INDEX

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-B testing</td>
<td>292, 298</td>
</tr>
<tr>
<td>accident data</td>
<td>283</td>
</tr>
<tr>
<td>discriminant analysis</td>
<td>283</td>
</tr>
<tr>
<td>naive Bayes</td>
<td>185</td>
</tr>
<tr>
<td>neural nets</td>
<td>261</td>
</tr>
<tr>
<td>activation function</td>
<td>253</td>
</tr>
<tr>
<td>additive seasonality</td>
<td>384</td>
</tr>
<tr>
<td>adjacency list</td>
<td>435</td>
</tr>
<tr>
<td>adjusted-$R^2$</td>
<td>149, 150</td>
</tr>
<tr>
<td>affinity analysis</td>
<td>16, 308</td>
</tr>
<tr>
<td>agglomerative</td>
<td>338, 347</td>
</tr>
<tr>
<td>agglomerative algorithm</td>
<td>347</td>
</tr>
<tr>
<td>aggregation</td>
<td>64, 65, 68, 71</td>
</tr>
<tr>
<td>airfare data</td>
<td>154</td>
</tr>
<tr>
<td>multiple linear regression</td>
<td>154</td>
</tr>
<tr>
<td>algorithm</td>
<td>9, 10</td>
</tr>
<tr>
<td>ALVNN</td>
<td>251</td>
</tr>
<tr>
<td>Amtrak data</td>
<td>367, 377</td>
</tr>
<tr>
<td>time series</td>
<td>367, 377</td>
</tr>
<tr>
<td>visualization</td>
<td>55</td>
</tr>
<tr>
<td>analytics</td>
<td>3</td>
</tr>
<tr>
<td>antecedent</td>
<td>311</td>
</tr>
<tr>
<td>appliance shipments data</td>
<td>375, 404, 423</td>
</tr>
<tr>
<td>time series</td>
<td>405, 425</td>
</tr>
<tr>
<td>autocorrelation</td>
<td>377, 384</td>
</tr>
<tr>
<td>average error</td>
<td>108</td>
</tr>
<tr>
<td>Apriori algorithm</td>
<td>308, 312</td>
</tr>
<tr>
<td>area under the curve</td>
<td>120</td>
</tr>
<tr>
<td>ARIMA models</td>
<td>389</td>
</tr>
<tr>
<td>AR models</td>
<td>390</td>
</tr>
<tr>
<td>artificial intelligence</td>
<td>6, 9, 83</td>
</tr>
<tr>
<td>artificial neural networks</td>
<td>250</td>
</tr>
<tr>
<td>association rules</td>
<td>16, 18, 308, 309</td>
</tr>
<tr>
<td>confidence</td>
<td>312, 317</td>
</tr>
<tr>
<td>confidence intervals</td>
<td>318</td>
</tr>
<tr>
<td>cutoff</td>
<td>316</td>
</tr>
<tr>
<td>data format</td>
<td>314</td>
</tr>
<tr>
<td>itemset</td>
<td>311</td>
</tr>
<tr>
<td>lift ratio</td>
<td>312, 314</td>
</tr>
<tr>
<td>random selection</td>
<td>317</td>
</tr>
<tr>
<td>statistical significance</td>
<td>320</td>
</tr>
<tr>
<td>support</td>
<td>312</td>
</tr>
<tr>
<td>assumptions</td>
<td>241</td>
</tr>
<tr>
<td>asymmetric cost</td>
<td>22, 129, 495</td>
</tr>
<tr>
<td>asymmetric response</td>
<td>495</td>
</tr>
<tr>
<td>attribute</td>
<td>9</td>
</tr>
<tr>
<td>AUC</td>
<td>120</td>
</tr>
<tr>
<td>Australian wine sales data</td>
<td>377, 384</td>
</tr>
<tr>
<td>time series</td>
<td>405, 425</td>
</tr>
<tr>
<td>autocorrelation</td>
<td>377, 384</td>
</tr>
<tr>
<td>average error</td>
<td>108</td>
</tr>
<tr>
<td>average linkage</td>
<td>349, 350, 354</td>
</tr>
<tr>
<td>average squared errors</td>
<td>143</td>
</tr>
<tr>
<td>back propagation</td>
<td>257</td>
</tr>
<tr>
<td>backward elimination</td>
<td>150, 231</td>
</tr>
<tr>
<td>bagging</td>
<td>210, 296</td>
</tr>
<tr>
<td>bag-of-words</td>
<td>451, 454</td>
</tr>
<tr>
<td>balanced portfolios</td>
<td>337</td>
</tr>
<tr>
<td>bankruptcy data</td>
<td>498</td>
</tr>
<tr>
<td>bar chart</td>
<td>55</td>
</tr>
<tr>
<td>bath soap data</td>
<td>491</td>
</tr>
<tr>
<td>batch updating</td>
<td>258</td>
</tr>
<tr>
<td>benchmark</td>
<td>108, 125</td>
</tr>
<tr>
<td>benchmark confidence value</td>
<td>314</td>
</tr>
<tr>
<td>best pruned tree</td>
<td>204</td>
</tr>
<tr>
<td>best subsets</td>
<td>37</td>
</tr>
<tr>
<td>betweenness in network</td>
<td>438</td>
</tr>
<tr>
<td>bias</td>
<td>27, 148, 149, 282</td>
</tr>
<tr>
<td>bias–variance trade-off</td>
<td>148</td>
</tr>
<tr>
<td>big data</td>
<td>7</td>
</tr>
<tr>
<td>binning</td>
<td>68, 70</td>
</tr>
<tr>
