Adaptive boosting, classification using, 189–191
Adjacency matrix, network data mining, 272–273
  friendship connections example, 283–292
Agglomerative algorithms, hierarchical clustering procedures, 212–217
  European protein consumption example, 217–218
Aggravation scale, logistic regression, death penalty data, 89–92
Aggregation, classification using, 188–191
  Alumni donations, data preprocessing case study, 7, 17–31
Analysis of variance analysis (ANOVA) deviance, decision tree splitting, 164–167
APGAR scores, birth data preprocessing case study, 13–17
  arules software package, 223–227
Assignment step, k-means clustering, 197
Association rules, market basket analysis, 221–234
Automatic data creation, data mining and, 4–5
Automobile fuel efficiency example, linear regression, 43–47
Bagging aggregation, classification using, 189–191
Bandwidth parameters, local polynomial regression, 56–65
Bar plots: alumni donations case study, 24–31
  birth data preprocessing case study, 14–17
Bayesian analysis:
  categorical predictor variables, 126–130
  cluster analysis, expectation-maximization algorithm, 206
  multinomial logistic regression, forensic glass case study, 141–147
  bayesm software:
    cluster analysis, expectation-maximization algorithm, 207
    data preprocessing case study, 7
    orange juice sales data, 31–39
    penalty-based variables, LASSO algorithm, orange juice sales data case study, 78–82
Benjamini-Hochberg-Yekutieli procedure, statistical modeling parsimony, false discovery and, 70
Bernoulli model, logistic regression, 85–86
Betweenness, network data mining, 275–278
Bigrams, text data mining, 259
Binary classification, decision problems, 108
Binary response variable, logistic regression, 83–85
  death penalty data, 88–92
Birth data, preprocessing case study using, 7–17
Boosting techniques, classification using, 189–191
Bootstrap aggregation, classification using, 189–191
Box plots:
- data preprocessing:
  - alumni donations case study, 20–31
  - birth data case study, 13–17
  - orange juice sales case study, 31–39
  - multinomial logistic regression, forensic glass case study, 134–147
- nearest neighbor analysis, forensic glass case study, 118–122
- penalty-based variables, LASSO algorithm, orange juice sales data case study, 81–82

Burt’s structural holes, network data mining, 278

Business analytics, data mining and, 5–6

Categorical outcome data, data mining and, 2

Categorical variables:
- alumni donations data preprocessing case study, 29–31
- decision trees, chi-square automatic interaction detection, 187–188
- hierarchical clustering procedures, 213–217
- naive Bayesian analysis, 126–130

Chi-square automatic interaction detection (CHAID), 186–188

Classification:
- discriminant analysis, 150–160
- Fisher’s linear discriminant function, 153–160
- nearest neighbor analysis, 115–125
- neural networks for, 192–193
- performance evaluation, 108–114
- support vector machines, 192

Classification and regression trees (CART) algorithm, decision tree predictions:
- basic principles, 161–167
- chi-square automatic interaction detection, 186–188
- prostate cancer data example, 170–179

Classification deviance, decision tree splitting, 165–167

Classification trees:
- chi-square automatic interaction detection, 186–188
- iris data example, 182–184
- software packages for, 185–195

Cluster analysis:
- defined, 2–3
- European protein consumption example, 197–201, 207–212, 217–218
- expectation-maximization algorithm, normal distribution mixtures, 204–207
- E-step, 206
- M-step, 206–207
- hierarchical procedures, 212–217
- k-means clustering, 196–197
- monthly U.S. unemployment rates example, 201–204, 218–219
- Complete-linkage clustering, 213–217
- Conditional probability, market basket analysis, 221
- Confidence parameters, market basket analysis, 221–222
- Continuous variables, decision tree splitting, 164–167
- Corpus, text data mining, 258–259
- Cost complexity, decision tree predictions, prostate cancer data, 171–179
- contourplot, birth data preprocessing case study, 15–17
- Covariance matrices, cluster analysis, expectation-maximization algorithm, European protein consumption example, 208–212

