Note: Page numbers in italics refer to Figures; those in bold to Tables

a priori algorithm
association rules
  minimum confidence, 609, 610
  Modeler results, 609, 611, 611
  one antecedent, 609, 610
  two antecedents, 609, 609
  two-step process, 606, 608
frequent itemsets, 607, 607–8
ADABoost algorithm
  final boosted classifier, 646, 647
  initial base classifier, 644, 645
  original dataset, 644, 644
  second base classifier, 644, 645
  third base classifier, 646, 646
adjusted cost matrix, 473, 473
  bank loan, 481, 481
  equivalent cost, 477, 478
  false negative cost, 477, 477
  false positive cost, 477, 477
  retailer cost, 476, 476
analysis of variance (ANOVA)
  MINITAB results, 154, 156, 156
  MSTR, 155
  multiple regression model, 240, 240–241
  R code, 158
  sample mean age, 153–4, 153–5, 154
  sum of squares, 155, 155
artificial neuron model, 338, 339, 342
association rules
  a priori property (see a priori algorithm)
  affinity analysis, 603
  antecedent and consequent, 603, 605–6
  business and research, 603
categorical data, 611–12, 612
  confidence and support, 603, 606
frequent itemsets, 606
  J-measure, 613–14
  lift ratio, 615–16
  market basket analysis, 603–5, 605
patterns and models, 617–18, 618
R code, 618
strong rules, 606
supervised/unsupervised learning, 616–17
worst case scenario, 614, 614–15, 615
attribute-relation file format (ARFF) file, 432, 433
back-propagation algorithm
  cross validation termination, 348–9
downstream node, 347–8
error propagation, 346–7
learning rate, 349–50, 350
momentum term, 350–352, 351
squared prediction error, 344
upstream node, 347–8
bagging model
  algorithm for, 641
  bootstrap samples, 641–3, 642
  vs. CART model, 647–8, 648
  prediction method, 642–3, 643
  R code, 649
  stable/unstable classification, 640, 640
balanced iterative reducing and clustering
  using hierarchies (BIRCH) clustering
  bank loans data set
    cost matrix, 576, 576
    data sorting, 577–9, 578, 579
    No Interest model, 572, 574, 575, 576
    With Interest model, 572, 573, 575, 575, 576
CF/CF tree
Additivity Theorem, 562
algorithm, 560
building process, 562–3
clustering sub-clusters, 564–5, 570
definition, 561
one-dimensional toy data set, 565–9, 566–9
radius, 564
tree structure, 562, 563
Modeler’s two-step algorithm, 571, 572
optimal number of clusters, 742–4, 743, 744
pseudo-F statistic method, 571, 571
R code, 579–80
two-step clustering, 560
baseline model
Captain Kirk’s situation, 167
regression model, 168
Bayesian approach see also Naïve Bayes classifier
balancing data set, 422–3
drawbacks, 416
frequentist/classical approach, 414
likelihood function, 416
MAP method (see maximum a posteriori (MAP))
marginal distribution, 415
MCMC methods, 416
posterior distribution, 415
posterior odds ratio, 420–422
prior distribution, 414–15
R code, 444–7
Bayesian belief networks (BBNs)
clothing purchase, 436–8, 437
conditional probability, 440
directed acyclic graph, 436
joint probability distribution, 440
prior probabilities, 439
WEKA
Explorer Panel, 441, 442
positive and negative classification, 443
prior probabilities, 442, 443
test set predictions, 442–4, 443
bias–variance trade-off, 164, 164–5, 165
boosting model
ADABOOST algorithm
final boosted classifier, 646, 647
initial base classifier, 644, 645
original dataset, 644, 644
second base classifier, 644, 645
third base classifier, 646, 646
vs. CART model, 647–8, 648
R code, 649
C4.5 algorithm
adult data set, 333–4, 335
candidate splits, 326–7, 327
capital gains, 334–5
categorical variables, 333
decision node A, 330–331, 330–331
entropy reduction, 326–8
initial split, 329–30, 330
marital status, 333–4
numerical variables, 333
savings split, 328, 331, 332
threshold partition, 328–9, 329
training data set, 320, 326
churn data set
account length, 383–4, 384
adult data set, 387
age predictor, 388, 388–90, 389
area code field, 72–3, 72–3
balanced data set, 422–3
categorical variables
clustered bar chart, 59, 60, 61
comparative pie chart, 59–60, 60, 61, 62
directed web graph, 63, 64, 65
International Plan, 57–9, 58, 59
marginal distribution, 58–9, 59
non-churners, 56, 58
row percentages, 60, 61
software packages, 56–7
two-way interaction, 63, 63, 64
voice mail plan, 62–3, 63, 63
clustering analysis
CART decision trees, 632, 633
churn proportion, 630, 632
contingency tables, 630, 632
international plan people, 630, 633
no-plan majority, 630, 632–3
voice mail plan people, 630, 633
conditional independence, 424–6, 425
continuous predictor (see continuous predictor)
correlation coefficient
account length, 80, 81
matrix plot, 79, 79
Minitab regression tool, 80, 80
INDEX 783