<td>binomial distribution</td>
<td>241</td>
</tr>
<tr>
<td>blackbox</td>
<td>250, 268, 297, 365, 456</td>
</tr>
<tr>
<td>boosted tree</td>
<td>186</td>
</tr>
</tbody>
</table>

boosting, 210, 296
bootstrap, 296
Boston housing, 84
Boston housing data, 84
multiple linear regression, 153
visualization, 54
boxplot, 52, 57, 63, 67, 108, 144, 501
side-by-side, 59
Bureau of Transportation statistics, 231
business analytics, 3, 15
business intelligence, 3

Cₚ, 149
C4.5, 186, 201
cab cancellations, 489
Canadian manufacturing workhours data, time series, 375, 397
CART, 186, 195, 201
case, 9
case deletion, 148
case updating, 258
catalog cross-selling case, 497
categorical variable, 188
centrality in a network, 437
centroid, 275, 277, 354
cereals data, 91
exploratory data analysis, 102
hierarchical clustering, 361
CHAID, 199
Charles Book Club case, 320, 468
CHIDIST, 242
chi-square distribution, 242
chi-square test, 199
city-block distance, 343
classification, 16, 55, 62, 68, 78, 107, 157, 169
discriminant analysis, 273
logistic regression, 218
classification functions, 277, 284
classification matrix, 113, 133, 196, 236, 242
classification methods, 229
classification performance, accuracy measures, 115
classification matrix, 113
training data, 114
validation data, 114
classification and regression trees, 100, 263, 268
classification rules, 187, 206
classification scores, 279, 282
classification trees, 186, 481
performance, 196
classifier, 107, 281, 501
cleaning the data, 23
clique, 437
closeness in a network, 438
cluster analysis, 55, 158, 336
allocate the observations, 355
average distance, 345
centroid distance, 345
initial partition, 354
labeling, 352
maximum distance, 345
minimum distance, 345
nesting, 350
normalizing, 340
outliers, 354
partition, 352, 355
randomly generated starting partitions, 356
stability, 354
sum of distances, 356
summary statistics, 352
unequal weighting, 341
validating clusters, 352
cluster centroids, 357
cluster dispersion, 354
clustering, 17, 18, 98, 326, 328
clustering algorithms, 347
clustering techniques, 25
cluster validity, 357
coefficients, 256
collaborative filtering, 17, 322
collaborative filtering using network data, 444
collinearity, 225
combining categories, 100
complete linkage, 348
conditional probability, 9, 10, 169, 174, 313
confidence, 9, 312
confidence interval, 9, 312
confidence levels, 312
confusion matrix, 113, 481
connected network, 437
consequent, 311
consistent, 225
consumer choice theory, logistic regression, 219
continuous response, 109
corpus, 452
correlation, 82, 277, 325, 341
correlation analysis, 88
correlation-based similarity, 343
correlation matrix, 88
correlation table, 60
correspondence analysis, 90
cosine similarity, 325
cosmetics data, affinity analysis, 333
association rules, 332
cost complexity, 202
cost/gain matrix, 481
costs, 22
course topics data, association rules, 332
covariance matrix, 277, 343
credit card data, neural nets, 271
credit risk score, 477
curse of dimensionality, 83, 166
customer segmentation, 337
classification tree, 195
deck, 115, 470, 481
cutoff, 229, 255
dashboards, 3
database marketing, 469
data driven method, 250, 366, 407
data exploration, 82
data mining, 5
data partitioning, 196, 364
data pipeline, 42
data projection, 92
data reduction, 17
data science, 8
data visualization, 18
decile chart, 127, 229
decile lift chart, 111, 127
decision making, 220, 269
deep learning, 270
deep learning networks, 270
degree distribution, 439
degree in a network, 437
delayed flight data, classification tree, 214
logistic regression, 231
naive Bayes, 169
deliminators for text, 453
dendrogram, 336, 350