Cross-validation:
- decision tree predictions:
  - deviance node impurity, 185–186
  - prostate cancer data, 170–179
- k-nearest neighbor algorithm, 117–125
- linear discriminant function analysis, forensic glass case study, 159
- local polynomial regression, 56–65
- logistic regression, delayed airplanes data, 93–100
- penalty-based variable selection, LASSO algorithm, 73–82
- prostate cancer data, 77–78
- standard linear regression, automobile fuel efficiency example, 46–47

Cross-validation cost (CV cost), decision tree partitioning, 177–179

Cutoff:
- probability-based decisions, 108–109
- sensitivity and specificity, 109
Data mining:
- network data, 272–291
- process of, 3–4
- size parameters, 1–2
- text data, 258–271
Data sets:
- current applications, 1–2
- data mining from, 1–2
Data warehousing, data mining and, 5
Death penalty data, logistic regression, 87–92
Decision trees, 161–184
- chi-square automatic interaction detection, 186–188
- ensemble construction methods, 188–191
- motorcycle acceleration example, 179–182
- prostate cancer data case study, 167–179
Delayed airplanes data:
- logistic regression, 92–100
- naive Bayesian analysis, 127–130
Dendrogram:
- European protein consumption example, 217–218
- hierarchical clustering procedures, 212–217
Density estimation:
- local polynomial regression, 58–65
- nitric oxide emissions, 62–65
- Old Faithful example, 59–61
- network data mining, friendship connections example, 285–292
Deviance estimation, logistic regression, 86
Deviance node impurity, decision trees:
- characteristics of, 162–167
- prostate cancer data, 171–179
- rpart and tree software systems, 185–186
Dimension reduction:
- European protein consumption example, 239–243
- factor models and components, 235–246
- monthly U.S. unemployment rates example, 243–246
- multicollinear regressions, 247–256
Discriminant analysis:
- classification and, 150–160
- German credit data case study, 154–156
Distance measure, hierarchical clustering procedures, 212–217
Divide-and-conquer partitioning, in decision trees, 163–167
Divisive algorithms, hierarchical clustering procedures, 212–217
Document comparison, text data mining, 259–271
Edge list, network data mining, 272–291
- friendship connections example, 283–292
Elasticity parameters, orange juice sales case study, data preprocessing and, 37–38
Ensemble classifiers, classification using, 188–191
Entropy criterion, decision tree analysis, 166–167
Epanechnikov weight function, local polynomial regression, 56–65
Error sum of squares, standard linear regression, 41–42
Euclidean distance:
- hierarchical clustering procedures, 213–217
- European protein consumption example, 217–218
- k-means clustering, 196–197
- k-nearest neighbor algorithm, 116–125
- European protein consumption example: cluster analysis, 197–201, 207–212, 217–218
- factor models, dimension reduction, 238–243
Evaluation (testing), logistic regression, delayed airplanes data, 93–100
Evaluation data set, graphical user interface, 195
Expectation-maximization (EM) algorithm, cluster analysis, normal distribution mixtures, 204–207
E-step, 206
M-step, 206–207
Expected cost of misclassification, discriminant analysis, 150–160
Explanatory variables, logistic regression, delayed airplanes data, 92–100
Factor models:
- dimension reduction, 235–246
Factor models (Continued)
European protein consumption example, 238–243
monthly U.S. unemployment rates example, 243–246
False discovery rate (FDR), statistical modeling, parsimony in, 67–70
False negative rate, probability-based decisions, 108–109
False positive rate, probability-based decisions, 108–109
Feature advertisement effects, orange juice sales case study, 34–37
Features, classification, nearest neighbor analysis, 115–125
Fisher’s linear discriminant function, 153–160
iris data, 156–157
decision tree analysis, 182–184
Fitted value:
  multinomial logistic regression, forensic glass case study, 137–147
  standard linear regression, 40
Folds, decision tree partitioning, 177–179
Forecasting:
  aggregated techniques for, 188–191
  multicollinear regressions, partial least squares, monthly U.S. unemployment rates example, 254–256
Forensic glass case study:
  linear discriminant function analysis, 157–159
  multinomial logistic regression, 134–147
  nearest neighbor analysis, 117–122
Forest-RI, classification using, 190–191
Frequency distribution table, birth data preprocessing case study, 10–17
Friendship connections, network data mining, 278–292
Generalized cross-validation (GCV), local polynomial regression, 56–65
Generalized linear models (GLM), logistic regression, 87