optimal solution, 80
p-values, 79–80, 80
thresholds, 78
customer service calls, 70–71, 71–2
data preparation
  contingency table, 75, 75
  HighDayEveMins_Flag variable, 76–7, 77
voice mail messages, 75–6, 76
z-score standardization, 77–8, 78
day minutes, 70–71, 71–2
dichotomous predictor (see dichotomous predictor)
education-num variable, 390–392, 391, 391–2
field values, 55–6, 56
flag variables, 535
hours-per-week, 392–4, 393, 393–4
income overlay, 387–8, 388
International Plan, 384–6, 535, 535, 537, 537
maximum a posteriori
  complement probabilities, 418, 419
  conditional probability, 417
International Plan, 417, 420
  joint conditional probabilities, 419, 419
  marginal and conditional probabilities, 418, 418
  posterior probabilities, 418, 418
Voice Mail Plan, 417, 420
  multivariate graphics, 69–70, 70–71
  numerical predictors
    binning methods, 72–5, 73–5, 74
    churn proportion, 65–6, 66
    churners vs. non-churners, 65, 66
    customer service call, 65, 65–7
    International Calls, 68, 68
    normalized and non-normalized histogram, 66–8, 66–8
test, 68–9, 69
  numerical variables, 535
  polychotomous predictor (see polychotomous predictor)
  posterior odds ratio, 420–422
  vs. variables, 535
  visualization, 56, 57
  voice mail plan, 384–7
  VoiceMail Plan adopters, 535, 536, 538, 538
classification and regression trees (CART)
  adult data set, 333–4, 335
  bank loans, 481, 481
  candidate splits, 321, 321
capital gains, 334–5
categorical variables, 333
classification error, 318, 323–4, 327
  components, 321, 322
  contingency table, 495, 495, 500, 501
  cost matrix, 500, 501
data-driven misclassification costs, 481–2, 481–2
  decision node A, 323, 323, 324
  decision node R, 323–4, 324, 325
decision tree output, 556, 556–7
  estimated revenue increase, 503
  evaluation measures, 459, 501–2, 502
  initial split, 322, 322
  lift chart, 510, 511
  marital status, 333–4
  maximum value, 321–2, 322
  numerical variables, 333
  optimal split, 320
  scaled cost matrix, 500, 501
  training data set, 320, 320–321
cluster feature (CF)
  Additivity Theorem, 562
  building process, 562–3
  clustering sub-clusters, 564–5, 570
  definition, 561
  one-dimensional toy data set, 565–9, 566–9
  radius, 564
  tree structure, 562, 563
cluster validation
  cross-validation
    loans data sets, 594–7, 595, 596
    methodology, 594
    prediction strength, 594
    R code, 598–9
  loans data sets, 594–7, 595, 596
  methodology, 594
  prediction strength, 594
  pseudo-F statistic method
    clustering model, 591
    distribution, 592
    Iris data set, 592–3
    R code, 598
SSB and SSE, 590
R code, 598–9
silhouette method
cluster validation (Continued)
  cohesion/separation, 583–4
  Iris data set, 585–9, 585–9, 587, 588
  mean silhouette, 584–5, 585
  positive/negative values, 583–4
  R code, 597–8