density of a network, 439
department store sales data, time series, 375, 401, 422
dependent variable, 10
de-seasonalizing, 411
de-trending, 411
deviance, 242
deviation, 108
dimension reduction, 17, 18, 82, 83, 186, 218, 328, 455
dimensionality, 82
directed vs. undirected network, 431
discriminant analysis, 6, 158, 209, 273
assumptions, 280
classification performance, 280
classification score, 277
confusion matrix, 281
cutoff, 279
distance, 275
expected cost of misclassification, 283
lift chart, 282
lift curves, 279
more than two classes, 283
multivariate normal distribution, 280
outliers, 280
prior/future probability of membership, 282
probabilities of class membership, 282
propensities, 282
unequal misclassification costs, 283
validation set, 282
discriminant functions, 206
discriminators, 277
disjoint, 311
distance, 340, 354
distance between clusters, 345
distance between observations, 345
distance between records, 158
distance matrix, 346, 347, 354
distribution plots, 52, 57
divisive, 338
domain-dependent, 341
domain knowledge, 24, 25, 83, 148, 347, 355
double exponential smoothing, 417
dummy variables, 23, 89, 90, 144, 234
East-West Airlines data, cluster analysis, 361
neural nets, 271
eBay auctions data, classification tree, 214
logistic regression, 248
edge in network, 431
edge weight, 437
efficient, 225
eigocentric network, 438
eigenvector centrality, 438
ensemble forecast, 366
ensembles, 210, 292
entity resolution, 442
entropy measure, 191, 208
impurity, 191
epoch, 258
equal costs, 22
error, average, 37, 108, 144
back propagation, 257
mean absolute, 108
mean absolute percentage, 108
overall error rate, 115
prediction, 108
RMS, 37
root mean squared, 37, 108
sum of squared, 37
total sum of squared, 108
error rate, 201
estimation, 10
Euclidean distance, 158, 275, 325, 340, 350, 354
evaluating performance, 207, 208
Excel template, 74
exhaustive search, 149, 150
expert knowledge, 83
explained variability, 149
exploratory modeling, 142
exploratory analysis, 280
explore, 20, 21
exponential smoothing, 407, 414
exponential trend, 67, 379
extrapolation, 268
factor analysis, 166
factor selection, 83
false negative rate, 120
false positive rate, 120
farm ads, 465
feature, 10
feature extraction, 83
field, 10
filtering, 53, 64, 67
finance, 337
financial applications, 273
Financial Condition of Banks data, logistic regression, 247
first principal component, 92
Fisher’s linear classification functions, 277
fitting the best model to the data, 142
forecasting, 55, 56, 71, 364
forward selection, 150, 231
fraudulent financial reporting data, 170
naive rule, 112
fraudulent transactions, 22
frequent itemset, 311
function, nonlinear, 221
fundraising data, 494, 495, 497
generalization, 27
German credit case, 476
Gini index, 191, 208
global pattern, 65-67, 78, 79, 371
goodness of fit, 107, 112
Gower, 344
graphical exploration, 53
heatmap, 60
hidden, 261
hidden layer, 250
hierarchical, 338
hierarchical clustering, 336, 347
hierarchical methods, 350
hierarchies, 64, 65
histogram, 52, 57, 70, 108, 144, 240
holdout data, 10, 29, 142
Holt–Winters exponential smoothing, 417
Home value data, 21
homoskedasticity, 143
impurity, 199, 207
impurity measures, 191, 207
imputation, 148
independent variable, 10
industry analysis, 337
input variable, 10
integer programming, 354
interactions, 251
interaction term, 231
interactive visualization, 52, 69
JMP, 72
software, 72
Spotfire, 72
Tableau, 72
iterative search, 150
Jaquard’s coefficient, 344
jittering, 68
kitchen-sink approach, 146
k-means clustering, 325, 336, 340, 354, 493
k-nearest neighbors, 6, 157, 325, 471, 473
algorithm, 157
labels, 67
labeling documents, 456