Generated data example, multicollinear regression, partial least squares, 249–251
Gentzkow/Shapiro estimation, text data mining, 268–271
Gephi software, 272
German credit data:
  cutoff and receiver operating characteristic function, 109–114
discriminant function analysis, 154–156
logistic regression, 103–107
nearest neighbor analysis, 122–125
Gini index, decision tree analysis, 167, 186
Graphical user interface, R package rattle, 193–195
Graphics techniques:
  network data mining, 274–278
  friendship connections example, 284–292
  preprocessing case studies, birth data case study, 10–17
Hamming distance, k-nearest neighbor algorithm, 117–125
Hard counts, cluster analysis, expectation-maximization algorithm, 204–207
Hat matrices, local polynomial regression, 57–65
Hidden layers, neural networks, 193
Hierarchical clustering procedures, 212–217
Histograms:
  data preprocessing:
    alumni donations case study, 18–31
    birth data case study, 10–17
    orange juice sales case study, 31–39
  local polynomial regression:
    nitric oxide emissions, 62–65
    Old Faithful example, 59–61
    smoothing, 58
igraph software, 272, 274–278
Income prediction example, market basket analysis, 227–234
Indicator variables, alumni donations data preprocessing case study, 25–31
Information criterion, decision tree analysis, 166–167
Initialization, k-means clustering, 197
In-sample performance, decision tree analysis, prostate cancer data, 176–179
Interaction effects, multinomial logistic regression, forensic glass case study, 146–147
Inverse multinomial logistic regression, text data mining, 259–260
Inverse prediction, text data mining, 260
Iris data:
  decision tree analysis, 182–184
  Fisher’s linear discriminant function, 156–157
K-direction partial least squares (PLS(K)), multicollinear regressions, 248–256
k-means clustering:
  basic principles, 196–197
  European protein consumption example, 207–212
  expectation-maximization algorithm, 204–207
  monthly U.S. unemployment rates example, 202–204, 243–246
k-nearest neighbor algorithm:
  classification applications, 116–125
  forensic glass case study, 117–122
  k-means clustering, 196–197
Lagrangian multiplier, least absolute shrinkage and selection operator algorithm, 72–82
Laplace transform, multinomial logistic regression, forensic glass case study, 141–147
Lattice (trellis) graphics, data preprocessing:
  birth data case study, 10–17
  orange juice sales data, 31–39
Leaf nodes:
  decision trees, 162
  prostate cancer data example, 170–179
Least absolute shrinkage and selection operator (LASSO) algorithm:
  multicollinear regressions, partial least squares, monthly U.S. unemployment rates example, 254–256
  multinomial logistic regression, 134
  penalty-based variable selection, multiparameter regression models, 71–82
prostate cancer example:
  decision trees, 167–179
  multiparameter regression models, 74–78
Least angle regression (LARS) algorithm, penalty-based variable selection, 72–82
Least squares estimates. See also Partial least squares (PLS)
  local polynomial regression, 56–65
  penalty-based variable selection, LASSO algorithm, 71–82
  standard linear regression, 40
levelplot, birth data preprocessing case study, 15–17
Leverage rules, market basket analysis, 222
Lift charts, logistic regression, delayed airplanes data, 97–100
Lift rules, market basket analysis, 221
Likelihood function, cluster analysis, expectation-maximization algorithm, 206–207
Linear discriminant function:
  classification, discriminant analysis, 152–160
  Fisher’s linear discriminant function, 153–160
  forensic glass data case study, 157–159
  MBA admission data, 159–160
Linear regression. See also Logistic regression
  automobile fuel efficiency example, 43–47
  decision trees, 161
  motorcycle acceleration example, 179–182
  overfitting effects, 53–54
  penalty-based variables, 71–82
  principal components analysis, multicollinear regressions, 247–256
  standard model, 40–54
  text data mining, Gentzkow/Shapiro estimation, 268–271
  Toyota used-car prices example, 47–51
Linkage criterion, hierarchical clustering procedures, 213–217
Loadings, factor models, dimension reduction, 236–246
Loan acceptance case study, logistic regression, 100–102
Local polynomial regression:
density estimation and histogram smoothing, 58
density estimation and histogram smoothing, 58
examples and software, 58–65
examples and software, 58–65
model selection, 56–57
model selection, 56–57
multiple regression model, 58
multiple regression model, 58
nitric oxide emissions, 62–65
nitric oxide emissions, 62–65