clustering analysis
  CART decision trees, 632, 633
  churn proportion, 630, 632, 632
  contingency tables, 630, 632
  definition, 523
  hierarchical clustering
    agglomerative clustering, 525
    complete-linkage clustering, 527–9, 528
    divisive clustering methods, 525
    single-linkage clustering, 526–7, 527
  international plan people, 630, 633
  k-means clustering algorithm
    data points, 530, 530
    definition, 529
    MSE, 529
    processing steps, 530–533, 531, 531–3, 533
    pseudo-F statistic method, 530
  SAS Enterprise Miner (see churn data set)
    statistics behavior, 533–4
    no-plan majority, 630, 632–3
    R code, 538–9
    voice mail plan people, 630, 633
  confidence interval
    customer service call, 135–6, 135–6
    lower bound, 136
    margin of error, 136–7
    population proportion, 137–8
    subgroup analyses, 135
    t-interval, 134–5
    upper bound, 136
  continuous predictor
    categorical predictor, 375, 377
    confidence intervals, 377
    day minute usage, 373, 373–5, 374
    deviance, 375, 375–6
    p-value, 375, 376
    test statistics, 148–9, 149, 157
    unit-increase interpretation, 376–7
  Cook’s distance, 193–5, 195–6
  correlation coefficient, 779
  account length, 80, 81
  matrix plot, 79, 79
  Minotab regression tool, 80, 80
  optimal solution, 80
  p-values, 79–80, 80
  PCA, 94–5
  thresholds, 78

