INDEX

Accuracy, 145
Agglomerative hierarchical clustering, 93–105, 204–205
Alternative hypothesis, 40. See also Hypothesis testing
Analysis of data, see Data analysis
Analysis of variance, see One-way analysis of variance
Antecedent in decision tree rule, 117. See also Association rules
Artificial neural network, see Neural network
Association rules, 85, 111–122, 205–206
antecedent part of, 117
confidence measure of, 116–120
consequence part of, 117
extraction of, 115–116
grouping of, 113–115
lift measure of, 116–120
support measure of, 116–120
Assumptions, linear regression, 158–160
homoscedasticity of errors, 160

Bar chart, 25–26. See also Graphs
Bimodal distribution, 28. See also Distributions
Binary, see Variables, binary
Black-box, see Model, transparency
Box-and-whisker plots, see Box plots
Box plots, 30–32. See also Graphs
Brushing, 85
Central limit theorem, 38
Central tendency, 22–24
mean, 24
median, 23
mode, 23
Charts, see Graphs
Chi-square, 79–81, 203. See also Hypothesis tests

© 2014 John Wiley & Sons, Inc. Published 2014 by John Wiley & Sons, Inc.
INDEX

Classification, 142
models, 142–143. See also Model
and regression trees (CART), see
Decision trees
trees, 172–178, 211
Cleaning data, 48–49, 195–196
Clustering, 85, 88–111, 204–205. See also
Grouping
agglomerative hierarchical, 93–105,
205
k-means, 105–111, 205
Coefficient of determination, 155–156
Concordance, 145
Confidence, 116–120. See also
Association rules
Confidence interval, 36–39, 202
Consequence, 117. See also Association
rules
Constant, 22, 49, 195, 197
Contingency tables
in Traceis software, 199
for validating models, 146
for visualizing relationships, 67–68
Correlation coefficient ($r$), 69–71, 204
CRISP-DM, 16
Cross-classification tables, 67–68, 199.
See also Contingency tables
Cross-disciplinary teams, 6–7
Cross-validation, 144–145
k-fold, 144–145
leave-one-out, 145
Customer relationship management
(CRM), 3
Data
cleaning, see Data preparation
describing, 17
segmentation of, see Data preparation
transformation of, see Data preparation
use of, 1
Data analysis, 3–13
problems in, 3–8
processes of, 3–13
Data matrix, see Data table
Data mining process, 3–13
Data preparation, 4, 8–9, 13–14, 47–57
cleaning, 48–49, 195–196
combining variables, 54
converting continuous data to
categories, 53
converting text to numbers, 52–53
data transformation, 49–52
removing variables, 49
segmentation, 54–55
Data sets, see Data tables
Data sources, 2–3
Data tables, 18–20
Data transformation
Box–Cox, 52
exponential, 52
log, 52
Data warehouse, 3
Decision trees, 85, 122–138, 206–207
child node, 126
parent nodes, 126
scoring splits for categorical response,
128–131
scoring splits for continuous response,
130, 134–136
splitting criteria, 128–136
Dendrogram, 97–105. See also
Agglomerative hierarchical
clustering
Deployment of project, 4, 11–13
Descriptive statistics, 17, 202
Discriminant analysis, 179. See also
Model
Distance matrix, 94–96
Distance metrics, see Similarity measures
Distributions, 24–36
bimodal, 28
frequency, 25–29, 34–36
normal, 28
sampling, 37
standard $z$-, 39
Diverse set, 198
Enterprise Resource Planning (ERP), 3
Entropy, 132–133
Errors
in classification models, 146
false negatives, 146
false positives, 146
in linear regression models, 154
type I, 40
type II, 40
INDEX  233

Euclidean distance, 91–92. See also
Distance metrics
Experimental data, use of, 2–3
Extracting rules, 115–116. See also
Association rules
False negatives, 146. See also Errors
False positives, 146. See also Errors
Frequency distribution, 24–29, 34–36. See
also Distribution
Frequency histogram, see Histogram
F-test, 157–158. See also Hypothesis tests
Gain, 133. See also Decision trees
Gini, 131. See also Decision trees
Graphs
bar charts, 25–26
box plots, 30–32, 49
decision tree, 123–125
dendrogram, 98
frequency histogram, 27–29
heatmap, 104–105
scatterplot matrix, 201–202
scatterplots, 60–62
small multiples, 85–87
Grouping, 14, 83–140
in association rules, 85
in clustering, 85, 88–111, 204–205
in decision trees, 85
supervised methods, 124
unsupervised methods, 88
Histogram, 27–29. See also Graphs
Hypothesis test, 40–42
alpha(α), 40
alternative hypothesis, 40
null hypothesis, 40
p-value, 42
Hypothesis tests, 40
chi-square, 79–81, 203
F-test, 77, 158
t-test, 41, 74
Impurity in decision trees, 131
Intercept in linear regression, 151
Interquartile range, 30
Jaccard distance, 92–93. See also Distance
metrics
Kendall Tau, 72–74, 204
k-Means clustering, 89, 105–111, 205. See
also Grouping
k-Nearest neighbors (kNN), 167–172,
210–211
learning, 170
prediction, 170–172
Kurtosis, 35–36. See also Shape
Law of large numbers, 38
Least squares method, 150. See also
Model
Lift, 116–120. See also Association rules
Linear regression, 149–161, 207–210. See
also Model
assumptions, 158–160
modeling, 149–161
Linear relationship, 60, 69
Linearity assumption, 158–159
Linkage rules, 95–100. See also
Agglomerative hierarchical
clustering
average linkage, 95–100
complete linkage, 95–100
single linkage, 95–100
Location, see Central tendency
Log transformation, 52
See also Model
classification, 167
coefficients, 165–166
fitting, 162–165
interpreting, 165
likelihood ratio test, 166
Wald test, 166
Lower quartile, 30. See also Box plots
Mean, 24. See also Central tendency
Median, 23. See also Central tendency
Mode, 23. See also Central tendency
Model
assessing fit of, 145–148
classification, see Classification
models
discriminant analysis, see
Discriminant analysis
k-nearest neighbors, see
k-Nearest-neighbors
INDEX