Old Faithful example, 59–61
Old Faithful example, 59–61
overview, 55–56
overview, 55–56
locfit R library, local polynomial regression, 58–65
locfit R library, local polynomial regression, 58–65
Logistic regression. See also Linear regression
binary response data, 83–85
decision trees, 161
delayed airplanes data, 92–100
German credit data, 103–107
loan acceptance analysis, 100–102
loan acceptance analysis, 100–102
multinomial techniques, 132–149
multinomial techniques, 132–149
computer software, 134
computer software, 134
forensic glass case study, 134–147
forensic glass case study, 134–147
simple triplet matrix specification, 147–149
simple triplet matrix specification, 147–149
new case classification, 86–87
new case classification, 86–87
overview, 83
overview, 83
regression coefficients, 85
regression coefficients, 85
R estimation, 87
R estimation, 87
statistical inference, 85–86
statistical inference, 85–86
Log-likelihood function, cluster analysis, expectation-maximization algorithm, 206
Log-likelihood function, cluster analysis, expectation-maximization algorithm, 206
Majority voting classification, k-nearest neighbor algorithm, 117–125
Majority voting classification, k-nearest neighbor algorithm, 117–125
Mallows’ Cp statistic:
local polynomial regression, 56–65
local polynomial regression, 56–65
standard linear regression, 41–42
standard linear regression, 41–42
Margin of separation, support vector machine classification, 192
Marginal of separation, support vector machine classification, 192
Market basket analysis, 220–234
Market basket analysis, 220–234
income prediction, 227–234
income prediction, 227–234
online radio example, 221–227
online radio example, 221–227
Markov Chain Monte Carlo (MCMC) methods, cluster analysis, expectation-maximization algorithm, 207
Markov Chain Monte Carlo (MCMC) methods, cluster analysis, expectation-maximization algorithm, 207
Marriage and power in fifteenth-century Florence, network data mining, 274–278
Marriage and power in fifteenth-century Florence, network data mining, 274–278
Maximum likelihood estimation:
cluster analysis, expectation-maximization algorithm, 207
cluster analysis, expectation-maximization algorithm, 207
logistic regression, 85–86
logistic regression, 85–86
death penalty data, 89–92
death penalty data, 89–92
multinomial logistic regression, 132–134
multinomial logistic regression, 132–134
standard linear regression, 40
standard linear regression, 40
Maximum pairwise distance, hierarchical clustering procedures, 213–217
Maximum pairwise distance, hierarchical clustering procedures, 213–217
MBA admission data, linear discriminant function, 159–160
MBA admission data, linear discriminant function, 159–160
Mean absolute percent error, standard linear regression, 42–43
Mean absolute percent error, standard linear regression, 42–43
automobile fuel efficiency example, 46–47
automobile fuel efficiency example, 46–47
Toyota used-car prices example, 50–51
Toyota used-car prices example, 50–51
Mean error, standard linear regression, 42–43
Mean error, standard linear regression, 42–43
automobile fuel efficiency example, 46–47
automobile fuel efficiency example, 46–47
Toyota used-car prices example, 50–51
Toyota used-car prices example, 50–51
Mean square error, linear regression, model overfitting effects, 53–54
Mean square error, linear regression, model overfitting effects, 53–54
Minimization, k-means clustering, 196–197
Minimization, k-means clustering, 196–197
Minimum pairwise distance, hierarchical clustering procedures, 213–217
Minimum pairwise distance, hierarchical clustering procedures, 213–217
Minitab program:
graphical user interface, 193–194
graphical user interface, 193–194
logistic regression, death penalty data, 88–92
logistic regression, death penalty data, 88–92
Misclassification proportion, naive Bayesian analysis, delayed airplanes data, 130
Misclassification proportion, naive Bayesian analysis, delayed airplanes data, 130
Mixed probability, cluster analysis, expectation-maximization algorithm, 204–207
Mixed probability, cluster analysis, expectation-maximization algorithm, 204–207
mixOmics software package, 249
mixOmics software package, 249
mixtools software package, 207
mixtools software package, 207
Model fitting:
dimension reduction, 235–246
dimension reduction, 235–246
logistic regression, delayed airplanes data, 93–100
logistic regression, delayed airplanes data, 93–100
Modeling problems, data preprocessing:
alumni donations case study, 31
birth data case study, 17
orange juice sales case study, 38–39
orange juice sales case study, 38–39
INDEX 347