cost-benefit analysis, 462, 462–3
  CART model
    contingency table, 495, 495, 500, 501
    cost matrix, 500, 501
    estimated revenue increase, 503
    evaluation measures, 501–2, 502
    scaled cost matrix, 500, 501
    cost matrix, 474–6, 476
    decision invariance
      binary classifier, 471–2, 472
      scaling, 476–8, 477, 478
    direct cost, 478
  k-nary classification
    accuracy, 504
    contingency table, 503, 504
    Loans data sets, 504–7, 505–7
    overall error rate, 504
    predicted/actual categories, 503, 503
    sensitivity, 503, 504
  Loans data set
    adjusted cost matrix, 481, 481
    assumptions, 480
    CART model, 481, 481–2, 482
    direct cost matrix, 480, 480
    simplified cost matrix, 481, 481
    strategies, 479
    opportunity cost, 478
    positive classification
      adjusted cost matrix, 473, 473
      C5.0 models, 474, 475
      R code, 485–7, 507–8
      rebalancing cost
        CART model, 484
        confidence and positive confidence, 485
        definition, 483
        network models, 483, 484
    trinary classification
      accuracy, 494, 498
      assumptions, 498–9
      contingency table, 491, 492
      cost calculation, 499
      cost matrix, 500, 500
      false negative, 494, 497
      false positive, 494, 497
number of customers, 495–6
number of records, 491–2
overall error rate, 494, 498
predicted/actual categories, 491, 492, 494, 495
principal and interest, 498
true negative, 493–4, 497
true positive, 493–4, 497
cross-industry standard process for data mining (CRISP-DM)
adaptive process, 6–7, 7, 707–8, 708
business understanding phase, 709–10
business/research phase, 7
clustering analysis
BIRCH clustering algorithm, 742–4, 743, 744
cluster profiles, 745–8, 746, 748
cross-validation, 745, 747
k-means clustering, 744, 744–5, 746
data phase, 7–8
data preparation phase
deriving flag variable, 719, 719–21, 720, 721
negative amounts, 714–16, 715
product uniformity, 716–17, 716–18
standardization, 717–19
data understanding phase
absolute pairwise correlation, 727, 727–30
continuous predictors, 712, 712
dataset, fields, 710–711
de-transformation, 730–731
lifestyle cluster types, 713, 713–14
missing values, 712, 713
predictors and response, 722–7
zip code fields, 711
deployment phase, 8
evaluation phase, 8
modeling and evaluation strategy
baseline model, 754, 754–5, 755
cost-benefit analysis, 750, 751–3, 752
high performance model, 762–6
input variables, 753, 753
misclassification cost, 755–6, 756
model voting, 757–8, 758
processing steps, 750–751
profitable classification model, 758–61, 759
propensity averaging, 757–8, 758, 758
rebalanced data set, 756, 756–7, 757
modeling phase, 8
principal components analysis
data set partitioning, 732–3, 733
input variables, 733–4, 734
low communality predictors, 734, 734, 736
principal component profiles, 737–42
rotated component matrix, 737, 738, 739
cross-validation, 161–3
customer service calls (CSC) see polychotomous predictor
data balancing, 166
data cleaning
age field, 22
American zip code, 21
data set, 21, 21
income field, 21–2
marital status field, 22
measures of center
customer service calls, 28, 28
measures of location, 27–8
measures of spread, 29
price/earning ratio, 28–9, 29
standard deviation, 28, 29
missing data
data imputation method, 25
field values, 22–3, 23
frequency distribution, 25–6, 26
random values, 24–5, 25
replacement values, 23–4, 24
variable brand, 24–5
outliers, 26, 26–7, 27
poverty, 22
R code, 45–7
transaction amount field, 22
data imputation method, 25
data preparation
contingency table, 75, 75
HighDayEveMins_Flag variable, 76–7, 77
voice mail messages, 75
z-score standardization, 77–8, 78
data summarization
bivariate relationship, 777–80
boxplot, 777, 777
discrete variable, 769–70
levels of measurement, 769
measures of center, 774–5
data summarization (Continued)
measures of position, 776–7
measures of variability, 775–6
qualitative/quantitative variable, 769
data transformation
binning methods, 41–2, 42
categorical variables
reclassification, 42
region_num variable, 40
survey_response variable, 40–41
correlated variables, 44
decimal scaling, 32
categorical variables, 333
donation_dollar field, 43–4
duplicate records, 44–5
flag variables, 39–40
ID fields, 45
index field, 43
min–max normalization, 30, 30–31
R code, 47–50
unary variables, 43
Z-score standardization
inverse_sqrt (weight) transformation, 36–7, 37
natural log transformation, 35–6, 36
negative standardization, 31
normal probability plot, 36, 38
normal Z distribution, 32, 32–3
outliers, 38–9
positive standardization, 31–2
skewness, 33–5, 33–5
square root transformation, 35, 35
weighted data, 30, 31–3, 33
data visualization
bar chart, 771, 771
bivariate relationship, 777–80
cumulative frequency distribution, 772, 772
dot plot, 774, 774
frequency distribution, 771, 772, 772
histogram, 773, 773
pie chart, 771, 772
skewness, 774, 774, 775
stem-and-leaf display, 773, 773
data-driven misclassification costs see
cost-benefit analysis
decision tree
C4.5 algorithm, information-gain
adult data set, 333–4, 335
candidate splits, 326–7, 327
capital gains, 334–5
categorical variables, 333
decision node A, 330–331, 330–331
entropy reduction, 326–8
initial split, 329–30, 330
marital status, 333–4
numerical variables, 333
savings split, 328, 331, 332
threshold partition, 328–9, 329
training data set, 320, 326
CART (see Classification and regression
trees (CART))
credit risk, 317–18, 318
decision rules, 332, 332
diverse attributes, 318, 318–19
R code, 335–7
requirements, 319
dichotomous predictor
reference cell coding, 368–9
voice mail plan, 366–8, 367–8
dimension-reduction method
applications, 93
factor analysis (see factor analysis)
houses data set
median income, 9, 96–8
predictor variables, 96, 97
multicollinearity, 92
PCA (see principal components analysis
(PCA))
R code, 119–23
user-defined composites
definition, 118
houses data set, 118–19
measurement error, 118
summated scales, 117
direct cost matrix, 480, 480
distance function
age variable, 306–7
Euclidean distance, 305–6, 306
min–max normalization, 305–7, 306
properties, 305
Z-score standardization, 305–7, 306
EDA see exploratory data analysis (EDA)
ensemble methods
bagging model
algorithm for, 641
bootstrap samples, 641–3, 642
vs. CART model, 647–8, 648
prediction method, 642–3, 643
R code, 649
stable/unstable classification, 640, 640
bias-variance trade-off, 640
boosting model
adaptive boosting (see ADABost algorithm)
algorithm for, 644
vs. CART model, 647–8, 648
R code, 649
model voting
alternative models, 654–5, 655
contingency tables, 657, 659–60
evaluative measures, 657, 660
majority classification, 653–4, 654
processing steps, 655–6, 657
R code, 666
working test data set, 656, 658
prediction error, 639, 639, 649
propensity averaging
evaluative measures, 663, 664
histogram model, 663, 663
m base classifiers, 661
processing steps, 661, 662
exploratory data analysis (EDA), 451–2
churn data set (see churn data set)
data understanding phase
absolute pairwise correlation, 727, 727–30
de-transformation, 730–731
predictors and response, 722–7
vs. hypothesis testing, 54
R code, 82–8
segmentation modeling
capital gains/losses, 627, 627
contingency tables, 628, 628
overall error rate, 628–9
factor analysis model
adult data set
Bartlett’s test, 112
correlation matrix, 111–12, 112
factor loadings, 113, 113–14, 114
KMO statistics, 112–13, 113
principal axis, 112, 113
factor rotation
oblique rotation method, 117
orthogonal rotation, 113, 116–17
percentage of variance, 115–16, 116
rotated vectors, 114–15, 116
unrotated vectors, 114, 115
varimax rotation, 114, 115
flag variables, 39–40, 149–50, 157
GAs see genetic algorithms (GAs)
gas mileage prediction
backward elimination, 273, 274
best subsets method, 274–5, 275
forward selection method, 271–3, 272
Mallows’ $C_p$ statistics
predictors, 275–6, 276
regression assumptions, 276–8, 277–8, 277–8
stepwise selection regression, 272, 273
target variable MPG, 270–271, 270–271
generalized rule induction (GRI) method, 612–14
neural networks
backpropagation, 682–3
feed-forward nature, 681
learning method, 682
modified discrete crossover, 683–4, 684, 685
random shock mutation, 684, 685, 686
sum of squared errors, 682
topology and operation, 682
R code, 692
selection operator
Boltzmann selection, 677
crowding phenomenon, 676
definition, 672
elitism, 677
fitness sharing, 676
rank selection, 677–8
sigma scaling, 677
tournament ranking, 678
terminologies, 671
WEKA
AttributeSelectedClassifier, 689, 689–90, 690
class distribution, 684, 687
initial population characteristics, 690, 691
Preprocess tab, 684, 687
genetic algorithms (GAs) (Continued)

