
INDEX

- A priori algorithm
 - generating association rules, 186–187
 - generating frequent itemsets, 185–186
- Adaptation, 165, 166
- Affinity analysis, 180–182
- Anomalous fields, 50–52
- Antecedent, 183
- Application of neural network modeling, 143–145
- A priori algorithm, 184–189
- A priori property, 184–185
- Association, 17
- Association rules, 17, 180–199
 - affinity analysis, 180–182
 - antecedent, 183
 - a priori algorithm, 184–189
 - a priori property, 184–185
 - generating association rules, 186–187
 - generating frequent itemsets, 185–186
 - confidence, 184
 - confidence difference method, 195–196
 - confidence ratio method, 195–196
 - consequent, 183
 - data representation for market basket analysis,
 - 182–183
 - tabular data format, 182–183
 - transactional data format, 182–183
 - definition of, 183
 - extension to general categorical data, 189–190
 - frequent itemset, 184
 - generalized rule induction (GRI), 190–196
 - application of GRI, 191–193
 - behavior of the J statistic, 191
 - J -measure, 190–191
 - itemset, 184
 - itemset frequency, 184
 - local patterns versus global models, 197–198
 - market basket analysis, 180–182
 - procedure for mining, 184
 - supervised or unsupervised learning, 196
 - support, 184
 - when not to use association rules, 193–196
- Back-propagation, 135
 - example of, 137–138
 - rules, 136–137
- Balancing the dataset, 104
- Bank of America, 1
- Bar chart, 46–48
- Between cluster variation (BCV), 149, 154
- Bhandari, Inderpal, 3
- Bias-variance trade-off, 93–95
 - mean square error, 95
- Binary trees, 109
- Binning (banding), 61–62
- Boston Celtics, 3
- Bremmer, Eric, 2
- Brown, Dudley, 3
- C4.5 algorithm, 116–127
- Candidate splits, 111
- CART, *see* Classification and regression trees
- Case studies
 - Daimler-Chrysler: analyzing automobile warranty claims, 8–9
 - mining association rules from legal data bases, 19–21
 - predicting abnormal stock market returns, 18–19
 - predicting corporate bankruptcies using decision trees, 21–22
 - profiling materials and market using clustering, 23–24
- CIO Magazine, 1
- Claritas, Inc., 16
- Classification, 14–15, 95–96, 107–127, 128–146
- Classification and regression trees (CART), 109–115, 122–126
 - optimality measure, 110
- Classification error, 114
- Clinton, President Bill, 2
- Cluster centroid, 153

218 INDEX

- Clustering, 16–17, 147–162, 163–179
 - between cluster variation (BCV), 149, 154
 - hierarchical methods, 149–153
 - agglomerative methods, 149–153
 - average linkage, 150, 152–153
 - complete linkage, 150–152
 - dendrogram, 149
 - divisive methods, 149–150
 - hierarchical clustering, 49
 - single linkage, 150–151
 - k*-means, *see k*-means clustering
 - within cluster variation (WCV), 149, 154
- Cluster membership for making predictions, 161
- Cluster profiles, 175–177
- Cluster validity, 170
- Combination function, 101–103
 - for neural networks, 132–133
- Competition, 165, 166
- Competitive learning, 163
- Completely connected network, 131–132
- Confidence, 122, 184
- Confidence interval estimate, 73–74
- Confidence level, 73
- Confluence of results, 19, 212
- Confusion matrix, 203–204
- Consequent, 183
- Cooperation, 165, 166
- Correlation, 53–54, 78
- Cross Industry Standard Process for Data Mining (CRISP-DM), 5–7
 - business research understanding phase, 8, 18–19, 21, 23
 - data preparation phase, 7–8, 18, 20–21, 23
 - data understanding phase, 2, 8, 18, 20–21, 23
 - deployment phase, 7, 9, 19, 21–22, 24
 - evaluation phase, 7, 9, 19, 20, 22, 24
 - modeling phase, 7, 9, 18, 20–21, 23
- Cross-tabulations, 47–48
- Cross-validation termination, 139