Monthly U.S. unemployment rates example:
  cluster analysis, 201–204, 218–219
  factor models, dimension reduction,
    243–246
  multicollinear regressions, partial least
    squares:
      out-of-sample predictions, 253–256
      predictions on past performance,
        251–253
mosaic plots, alumni donations data
preprocessing case study, 25–31
Motorcycle acceleration, decision tree
analysis, 179–182
Multicollinearity:
  penalty-based variable selection, LASSO
    algorithm, 73–82
  principal components regression,
    247–256
  generated data example, 249–251
  monthly U.S. unemployment example,
    vector autoregression model,
      251–253
  standard linear regression, 41–42
Multinomial logistic regression,
  132–149
  computer software, 134
  decision trees, 161
  forensic glass case study, 134–147
  inverse, text data mining, 259–260
  simple triplet matrix specification,
    147–149
Multiparameter regression models:
  multinomial logistic regression, forensic
    glass case study, 141–147
  penalty-based variable selection, 71–82
Multiple regression model, local
  polynomial regression and, 58
Multiplicity problem, statistical modeling
  parsimony, false discovery and, 67–70
Multivariate data sets:
  cluster analysis,
    expectation-maximization algorithm,
      205–207
  decision trees, chi-square automatic
    interaction detection, 186–188
  dimension reduction, 236–246
  multicollinear regressions, partial least
    squares, monthly U.S. unemployment
    rates example, 254–256
Naive Bayesian analysis, categorical
  predictor variables, 126–130
Nearest neighbor analysis:
  classification applications, 115–125
  forensic glass case study, 117–122
  German credit data, 122–125
Network data mining, 272–291
  friendship connections example,
    278–292
  marriage and power in fifteenth-century
    Florence, 274–278
Neural networks, classification using,
  192–193
Nitric oxide emissions, local polynomial
  regression, 62–65
Nodes and node impurity:
  decision trees:
    characteristics of, 162–167
    R software systems, 185–186
Nodes in network data mining, 272–291
  marriage and power in fifteenth-century
    Florence example, 275–278
Nonparametric regression:
  decision tree predictions, prostate cancer
    data example, 170–179
  local polynomial regression, 55–65
Normal distribution:
  classification, discriminant analysis,
    151–160
  cluster analysis,
    expectation-maximization algorithm,
      204–207
  European protein consumption
    example, 208–212
NP hard problem, 196–197
n × n distance matrix, hierarchical
  clustering, 213–217
  monthly U.S. unemployment rate
    example, 218–219
Odds ratio, logistic regression, 85
Old Faithful example, local polynomial
  regression, 59–61
One-step-ahead forecast errors,
  multicollinear regressions, partial least
    squares, monthly U.S. unemployment
    rates example, 254–256
Online radio, market basket analysis,
  221–227
INDEX

