

Index

- Accuracy, 9, 158–165, 167
- Agglomerative hierarchical clustering,
 - 111–120, 154
 - adjusting cut-off distances, 116
 - creating clusters, 114–116
 - example, 116–120
 - grouping process, 111–113
- Aggregate table, 39
- Aggregation, 31
- Alternative hypothesis, 73–74
- Anomaly detection, 211
- Artificial neural network, *see* Neural network
- Association rules, *see* Associative rules
- Associative rules, 129–139
 - antecedent, 134
 - confidence, 134–135
 - consequence, 134
 - example, 137–139, 230
 - extracting rules, 132–137
 - grouping, 130–132
 - lift, 135–137
 - support, 134
- Analysis of variance, *see* One-way analysis of variance
- Average, *see* Mean

- Bagging, 168
- Bar chart, 41
- Bin, 30
- Binary, *see* Variable, binary
- Binning, 30
- Black-box, 197
- Boosting, 168
- Box plots, 25, 45–46, 52, 233
- Box-and-whisker plots, *see* Box plots
- Budget, 12, 14–15
- Business analyst, 10

- Case study, 12
- Central limits theorem, 63
- Central tendency, 55–57, 96
- Charts, *see* Graphs
- Chi-square, 39, 67, 82–84, 91
 - critical value, 83–84
 - degrees of freedom, 84
 - distribution, 243
 - expected frequencies, 83
 - observed frequencies, 83
- Churn analysis, 210
- Claim, 72
- Classification, 158–162
- Classification and regression tree (CART),
 - see* Decision trees
- Classification models, 158, 182, 199,
 - 202, 233
- Classification trees, 181–184,
 - 203
- Cleaning, 24–26, 32, 219–220
- Clustering, 25, 110–129, 168
 - agglomerative hierarchical clustering, 111–120, 154
 - bottom-up, 111
 - hierarchical, 110
 - k-means clustering, 120–129, 154
 - nonhierarchical, 110, 120
 - outlier detection, 25
 - top-down, 120
- Common subsets, 49–51
- Concordance, 160
- Confidence, 158, 167
- Confidence intervals, 67–72
 - categorical variables, 72
 - continuous variables, 68–72
 - critical *t-value*, 69–71
 - critical *z-score*, 68–72

- Comparative statistics, 55, 88–96
 - correlation coefficient (r), 92–95, 97
 - direction, 90
 - multiple variables, 94–96
 - r^2 , 95–96
 - shape, 90
 - visualizing relationships, 90–91
- Constant, 21
- Consumer, 11
- Contingencies, 12, 14, 16
- Contingency tables, 36–39, 52, 91, 159–160
- Correlation coefficient (r), 92–94, 97
- Correlation matrix, 94–95, 46–48
- Cost/benefits, 12, 14, 16, 218
- CRISP-DM, 7
- Cross validation, 167, 233
 - leave-one-out, 167
- Cross-disciplinary teams, 11
- Customer relationship management (CRM), 18, 208
- Data, 19
- Data analysis, 1
 - process, 1, 6, 213–216
- Data analysis/mining expert, 10
- Data matrix, *see* Data table
- Data mining, 1
 - process, 1, 6, 213–216
- Data preparation, 2, 5–6, 17–35, 166, 168
 - cleaning, 24–26
 - data sources, 17–19
 - data transformations, 26–31
 - data understanding, 19–24
 - example, 217–225
 - planning, 14
 - removing variables, 26
 - segmentation, 31–32
 - summary, 32–33, 213
- Data quality, 216
- Data sets, 7
- Data smoothing, 30
- Data sources, 17–19
- Data tables, 19–20, 32, 36, 52
- Data visualization, 36–53
- Data warehouse, 18
- Decision trees, 139–154, 181–187
 - child node, 142
 - example, 151–153, 184–187
 - generation, 142–144
 - head, 143
 - leaf node, 142, 151
 - optimization, 141, 184
 - parent node, 142
 - parent-child relationship, 142
 - predicting, 182–184
 - rules, 151–152, 184–187
 - scoring splits for categorical response, 146–149
 - scoring splits for continuous response, 149–151
 - splitting criteria, 144–151
 - splitting points, 142–143
 - terminal node, 182
 - two-way split, 144
- Definition, 8–16
 - case study, 12–14
 - objectives, 8–9
 - deliverables, 9–10
 - roles and responsibilities, 10–11
 - project plan, 11–12
 - summary, 14, 16
- Deliverables, 9–10, 13, 16, 217, 223, 225
- Deployment, 208–212
 - activities, 209–210, 211
 - deliverables, 208–209, 211
 - example, 14, 218, 235
 - execution, 209, 211
 - measuring, 209, 211
 - monitoring, 209, 211
 - planning, 10, 209, 211
 - scenarios, 210–211
 - summary, 2, 5–6, 214–216
- Descriptive statistics, 4, 55–63
 - central tendency, 56–57
 - example, 62–63
 - shape, 61–62
 - variation, 57–61
- Discretization, 30–31
- Distance, 104–108, 111, 123, 154, 178
- Diverse set, 31, 102
- Double blind study, 18
- E-commerce, 210
- Embedded data mining, 209
- Entropy, 147–148
- Errors, 25, 82, 160
- Estimate, 156–157