- Daimler-Chrysler, 5, 8–9
- Data cleaning, *see* Data preprocessing
- Data mining
 - case studies, *see* Case studies
 - cross industry standard process (CRISP-DM), 5–7
 - definition of, 2
 - easy to do badly, xii, 5
 - examples of
 - Bank of America, 1
 - Boston Celtics, 3
 - brain tumors, 2
 - New York Knicks, 3
 - Clinton, President Bill, 2
 - fallacies of, 10–11
 - need for human direction, 4, 10
 - software
 - Advanced Scout by IBM, 3
 - Clementine by SPSS, Inc., 3
 - Enterprise Miner by the SAS Institute, 158
 - Insightful Miner by Insightful Corp., 31
 - Minitab, 12
 - tasks, *see* Tasks, data mining
 - why data mining, 4
- Data preprocessing, 27–40
 - data cleaning, 28–30
 - ambiguous coding, 28–30
 - anomalous values, 28–29
 - character versus numeric formatting, 28–29
 - min-max normalization, 36–37
 - z*-score standardization, 37–38
 - identifying misclassifications, 33–34
 - missing data, 30–33
 - replace with constant, 31
 - replace with mean or mode, 31–32
 - replace with random value from distribution, 31–33
 - outliers, graphical methods for identifying, 34–35
 - definition of, 34
 - histogram, 34–35
 - interquartile range, 39
 - quartiles, 39
 - scatterplot, 35
 - outliers, numerical methods for identifying, 38–39
 - why preprocess data, 27–28
- Data set
 - adult*, 143
 - cereals*, 75
 - churn*, 42
- Data transformation, *see* Data preprocessing
- Decision cost/benefit analysis, 207–208
- Decision nodes, 107–108
- Decision rules, 121–122
- Decision tree pruning, 114–115, 121
- Decision trees, 107–127
 - C4.5 algorithm, 116–127
 - entropy, 116
 - entropy as noise, 117
 - entropy reduction, 116
 - information as signal, 117
 - information gain, 116
 - classification and regression trees (CART), 109–115, 122–126
 - binary trees, 109
 - candidate splits, 111
 - classification error, 114
 - optimality measure, 110
 - tree pruning, 114–115

- comparison of the CART and C4.5 algorithms, 122–126
 - minimum records per node, 125
- decision nodes, 107–108
- decision rules, 121–122
 - confidence, 122
 - support, 122
- group node, 107–108
- leaf nodes, 107–108
- requirements for, 109
- Democratic Leadership Council, 2
- Dendrogram, 149
- Description, 11
- Description task, model evaluation techniques, 201
- “Different from” function, 100
- Distance function (distance metric), 99–101
 - city block distance, 148
 - Euclidian distance, 99, 148
 - Minkowski distance, 148
- Draftsman’s plot, 83–84

- Entropy, 116
- Entropy reduction, 116
- Error rate, classification, 203–204
- Error responsibility, 137
- Estimated regression equation (ERE), 76
- Estimation, 12–13, 67–88, 104–105, 131
- Estimation and prediction using neural networks, 131
- Estimation error, 77, 201
- Estimation task, model evaluation techniques, 201–202
- Euclidian distance, 99, 148
- Exploratory data analysis, 41–66
 - anomalous fields, 50–52
 - binning (banding), 63
 - categorical variables, 45–50
 - comparison bar chart, 46–48
 - cross-tabulations, 47–48
 - directed web graph, 50
 - two-way interactions among categorical variables, 48–50
 - dealing with correlated variables, 44–45
 - getting to know the data set, 42–44
 - multivariate relationships, 59–61
 - interaction, 59–60
 - three dimensional scatterplot, 60–61
 - numerical variables, 52–59
 - correlation, 53–54
 - graphical analysis of numerical variables, 54–59
 - normalized histogram, 55–58
 - retaining variables in model, 58–59
 - selecting interesting subsets of the data, 61–62
 - versus hypothesis testing, 41–42
- Extension to general categorical data, 189–190
- Extrapolation, 79

- False negative rate, 204
- False negatives, 204
- False positive rate, 204
- False positives, 204
- FBI, 2
- Feedforward network, 131–132

- Gains charts, 208–211
- Gartner Group, 2
- Generalized rule induction (GRI), 190–196
 - application of, 191–193
- Global minimum, 139
- Gradient descent method, 135–136
- GRI, *see* Generalized rule induction
- Grinstein, Georges, 5
- Group node, 107–108

- Hidden layer, 132
 - size of, 132
- Hierarchical clustering, 149
- Hipp, Jochen, 8
- Histogram, normalized, 55–58

- ID3 algorithm, 116
- Identifying misclassifications, *see* Data preprocessing
- Indicator variables for neural networks, 130
- Information gain, 116
- Input and output encoding, neural networks, 129–131
- Input layer, 131–132
- Insider trading, 18
- Instance-based learning, 96
- Intelligent Data Analysis* (journal), 19
- Interquartile range, 39
- Itemset, 184
 - frequency, 184
 - frequent, 184

