variables, 210–212
independent, 208
multiplying, 137–144
vector math, R, 364–367
VLOOKUP formulas, 12
VLOOKUP function, 116
Voronoi function, 32
W–Z

weak learners, 254–255
weighted errors, 272
reweighting, 277–278
WineKMC.xlsx, 36
WineNetwork.xlsx, building graph, 170–172

workbooks
Ensemble.xlsm, 252
Mandrill.xlsx, 87
OrangeJuiceBlending.xlsx, 118
Pregnancy Duration.xlsx, 336
SwordForecasting.xlsx, 286
WineKMC.xlsx, 36