- Euclidean distance, 105–107
- Experiments, 18
- Experimental analysis, 210
- Experimental design, 210
- Explanation, 158, 167–168
- Exploratory data analysis, 1

- False negatives, 159–160, 233–235
- False positives, 160, 233
- Finding hidden relationships, 2–5, 102, 217–218, 214–215, 230–232
- Forecast, 156
- Frequency distribution, 23
 - peak, 62
 - shape, 61–62
 - symmetry, 61–62
- Frequency polygrams, 40–41, 52

- Gain, 148–149
- Gaussian distribution, 23
- Gini, 147
- Graphs, 3, 36, 40–52
- Grouping, 4, 102–155
 - approaches, 108–109
 - associative rules, 129–139
 - by ranges, 103–104
 - by value combinations, 130–132
 - by values, 103–104
 - clustering, 110–129
 - decision trees, 139–153
 - methods, 153
 - overlapping groups, 109, 154
 - supervised, 108, 140, 154
 - unsupervised, 108, 129, 154

- Histograms, 23, 25, 41–43, 52
- Historical databases, 19
- Holdout set, 167
- Hypothesis test, 67, 72–82, 97, 104, 228–230
 - alpha (α), 74–75
 - alternative hypothesis, 73–74
 - assessment, 74–75
 - critical *z-score*, 74–76
 - null hypothesis, 73–74
 - paired test, 81–82
 - p-value*, 75–76
 - single group, categorical data, 78
 - single group, continuous data, 76–78
 - two groups, categorical data, 80–81
 - two groups, continuous data, 78–79

- Implementation, 2–6, 14–16, 214–215, 217–218, 225–237
- Impurity, 146
- Inconsistencies, 24–25
- Inferential statistics, 4, 55, 63–88
 - chi-square, 82–84
 - confidence intervals, 67–72
 - hypothesis tests, 72–82
 - one-way analysis of variance, 84–88
- Integration, 208–209, 211
- Intercept, *see* Intersection
- Interquartile range, 58
- Intersection, 169–172
- Inverse transformation, 26, 174
- IT expert, 11, 13, 16

- Jaccard distance, 107–108

- k-means clustering, 120–129, 154
 - example, 127–129
 - grouping process, 122–125
 - cluster center, 125–127
- k-nearest neighbors (kNN), 176–181, 203, 233
 - learning, 178–179
 - prediction, 180–181
- Kurtosis, 62–63

- Least squares, 172–173
- Legal issues, 11–12, 16
- Linear relationship, 44, 90–91, 169–173, 162
- Linkage rules, 113–114
 - average linkage, 113–114
 - complete linkage, 113–114
 - single linkage, 113–114
- Logistic regression, 202
- Lower extreme, 45
- Lower quartile, 45

- Mathematical models, 4
- Maximum value, 39, 57
- Mean, 39–40, 45, 57, 96
- Median, 39, 45, 96
- Minimum value, 39, 57
- Misclassification, 147
- Missing data, 25–26

- Mode, 56, 96
- Model parameters, 166
- Modeling experiment, 166–167
- Multiple graphs, 46–52
- Multiple linear regression, 199–202
- Multivariate models, 166

- Naïve Bayes classifiers, 202
- Negative relationship, 44–45, 90
- Neural networks, 187–199, 203, 233–236
 - activation function, 189–190
 - backpropagation, 192–196
 - calculations, 188–190
 - cycles, 197
 - epoch, 197
 - error, 191–196
 - example, 194–196, 197–199, 233–236
 - feed forward, 191
 - hidden layers, 187, 190, 193, 196
 - input layer, 187–188
 - layers, 187–188
 - learning process, 191–192
 - learning rate, 194–196
 - nodes, 187–188
 - optimize, 196–197
 - output layers, 187–188, 192–196
 - prediction, 190–191
 - topology, 194
 - using, 196–197
 - weights, 188–197
- Nonlinear relationships, 44–45, 90–91, 172–176
- Nonnumeric terms, 25
- Nonparametric procedures, 23–24
- Normal distribution, 23–24, 239
- Normalization, 26–29
- Null hypothesis, 73–74

- Objectives, 8–9, 16, 213–216
- Objects, 19
- Observational study, 18
- Observations, 19–20, 36
- Occam's Razor, 167
- One-way analysis of variance, 67, 84–89, 97
 - between group variance, 87
 - degrees of freedom, 88
 - F-distribution, 247
 - F-statistic, 86–88
 - group means, 86
 - group variances, 86
 - mean square between, 87
 - mean square within, 86–87
 - within group variance, 86–87
- On-line Analytical Processing (OLAP), 1
- Operational databases, 18
- Outliers, 25, 109–110, 121–122