laptop sales data, visualization, 80
large dataset, 6, 67
latent semantic indexing, 455
lazy learning, 166
learning rate, 257, 257, 256, 257, 286
level, 364, 367
lift, 236, 485
lift chart, 106, 109, 124, 127, 134, 242, 497
lift (gains) chart, 229
lift curve, 236
lift ratio, 312
line graph, 55
linear, 255
linear classification rule, 274
linear combination, 90, 92
linear regression, 6, 18, 26, 32, 100, 157, 240, 366
linear regression models, 251
linear relationship, 88, 140, 143, 148
linked plots, 71
link prediction, 441
links in a network, 431
local optimum, 269
local pattern, 67, 78, 79, 371
log, 144
logistic regression,
classification, 224
classification performance, 229
confidence intervals, 225
dummy variable, 228
iterations, 225
maximum likelihood, 225
model interpretation, 234
model performance, 236
negative coefficients, 226
nominal classes, 245
ordinal classes, 244
parameter estimates, 225
positive coefficients, 226
prediction, 224
profiling, 241
p-value, 242
training data, 242
variable selection, 231, 237
logistic regression, 6, 100, 218, 251, 255, 273, 280, 471, 476, 481, 484
logistic response function, 221
logit, 220, 221, 229
log-scale, 274
log transform, 256
MA, 408
machine learning, 5, 6, 9, 23
MAE/MAD, 108
Mahalanobis distance, 277, 343
majority class, 160
majority decision rule, 158
majority vote, 163
Manhattan distance, 341, 343
MAPE, 106, 108
market basket analysis, 308
marketing, 196, 199, 471
market segmentation, 337
market structure analysis, 337
matching, 440
matching coefficient, 344
maximum coordinate
distance, 343
maximum likelihood, 225, 256, 257
mean absolute error, 108
mean absolute percentage
error, 108
measuring impurity, 208
minimum error tree, 202, 204
minimum validation error, 263
minority class, 269
misclassification,
asymmetric costs, 121
average misclassification
cost, 123
estimated misclassification
rate, 115
misclassification costs, 117
misclassification error, 112, 158
misclassification rate, 22, 114
missing data, 20, 26
missing values, 26, 96, 148, 186
model, 10, 22, 24, 30
logistic regression, 220
model complexity, 231
model performance, 132
model validity, 142
momentum, 267, 269
moving average, 407, 408
centered moving average, 408
trailing moving average, 408, 409
multilevel forecaster, 366
multicollinearity, 88, 96, 144, 148
multilayer feedforward
networks, 251
multiple linear regression,
36, 140, 220, 231, 242, 286, 484
multiple linear regression
model, 144
multiple panels, 53
multiple R², 242
multiplicative factor, 227
multiplicative model, 227
multiplicative seasonalitiy, 384
naive Bayes, 169
naive classification, 122
naive forecast, 372
naive model, 241
naive rule, 112, 160
natural gas sales data, time series, 424
natural hierarchy, 338
nearest neighbor, 158
almost, 165
recommendation system, 322
recommender systems, 309
record, 10, 25
recursive partitioning, 186, 188
redundancy, 92
reference category, 234
reference line, 125
regression, 140
time series, 377
regression trees, 152, 186, 207, 484
rescaling, 64, 65, 256
residuals, histogram, 241
residual series, 388
response, 10
response rate, 22
reweight, 131
RFM segmentation, 471
riding-mower data, CART, 188
discriminant analysis, 274
k-nearest neighbor, 158
logistic regression, 248
visualization, 80
right-skewed distribution, 144, 256
RMSE, 106, 108
robust, 106, 225, 347, 354
robust distances, 341
robust to outliers, 209
ROC curve, 120
root-mean-squared error, 108
row, 10
rules, association rules, 309
sample, 6, 9, 10, 21
sampling, 21, 40, 67
satellite radio customer data, association rules, 332
scale, 27, 97, 263, 340
scatter plot, 55, 62, 67, 98
animated, 64
color-coded, 62