- WrapperSubsetEval evaluation method, 688, 689, 690
- gradient-descent method, 345, 345–6
- graphical evaluation
  - gains charts, 510–512, 512
  - lift chart, 510–511, 511, 511
  - profits charts, 512–14, 513–15
  - R code, 516–17
  - response charts, 511, 512
  - return-on-investment charts, 514, 516

hierarchical clustering

- agglomerative clustering, 525
- complete-linkage clustering, 527–9, 528
- divisive clustering methods, 525
- single-linkage clustering, 526–7, 527

hypothesis testing

- confidence interval, 141, 141–3, 142, 143
- criminal trial, outcomes, 138, 138–9
- null hypothesis, 140–141, 141
- p-value, 139, 139–40
- population proportion, 143, 143–4
- standard error, 139
- treatment, 139

indicator variable
cereals, y-intercepts, 252–3, 253
estimated nutritional rating, 253–4, 254
p-values, 255–6, 256
parallel planes, 251–2, 252
reference category, 250
regression coefficient values, 250, 250–251, 251
relative estimation error, 255, 255
shelf effect, 249, 249–50
instance-based learning
issues, 304
sodium/potassium ratio, 302–3, 303
training data points, 302–3, 303
voting, 304, 304
k-means clustering algorithm
data points, 530, 530
definition, 529
MSE, 529
processing steps, 530–533, 531, 531–3, 533
pseudo-F statistic method, 530
SAS Enterprise Miner (see churn data set)

k-nary classification
accuracy, 504
contingency table, 503, 504
Loans data sets, 504–7, 505–7
overall error rate, 504
predicted/actual categories, 503, 503
sensitivity, 503, 504

k-nearest neighbor (KNN) algorithm
classification
- data set, 301–2, 302
- income bracket, 301–2
ClassifyRisk data set, 312, 312
combination function
  - simple unweighted voting, 303, 304, 307–8
  - weighted voting, 303, 304, 308–9, 309
- cross-validation approach, 309–10
database, 303, 310
distance function
  - age variable, 306–7
  - Euclidean distance, 305–6, 306
  - min–max normalization, 305–7, 306
  - properties, 305
  - Z-score standardization, 305–7, 306
instance-based learning
issues, 304
sodium/potassium ratio, 302–3, 303
training data points, 302–3, 303
voting, 304, 304
locally weighted averaging, 310–311, 311
modeler’s results, 312, 312
outliers/unsual observations, 311–12
R code, 312–15