- Parameters, 54
- Parametric procedures, 23
- Partial least squares, 202
- Placebo, 80
- Point estimate, 67–68
- Polls, 18
- Pooled standard deviation, 79
- Pooled variance, 79
- Population variance, 59
- Populations, 9, 54, 63
- Positive relationship, 44–45, 90
- Prediction, 3–6, 156–207. *See also* Predictive models
- Prediction models, *see* Predictive models
- Predictive models, 9, 31, 156, 217
 - applying, 158, 167–168, 203
 - building, 158, 166–167, 203
 - classification and regression trees, 181–187
 - classification models, 158–162
 - defined, 156–158
 - grouping prior to building, 102
 - k-nearest neighbors, 176–181
 - methods, 158, 199–202
 - neural networks, 187–201
 - regression models, 162–165
 - simple regression models, 169–176
 - specification, 9–10
- Predictors, *see* Variables, descriptors
- Preparation, *see* Data preparation
- Principal component analysis, 35
- Privacy issues, 11–12, 16
- Probability, 65
- Problem definition, *see* Definition
- Project leader, 10, 214
- Project management, 16
- Project plan, 11–12
- Proportion, 66–67, 72, 78, 80–81
- Purchased data, 19
- p-value*, 75–76

- Quality, 17, 159
- Quality control, 210
- Quartiles, 57–58, 96
- Query, 103–104

- Random forest, 202
- Random subset, 31, 63
- Range, 57, 96
- Regression, 25, 162–165
- Regression model, 158, 162–165, 199–203
- Regression trees, 181–187, 203
- Report, 208, 211, 217
- Research hypothesis, *see* Alternative hypothesis
- Residuals, 163–165
- Review, 12, 210–211, 235
- Risks, 12, 14, 16
- Roles and responsibilities, 10–11, 13, 16, 217
- Root mean square, 59
- r^2 , 95–96, 163–165
- Rule-based classifiers, 202
- Rules, *see* Associative rules

- Sample standard deviation, 59–60
- Sample variance, 58
- Samples, 54, 63
- Sampling distribution, 63–67
- Sampling error, 63
- Scale, 21–22
 - interval, 22
 - nominal, 21
 - ordinal, 21–22
 - ratio, 22
- Scatterplot matrix, 48, 94–95
- Scatterplots, 25, 43–45, 52, 91, 94–95, 162–165
- Searching, 4, 103–104
- Segmentation, 31–32, 102, 168
- SEMMA, 7
- Sensitivity, 160–162, 233–237
- Sequence data mining, 238
- Sigmoid function, 189–190
- Similarity measures, 104–108
- Simple linear regression, 169–172, 203
- Simple models, 166
- Simple nonlinear regression, 172–176
- Skewness, 61–62

- Slope, 169–172
- Specificity, 160–162, 233–236
- Standalone software, 209, 211
- Standard deviation, 39, 59–60, 96
 - of the proportions, 67, 72
 - of the sample means, 65–66
- Standard error
 - of proportions, 67, 72
 - of the means, 66
- Statistical tables, 239–255
- Statistics, 54–101
- Student's *t*-distribution, 69–71, 239
- Subject matter experts, 10, 16
- Subsets, *see* Segmentation
- Success criteria, 8, 16
- Sum, 39
- Sum of squares of error (SSE), 149–151, 179
- Summarizing the data, 2–5, 217, 225–230
- Summary tables, 3–4, 39–40, 52
- Support vector machines, 202
- Surveys, 18, 211

- Tables, 19–20, 36–40, 49, 52
- Tanh function, 189–190
- Targeted marketing campaigns, 210
- Test set, 167
- Text data mining, 237
- Time series data mining, 238
- Timetable, 12, 14–16, 217–218
- Training set, 167
- Transformation, 26–32, 221–223
 - Box-Cox, 28–29
 - decimal scaling, 27
 - exponential, 28
 - min-max, 27–28, 221–222
 - z*-score, 27
- True negatives, 160
- True positives, 159–160
- t*-value, 69–71
 - degrees of freedom, 71
- Two-way cross-classification table, 36–39
- Type I error, 82
- Type II error, 82

- Units, 21, 26
- Upper extreme, 45, 57
- Upper quartile, 45, 57

Value mapping, 29–30

Variables, 19–24, 36

 binary, 21

 characterize, 20–24

 comparing, 88–97

 constant, 21

 continuous, 20–21

 descriptors, 22–23

 dichotomous, 21

 discrete, 20–21

 dummy, 29–30

 labels, 22–23

 removing, 26, 32, 220

 response, 22, 38

 roles, 22–23, 217

Variance, 58–59, 96

Variation, 55, 57–61, 96

Visualizing relationships,
 90–91

Voting schemes, 168

χ^2 , *see* Chi-square

X variables, *see* Variables,
 descriptors

y-intercept, *see* Intersection

Y variables, *see* Variables,
 response

z-score, 25, 60–61, 96