- J*-measure, 190–191
- J*-statistic, behavior of, 191

- Kelly, Chris, 1
- k*-means clustering, 153–162
 - application of, using SAS Enterprise Miner, 158–161
 - choosing *k*, 157
 - cluster centroid, 153
 - example of, 153–158
 - using cluster membership to make predictions, 161

220 INDEX

- k*-nearest neighbor algorithm, 90–106
 - choosing *k*, 105–106
 - combination function, 101–103
 - simple unweighted voting, 101–102
 - weighted voting, 102–103
 - database considerations, balancing the dataset, 104
 - distance function (distance metric), 99–101
 - “different from” function, 100
 - Euclidian distance, 99
 - similarity, 99–101
 - triangle inequality, 99
 - estimation and prediction, locally weighted
 - averaging, 104–105
 - instance-based learning, 96
 - stretching the axes, 103–104
- Kohonen learning, 165
- Kohonen networks, 163–179
 - adaptation, 165, 166
 - algorithm, 166
 - application of clustering using, 170–177
 - cluster membership as input to downstream
 - models, 177
 - cluster profiles, 175–177
 - cluster validity, 170
 - competition, 165, 166
 - cooperation, 165, 166
 - example of a Kohonen network study, 166–170
 - learning, 165
 - neighborhood size, 167
 - self-organizing maps (SOMs), 163–165
 - competitive learning, 163
 - scoring function, 163–164
 - winning node, 165
 - weight adjustment, 167–169
- Kohonen, Tuevo, 163
- Layered network, 131–132
- Leaf nodes, 107–108
- Learning rate, 139–140
- Least squares, 78
- Lift, 208–209
- Lift charts, 208–211
- Lindner, Guido, 8
- Linkage
 - average, 150, 152–153
 - complete, 150–152
 - single, 150–151
- Local minimum, 139
- Local patterns versus global models, 197–198
- Louie, Jen Que, 10
- Margin of error, 73–74
- Market basket analysis, 180–182
 - data representation, 182–183
- Mean, 69–70
- Mean square error (MSE), 95, 201
- Measures of variability, 70
- Median, 70
- Minimum descriptive length principle, 201
- Misclassification cost adjustment, 205–207
- Missing data, *see* Data preprocessing
- Mode, 70
- Model complexity, 92–93
- Model evaluation techniques, 200–212
 - confluence of results, 212
 - classification task, 203–211
 - confusion matrix, 203–204
 - decision cost/benefit analysis, 207–208
 - error rate, 203–204
 - false negative rate, 204
 - false negatives, 204
 - false positive rate, 204
 - false positives, 204
 - gains charts, 208–211
 - lift, 208–209
 - lift charts, 208–211
 - misclassification cost adjustment, 205–207
 - type I error, 205
 - type II error, 205
 - description task, 201
 - minimum descriptive length principle, 201
 - Occam’s razor, 201
 - estimation and prediction tasks, 201–202
 - estimation error, 201
 - mean square error (MSE), 201
 - residual, 201
 - standard error of the estimate, 202
 - interweaving model evaluation with model
 - building, 211–212
- Mohammed Atta, 2
- Momentum term, 140–142
- Multicollinearity, 84
- Naisbitt, John, 4
- NCR, 5
- Neighborhood size, 167
- Neural networks, 128–146
 - application of neural network modeling, 143–145
 - back-propagation, 135
 - example of, 137–138
 - minimizing SSE, 135
 - stochastic back-propagation, 137
 - back-propagation rules, 136–137
 - error responsibility, 137
 - estimation and prediction, 131
 - gradient descent method, 135–136
 - provides direction for adjusting weights, 135

- learning rate, 139–140
 - helps move weights to global minimum, 139
 - reducing the learning rate, 140
- momentum term, 140–142
 - momentum represents inertia, 141
- neurons, 128–129
 - indicator variables, 130
 - input and output encoding, 129–131
- sensitivity analysis, 142–143
 - opacity of neural networks, 142
- sigmoid activation function, 134
 - squashing function, 134
- simple example of a neural network, 131–134
 - combination function, 132–133
 - completely connected network, 131–132
 - feedforward network, 131–132
 - hidden layer, size of, 132
 - input layer, 131–132
 - layered network, 131–132
 - nonlinear behavior, 133
 - output layer, 132
 - sigmoid function, 133
 - weights, connection, 132
- termination criteria, 139
 - cross-validation termination, 139
 - global minimum, 139
 - local minimum, 139
- Neurons, 128–129
- New York Knicks, 3
- Nonlinear behavior of neural networks, 133
- Normal plot of the residuals, 85