Kaiser–Meyer–Olkin (KMO) statistics, 112–13, 113

Kohonen networks
age and income data set, 545, 545–9, 549
algorithm, 544–5
CART decision tree model, 556, 556–7
cluster profiles, 554–6, 556
flag variables, 550
International Plan adopters, 551, 551
mean analysis, 552–3, 554
numerical variables, 550
R code, 557–8
SOM
architecture, 542, 543
characteristic processes, 544
goal, 542
networks connection, 542–3
topology, 548, 548
validation, 549
variables distribution, 552, 553
VoiceMail Plan adoption, 551–2, 552

logistic regression model
conditional mean, 359
disease vs. age, 358–9, 359, 359
linear regression model, 360
logit transformation, 360
maximum-likelihood estimation
confidence interval, 364
interpretation, 361–2
likelihood ratio test, 362, 362–3
log-likelihood estimators, 361
mean square regression, 362
negative response, 360–361
parameters, 360
positive response, 360–361
saturated model, 362
Wald test, parameters, 363–4
odds ratio (see odds ratio (OR))
R code, 403–8
sigmoidal curve, 359–60
training data set
education variable, 397–8
marital status, 394–7, 395, 396
WEKA
explorer panel, 400, 400
RATING field, 399
regression coefficients, 400–401, 401
test set prediction, 401–2, 402
training file, 398–9, 399

market basket analysis, 603–5, 605
Markov chain Monte Carlo (MCMC)
moments, 416
maximum a posteriori (MAP), churn data set
complement probabilities, 418, 419
conditional probability, 417
International Plan, 417, 420
joint conditional probabilities, 419, 419
marginal and conditional probabilities, 418, 418
posterior probabilities, 418, 418
Voice Mail Plan, 417, 420
McKinsey Global Institute (MGI) report
association task, 16–17

classification
income bracket, 12–13, 13
sodium/potassium ratio, 13–14, 14
clustering, 15, 15–16
continuous quality monitoring, 9
CRISP-DM
adaptive process, 6–7, 7
business/research phase, 7
data phase, 7–8
deployment phase, 8
evaluation phase, 8
modeling phase, 8
estimation model, 11–12, 12

Forbes magazine, 4
HMO, 4
patterns and trends, 10–11
prediction, 12
problem solving, human process, 6
profitable results, 10
R code, 17–18
software packages, 9
tools, 9

mean absolute error (MAE), 454
mean square error (MSE), 452, 453
mean square treatment (MSTR), 155

missing data imputation
CART model, 700, 700–701
data weighting, 701
flag variable, 701
multiple regression model, 696–9, 697
R code, 702–3
SEI formula, 699–700