Model (Continued)

- linear regression, see Linear regression
- logistic regression, see Logistic regression
- multiple linear regression, see Multiple linear regression
- multiple logistic regression, see Multiple logistic regression
- naïve Bayes, see Naïve Bayes classifiers
- neural networks, see Neural networks
- random forest, see Random forest
- regression tree, see Regression tree
- selection of variables in, 143
- significance of coefficients, 157, 165–168
- support vector machines, see Support vector machines
- transparency, 143
- Multiple linear regression, 153–161. See also Model
- Multiple logistic regression, 165
- Naïve Bayes classifiers, 179. See also Model
- Negative relationship between variables, 60
- Neural networks, 178. See also Model
- Nonlinear relationships between variables, 60–62
- Normal distribution, 28. See also Distribution
- Normalization, 8, 50–51
- Null hypothesis, 40
- Observations, 18
- One-way analysis of variance, 76–79, 203–204
- Operational databases, 1, 3
- Outliers, 61
- Parameters, 37
- Polls, 2
- Populations, 36
- Positive relationship between variables, 60
- Predictive models, 14–15, 141–183. See also Model
- Preparation, see Data preparation
- Principal component analysis, 182
- Process of analysis, 3–13
- Project management, 16
- p-Value, 42
- Quartiles, 29–30. See also Box plot
- Random forest, 179. See also Model
- Random subset, 198
- Range, 29
- Regression model, 143. See also Model
- Regression tree, 172–178, 211
- Relationships in data, 14, 59–82
- visualizing, 59–82
- Residuals, 148, 154
- Roles of individuals involved in analysis, 7
- Rules, see Association rules
- Sample standard deviation, 33–34
- Sample variance, 32
- Samples, 36
- Sampling distribution, 37. See also Distribution
- Scale, 21
- interval, 21
- nominal, 21
- ordinal, 21
- ratio, 21
- Scaling, decimal, 50
- Scatterplot matrix, 201–202. See also Graphs
- Scatterplots, 60–62. See also Graphs
- Segmentation, doing, 198–199
- SEMMA, 16
- Sensitivity, 146–148. See also Errors
- Shape, 34–36
- kurtosis, 34–36
- skewness, 35–36
- Similarity measures, 91–93
- Euclidean, 91
- Jaccard, 92
- Simple linear regression, 149–153. See also Linear regression; Model
- Skewness, 34–35
Slope, 149–151
Small multiples, 85–87. See also Graphs
Spearman Rho, 204
Specificity, 146–148. See also Errors
Splitting decision trees, see Decision trees
Standard deviation, 33–34
Standard error, 37–38
Stepwise regression, 161
Subsets, see Segmentation
Success criteria, 5
Summary charts, 63–67
Summary tables, 62–63, 84–85, 200
Support, 116–120. See also Association rules
Support vector machines, 178
Surveys, 2
Tables, 18–20
  cross-classification, 67–68
  summary, 62–63
Test set, 144
Traceis, 191–226
Training set, 144
Transformation, 49–54, 196–198
  Box–Cox, 52, 196
  decimal scaling, 50
  exponential, 52
  min–max, 50
  z-score, 34
Transparency of model, see Model, transparency
True negatives, 146. See also Errors, in classification models
True positives, 146. See also Errors, in classification models
t-Test, see Hypothesis tests
Two-way cross-classification table, see Contingency tables
Type I error, 40. See also Errors
Type II error, 40. See also Errors
Unstructured data, 55
Upper extreme, 30. See also Box plots
Upper quartile, 31. See also Box plots
Variables, 18, 20–22
  binary, 21
  constant, 22
  continuous, 20
  dichotomous, 21
  discrete, 20
  dummy, 52–53
  independent, 142
  labels for, 22
  response, 142
Variables, measuring relationships between, 69–81
  chi-square, 79–81
  correlation coefficient, 69
  Kendall Tau, 72–74
  t-test, 74–75
  F-test, 77
Variance, 32–33
X variables, see Variables, independent
y-intercept in linear regression, 151
X variables, see Variables, response
z-scores, 34
z-distribution, 39