Orange juice sales data case study:  
data preprocessing, 31–39  
penalty-based variable selection, LASSO algorithm, 78–82  
Outliers, hierarchical clustering, monthly U.S. unemployment rate example, 219  
Out-of-sample prediction:  
linear regression, Toyota used-car prices example, 50–51  
multicollinear regressions, partial least squares, monthly U.S. unemployment rates example, 253–256  
Overfitting:  
decision tree predictions, prostate cancer data example, 170–179  
linear regression, mean square error, 53–54

Pairwise correlations:  
alumni donations data preprocessing case study, 28–31  
dimension reduction, 236–246  
Pajek software, 272  
Parametric properties, regression models, 161  
Parsimony, statistical modeling, 67–70  
Partial least squares (PLS):  
principal components analysis, 247–256  
multicollinear regressions, generated data example, 249–251  
monthly U.S. unemployment, vector autoregressive model, 251–253  
text data mining, 268–271  
Partitioning, in decision trees, 163–167  
prostate cancer data, 177–179  
Pearson chi-square statistic, chi-square automatic interaction detection, 187–188  
Penalty-based variables:  
multiparameter regression models, 71–82  
orange juice sales data case study, 78–82  
Political sentiment example, text data mining, 266–268  
Porter stemming algorithm, text data mining, 258–271  
Posterior probability, classification, discriminant analysis, 151–160  
Predictions:  
decision trees for, 161–167  
prostate cancer data example, 170–179  
market basket analysis, income prediction, 227–234  
Predictor variables:  
logistic regression, 84–85  
a naive Bayesian analysis, delayed airplanes data, 128–130  
penalty-based variable selection, LASSO algorithm, 73–82  
standard linear regression, automobile fuel efficiency example, 45–47  
Preprocessing of data:  
alumni donations case study, 17–31  
birth data case study, 7–17  
case studies in, 7–39  
orange juice sales data case study, 31–39  
text data, 258–271  
Price variables, linear regression, Toyota used-car prices example, 48–51  
Principal components analysis (PCA):  
European protein consumption example, 238–243  
factor models, dimension reduction, 237–246  
monthly U.S. unemployment rates example, 243–246  
multicollinear regressions, partial least squares, 254–256  
multicollinear regressions, 247–256

Probabilities:  
Bayesian analysis, categorical predictor variables, 126–130  
binary classification and, 108–114  
decision problems, 108–109  
logistic regression, 84–85  
loan acceptance case study, 100–102  
market basket analysis, 221  
Probability cutoff, logistic regression, delayed airplanes data, 93–100  
Probit model, logistic regression, 85  
Prostate cancer data:  
decision tree analysis, 167–179  
least absolute shrinkage and selection operator algorithm and, 74–78  
Pruning, decision tree predictions, prostate cancer data, 171–179  
Purchase coincidence, market basket analysis, 220–234
INDEX

p-values, decision tree partitioning, chi-square automatic interaction detection, 188

Quadratic discriminant function, classification, discriminant analysis, 152–160
Query, k-nearest neighbor algorithm, 116–125

Random attribute selection, bagging classification and, 190–191
randomForest method, ensemble classification, 190–191
Random noise, decision tree predictions, prostate cancer data, 170–171
rattle library, graphical user interface, 194–195
Rcmdr software, graphical user interface, 194–195
Receiver-operating characteristic (ROC) function:
German credit data, 109–114
logistic regression, delayed airplanes data, 93–100
sensitivity and specificity, 109
text data mining, 261–266
Recommender systems, market basket analysis, 220–221
Recursive partitioning:
in decision trees, 163–167
rpart software, 185–186
Reference trees, decision tree analysis, prostate cancer data, 177–179
Regression. See Linear regression; Logistic regression
Regression coefficients, logistic regression, 85
Regression deviance:
decision tree partitioning, 163–167
standard linear regression, 40
Regression trees:
basic characteristics, 161–167
chi-square automatic interaction detection, 186–188
motorcycle acceleration example, 179–182
software packages for, 185–195
Removal rules, text data mining, 258–271
Residuals, standard linear regression, 40
Toyota used-car prices example, 51–52
Restaurant reviews example, text data mining, 261–266
R function:
logistic regression, 87
penalty-based variable selection, LASSO algorithm, 73–82
standard linear regression, 43
automobile fuel efficiency example, 44–47
statistical modeling parsimony, false discovery and, 70
Ridge regression, 72–82
R in a Nutshell: A Desktop Quick Reference, data preprocessing case study, 7–17
Root mean square error:
multicollinear regressions, partial least squares, monthly U.S. unemployment rates example, 254–256
standard linear regression, 42–43
automobile fuel efficiency example, 46–47
Toyota used-car prices example, 50–51
rpart software, tree construction using, 185–186
R–square:
multicollinear regressions, partial least squares, 249–251
standard linear regression, 40–41
R statistical software packages, 6
arules package, 223–228
cluster package, 217
decision tree construction, 163–167, 185–186
pruning applications, 171–179
igraph package, 272, 274–278
local polynomial regression, locfit R library, 58–65
logistic regression:
death penalty data, 89–92
German credit data, 103–107
mixOmics package, 249
mixtools package, 207
multinomial logistic regression, 134
preprocessing case studies using, 7
birth data case study, 10–17
randomForest package, 191
R statistical software packages (Continued)