model evaluation techniques
classification task
accuracy, 456
building and data model, 466
C5.0 model, 454
contingency table, 455, 455
cost/benefit analysis, 462, 462–3
error rate, 457
false negative, 459
false-negative rate, 458
false-positive, 459
false-positive rate, 458
financial lending firm, 455
gains chart, 463–6, 465
income classification, 466–7, 467
lift charts, 463–6, 464, 465
model evaluation techniques (Continued)
misclassification cost adjustment, 460–461, 461
true negative, 457–9
true positive, 457–9
description task, 451–2
estimation and prediction tasks
MAE, 454
MSE, 452, 453
standard error of the estimate, 452, 453
R code, 467–8
model voting process
alternative models, 654–5, 655
contingency tables, 657, 659–60
evaluative measures, 657, 660
majority classification, 653–4, 654
processing steps, 655–6, 657
R code, 666
working test data set, 656, 658
multicollinearity
correlation coefficients, 260–261, 261
fiber variable, 264–5, 265
matrix plot, 261, 261
potassium variable, 258–9, 264–5, 265
stability coefficient, 259–60, 260
user-defined composite, 263–4, 264
variable coefficients, 259–60, 260
variance inflation factor, 261–3, 263
multinomial data
chi-square test, 152–3, 153
expected frequency, 151, 151
observed frequency, 150, 150–151
R code, 158
test statistics, 151, 152
multiple regression model
ANOVA table, 240, 240–241
coefficient of determination, \( R^2 \), 241,
241–2, 253, 256, 256–7, 261
confidence interval
mean value, \( \mu \), 239, 248
particular coefficient, \( \beta_i \), 239, 247–8
estimation error, 239–40, 240
indicator variable
cereals, y-intercepts, 252–3, 253
estimated nutritional rating, 253–4, 254
\( p \)-values, 255–6, 256
parallel planes, 251–2, 252
reference category, 250
regression coefficient values, 250, 250–251, 251
relative estimation error, 255, 255
shelf effect, 249, 249–50
inference
\( F \)-test, 240, 245, 245–7
\( t \)-test, 239, 243–5
multicollinearity
correlation coefficients, 260–261, 261
fiber variable, 264–5, 265
matrix plot, 261, 261
potassium variable, 258–9, 264–5, 265
stability coefficient, 259–60, 260
user-defined composite, 263–4, 264
variable coefficients, 259–60, 260
variance inflation factor, 261–3, 263
nutritional rating vs. sugars, 237, 238, 239
population, 242–3
prediction interval, 239, 248
predictor variables, 237–8
principal components
Box–Cox transformation, 279, 282,
283–4
component values, 279–80, 280
unrotated and rotated component weights, 279, 282
varimax-rotated solution, 279, 281
R code, 284–92
regression plane/hyperplane, 238
slope coefficients, 238, 238
Spoon Size Shredded Wheat, 239
SSR, 257–8, 258, 259
three-dimensional scatter plot, 236, 236
variable selection method (see variable selection method)
Naïve Bayes classifier see also Bayesian approach
conditional independence, 424–6, 425
posterior odds ratio, 426–8
predictor variables, 429–32, 430
WEKA
ARFF, 432, 433
conditional probabilities, 434, 435
Explorer Panel, 432, 434
load training file, 432
test set predictions, 434–6, 435
zero-frequency cells, 428
neural network model
adult data set, 352, 353
artificial neuron model, 338, 339
back-propagation algorithm
cross validation termination, 348–9
downstream node, 347–8
error propagation, 346–7
learning rate, 349–50, 350
momentum term, 350–352, 351
squared prediction error, 344
upstream node, 347–8
combination function, 342–3
data preprocessing, 354–5
estimation and prediction, 341
hidden layer, 341–2, 342
input and output encoding
categorical variables, 339–40
dichotomous classification, 340, 340–341
drawback, 338–9
min–max normalization, 339
thresholds, 340
input layer, 342
output layer, 342, 342
prediction accuracy, 353–4, 354
R code, 355–6
real neuron, 338, 339
sensitivity analysis, 352, 354
sigmoid function, 343–4, 344
neural networks
backpropagation, 682–3
feed-forward nature, 681
learning method, 682
modified discrete crossover, 683–4, 684, 685
random shock mutation, 684, 685, 686
sum of squared errors, 682
topology and operation, 682
odds ratio (OR)
assumptions
capnet variable, 379–81, 380
churn overlay, 377–9, 378, 378–9
customer service calls, 377–9, 378, 378–9
continuous predictor (see continuous predictor)
dichotomous predictor (see dichotomous predictor)
estrogen replacement therapy, 365
interpretation, 364–5
polychotomous predictor (see polychotomous predictor)
relative risk, 365
response variable, 365
zero-count cell, 381, 381–3
overfitting
complexity model, 163, 163–4
provisional model, 163, 163
partitioning variable, 166–7, 169
PCA see Principal components analysis (PCA)
polychotomous predictor
certainty interval, 370, 372–3
estimated probability, 369, 370–371
medium customer service call, 369, 370–371
reference cell encoding, 369, 369–70
standard error, 372
Wald test, 371–2
principal components analysis (PCA)
communality, 101, 103, 108–10
component matrix, 98–9, 101
component size, 106
component weights, 107–8, 108
coordinate system, 94
correlation coefficient, 94–5
correlation matrix, 98, 100
covariance matrix, 94
data set partitioning, 732–3, 733
eigenvalues, 95–6, 102
eigenvectors, 95–6
geographical component, 99, 100, 106
housing median age, 105, 106–7, 107
input variables, 733–4, 734
linear combination, 93–4
low communality predictors, 734, 734, 736
matrix plot, 98, 99
median income, 105, 106–7, 107
multiple regression analysis, 98
orthogonal vectors, 101–2
principal component profiles, 737–42
rotated component matrix, 737, 738, 739
scree plot, 103–5, 104, 105
standard deviation matrix, 94
validation, 110, 110
variance proportion, 99, 101, 101, 103
profits charts, 512–14, 513–15
propensity averaging process
evaluative measures, 663, 664
histogram model, 663, 663
propensity averaging process (Continued)
m base classifiers, 661
processing steps, 661, 662
pseudo-\( F \) statistic method
clustering model, 591
distribution, 592
Iris data set, 592–3
R code, 598
SSB and SSE, 590
regression modeling
ANOVA table, 186, 186, 187
baseline model, 168
Box–Cox transformation, 220
cereals data set, 171–2, 172
coefficient of determination, \( r^2 \)
data points, 179–80, 180
distance and time estimation, 179, 179
estimation error, 180, 181
maximum value, 182
minimum value, 182
predicted score column, 179, 179
prediction error, 178–9
predictor and response variables, 182
predictor information, 179
residual error, 178–9
sample variance, 180
standard deviation, 180
sum of squares regression, 181–2
sum of squares total, 180–181, 181
Cook’s distance, 193–5, 195–6
correlation coefficient, \( r \)
confidence interval, 208, 208–10, 209
linear correlation, 186
negative correlation, 185
positive correlation, 185
quantitative variables, 184