- Occam's razor, 201
- Outliers, methods for identifying, *see* Data preprocessing
- Output layer, 132
- Overfitting, 92–93

- Parameter, 71
- Plot of standardized residuals versus fitted values, 86
- Point estimate, 72
- Point estimation, 72
- Population, 71
- Precision, 74
- Prediction, 13, 67–88, 104–105, 131
- Prediction error, 77
- Prediction task, *see* Model evaluation techniques
- Procedure for mining, 184

- Quartiles, 39
- Quinlan, Ross, 116

- Range, 71
- Regression coefficients, 76
- Regression line, 76–77
- Regression, simple linear, 12
 - requirements for, 109
- Residual, 77, 201

- Sample, 71
- Sampling error, 73
- Scatterplot, three dimensional, 60–61
- Scoring function, SOMs, 163–164
- Self-organizing maps (SOMs), 163–165
- Sensitivity analysis, 142–143
- Sigmoid activation function, 134
- Sigmoid function, 133
- Similarity, 99–101
- Simoudis, Evangelos, 2
- Slope, 76
- Sodium/potassium ratio, 14–15
- SPSS, Inc., 2, 5
- Squashing function, 134
- Standard deviation, 71
- Standard error of the estimate, 202
- Statistic, 71
- Statistical approaches to estimation and prediction, 67–89
 - bivariate methods, 75–82
 - confidence in our estimates, 73
 - sampling error, 73
 - confidence interval estimation, 73–75
 - confidence interval estimate, 73–74
 - confidence level, 73
 - margin of error, 73–74
 - precision, 74
 - t -interval for the mean, 74–75
 - confidence intervals for the mean value of y given x , 80–82
 - extrapolation, 79
 - dangers of, 79–80
 - measures of center, 69–70
 - mean, 69–70
 - measures of location, 69
 - mode, 70
 - measures of spread, 70–71
 - measures of variability, 70
 - range, 71
 - standard deviation, 71
 - multiple regression, 83–88
 - draftsman's plot, 83–84
 - multicollinearity, 84
 - prediction intervals for a randomly chosen value
 - of y given x , 80–82
 - unusual observations, 82
 - simple linear regression, 75–82
 - correlation, 78
 - estimated regression equation (ERE), 76
 - estimation error, 77

222 INDEX

- Statistical approaches to estimation and prediction (*Continued*)
 - least squares, 78
 - prediction error, 77
 - regression coefficients, 76
 - regression line, 76–77
 - residual, 77
 - slope, 76
 - y-intercept, 76
- statistical inference, 71–75
 - estimation, 72
 - parameter, 71
 - point estimate, 72
 - point estimation, 72
 - population, 71
 - representative sample of population, 71–72
 - statistic, 71
- univariate methods, 69–75
- verifying model assumptions, 85–86
 - normal plot of the residuals, 85
 - plot of standardized residuals versus fitted values, 86
- Stochastic back-propagation, 137
- Supervised methods, 91
- Supervised modeling, methodology for, 91–93
 - model complexity, 92–93
 - overfitting, 92–93
 - test data set, 91–92
 - training data set, 91–92
 - underfitting, 92–93
 - validation data set, 92
- Supervised versus unsupervised learning, 90–91, 196
 - supervised methods, 91
 - unsupervised methods, 90
- Support, 122, 184
- Tabular data format, 182–183
- Target variable, 14
- Tasks, data mining, 11–17
 - association, 17
 - classification, 14–15
 - clustering, 16–17
 - description, 11
 - estimation, 12–13
 - prediction, 13
- Termination criteria, 139
- Terrorism, 2
- Test data set, 91–92
- t*-interval for the mean, 74–75
- Training data set, 14, 91–92
- Transactional data format, 182–183
- Triangle inequality, 99
- Type I error, 205
- Type II error, 205
- UCI Repository of Machine Learning Databases, 42, 122
- Underfitting, 92–93
- Unsupervised methods, 90
- Unusual observations, 82
- Validation data set, 92
- Voting
 - simple unweighted, 101–102
 - weighted, 102–103
- Web graph, 50
- Weight adjustment, 167–169
- Weights, connection, 132
- Within cluster variation (WCV), 149, 154
- y-intercept, 76