rattle package, graphical user interface, 193–195
statnet package, 272, 278–292
stats package, 197, 217
support vector machine classification, 192
textir library, 260–268

Sales data preprocessing, orange juice case study, 31–39
Scale of data mining, defined, 2
Scatter plots:
cluster analysis, European protein consumption example, 199–201
data preprocessing:
alumni donations case study, 28–31
birth data case study, 11–17
orange juice case study, 34–39
decision trees, motorcycle acceleration example, 179–182
factor models, European protein consumption example, 239–243
standard linear regression, Toyota used-car prices example, 51–52
Sensitivity, classification problems, 109
Simple triplet matrix, multinomial logistic regression, 147–149
Single-linkage clustering, 213–217
“Slant” analysis, text data mining, 268–271
Smoothed density plots, data preprocessing, orange juice case study, 31–39
Smoothing function, local polynomial regression:
histogram smoothing, 58
nitric oxide emissions, 62–65
Soft counts, cluster analysis, expectation-maximization algorithm, 204–207
Software systems:
local polynomial regression, 58–65
multinomial logistic regression, 134
network data mining, 272
regression and classification trees, 185–195
Specificity, classification problems, 109
Speed of data mining, defined, 2
Splitting of data:
decision tree partitioning, 163–167

data mining and limitations of, 2–3
parsimony, false discovery, 67–70
Statistical inference, logistic regression, 85–86
statnet software package, 272, 278–292
stats software package, 197
Stemming, text data, 258–271
Stepwise regression techniques, standard linear regression, 42–43
Stop words, text data mining, 258–271
Sum of squares criterion:
cluster analysis, monthly U.S. unemployment rates example, 201–204
deviance node impurity, decision trees, 185–186
standard linear regression, 40
Supervised learning, defined, 3
Support vector machines (SVM), classification using, 192
tapply function:
alumni donations case study, 22–31
birth data preprocessing case study, 14–17
Targeted marketing, market basket analysis, 220–221
Terminal node, decision trees, 162
Test point, k-nearest neighbor algorithm, 116–125
Text data mining:
Gentzkow/Shapiro “slant” and partial least squares estimates, 268–271
inverse multinomial logistic regression, 259–260
political sentiment example, 266–268
restaurant reviews example, 261–266
sentiment analysis, 258–271
textir library, 260–268
Time sequence plots:
- cluster analysis, monthly U.S. unemployment rates example, 201–204
- orange juice sales data preprocessing, 31–39

Tokenization, text data, 258–271

Toyota used-car prices example, linear regression, 47–51

Tree-logic, predictions using, 161–162

Tumor log volume prediction, least absolute shrinkage and selection operator algorithm and, 74–78

Univariate autoregressive model, multicollinear regressions, partial least squares, monthly U.S. unemployment rates example, 253–256

Unsupervised learning, defined, 2–3

Update step, \( k \)-means clustering, 197

Vector autoregressive (VAR) models, multicollinear regressions, partial least squares, 251–253

V-fold cross-validation, decision tree analysis, prostate cancer data, 177–179

VGAM software package, multinomial logistic regression, forensic glass case study, 134–147

Weight functions:
- ensemble classification, 190–191
- local polynomial regression, 56–65

Zero correlation, partial least squares, multicollinear regressions, 248–256