scatter plot matrix, 63
score, 10, 21, 40, 470
seasonal variable, 384
seasonality, 364, 367
second principal component, 92
segmentation, 337
segmenting consumers of bath soap case, 491
self-proximity, 340
SEMMA, 21
sensitivity, 119
sensitivity analysis, 268
separating hyper-plane, 277
separating line, 277
Sept 11 travel data, time series, 374, 396, 420
shampoo sales data, time series, 376, 424
similarity measures, 343
simple linear regression, 223
color-coded, 62
single linkage, 348, 350
singleton, 437
singular value decomposition, 166, 328
smoothing, 160, 366
time series, 407
smoothing constants, 407
smoothing parameter, 414, 415
souvenir sales data, time series, 376, 403
spam e-mail data, discriminant analysis, 290
spelling errors, 256
terminal node, 201
test data, 20
test partition, 29
test set, 10, 11, 231
text mining, 450
TF-IDF, 455
time series, dummies, 383
lagged series, 387
residuals, 378
RMS error, 379
window width, 411, 415
time series forecasting, 78, 364
time series partitioning, 371
tokenization, 453
total SSE, 108
total sum of squared errors, 108
total variability, 92
Toyota Corolla data, 49, 143
best subsets, 150
classification tree, 207, 215
forward selection, 151
multiple linear regression, 156
neural nets, 271
principal components analysis, 103
Toys R Us revenues data, data reduction, 90
time series, 398
training, 257
training data, 18, 20
sum of squared deviations, 143, 208
sum of squared errors, 242
sum of squared perpendicular distances, 92
supervised learning, 11, 18, 52, 55, 63, 106
system administrators data, discriminant analysis, 290
logistic regression, 247
target variable, 10
Tayko data, 482
multiple linear regression, 154
Tayko software catalog case, 482
Term Frequency-Inverse Document Frequency, 455
TP, 16
standard error of estimate, 107
standardization, 96
standardize, 27, 96, 158
statistical distance, 277, 343, 354
statistics, 6
stemming, 454
steps in data mining, 19
stepwise, 151
stepwise regression, 150, 231
stop list, 453
stopping tree growth, 199
stopword, 453
stratified sampling, 129
subsets, 160
subset selection, 100, 150
subset selection in linear regression, 146
success class, 11
summary statistics, 148
training partition, 28
training set, 11, 106, 114, 142, 364
transfer function, 253
transformation of variables, 209
transformations, 89, 251
transform, 256
transpose, 277
tree depth, 199
trees, 28
search, 165
trend, 65, 67, 364, 367, 379
trend lines, 67, 79, 368
trial, 258
triangle inequality, 340
unbiased, 143, 149
undersampling, 129
unequal importance of classes, 119
Universal Bank data, 167
classification tree, 196, 204
discriminant analysis, 274, 289
logistic regression, 225
university rankings data, cluster analysis, 360
principal components analysis, 102
unsupervised learning, 11, 18, 52, 55, 63, 69, 79, 328
UPGMA, 349
uplift modeling, 292, 297
validation data, 20, 201
validation partition, 28
validation set, 10, 11, 106, 114, 131, 140, 229, 364
variability, between-class, 277
within-class variability, 277
variable, 11
binary dependent, 219
categorical, 23
continuous, 23
nominal, 23
numerical, 23
ordinal, 23
selection, 24, 146
text, 23
variation, between-cluster, 352
within-cluster, 352
variable selection, 140, 231
vertex in a network, 431
visualization, animation, 64
color, 62
hue, 62
map chart, 76
multiple panels, 62, 63
networks, 73
shape, 62, 63
size, 62, 63
treemaps, 75
Walmart stock data, time series, 400
Ward’s method, 349
weighted average, 163
weighted sampling, 495
West Roxbury housing data, 32
wine data, principal components analysis, 103
within-cluster dispersion, 357
zooming, 53, 64, 66, 71
z-score, 27, 276, 277, 340