dangers of extrapolation
chocolate frosted sugar bombs, 177
observed and unobserved points, 177
policy recommendations, 178
prediction error, 177–8
predictor variable, 177
end-user
confidence interval, 210
prediction interval, 211–13, 214
field values, 172
high leverage point
characteristics, 191
distance vs. time, 189, 190

hard-core orienteer, 189–91, 191
mild outlier, 192–3, 193
observation, 192–3
regression results, 191–2, 192
standard error, 192, 193
inference, 203–4
least-squares estimation
error term \( \varepsilon \), 174
estimated nutritional rating, 176–7
nutritional rating vs. sugar content, 172–3
prediction error, 173
statistics, 175–6, 176
sum of squared errors, 174–5
\( y \)-intercept \( b_0 \), 176–7
linearity transformation
bulging rule, 215, 216, 217
log transformation, 218, 218
point value vs. letter frequency, 215, 216, 217
response variable, 213–14
Scrabble\textsuperscript{®}, 214–15, 215
square root transformation, 217, 217
standardized residual, 218, 218, 219, 219
normal probability plot
Anderson–Darling (AD) statistics, 199–200, 200
assumptions, 201–3, 202
chi-square distribution, 198–9, 199
distance vs. time, 200, 216
horizontal zero line, 200–201, 201
normal distribution, 198, 200, 200
\( p \)-value, 199–200, 200
Rorschach effect, 202
uniform distribution, 198–9, 199
outliers
Minitab, 187, 188
nutritional rating vs. sugars, 186, 188
positive and negative values, 188–9
standardized residuals, 188
population regression equation
assumptions, 197
bivariate observation, 195–6
constant variance, 197, 197–8
true regression line, 197, 197–8
\( R \) code, 220–226
regression equation, 172–3
standard error
mean square error, 183
standard deviation, response variable, 179, 183–4
sum of squares regression, 184
sum of squares total, 184
time and distance calculation, 184
t-test
assumptions, 204, 204
confidence interval, 206–7, 207
null hypothesis, 205
nutritional rating vs. sugar content, 205–6, 206
p-value method, 205–6
sampling distribution, 205
response charts, 511, 512
return-on-investment (ROI) charts, 514, 516
scatter plot, 710, 779
segmentation modeling
clustering analysis
CART decision trees, 632, 633
customer service call, 135–6, 135–6
churn proportion, 630, 632, 632
contingency tables, 630, 632
customer service call, 135–6, 135–6
class proportion, 630, 632–3
contingency tables, 630, 632–3
class proportion, 630, 633
customer service call, 135–6, 135–6
churn proportion, 630, 632–3
class proportion, 630, 633
customer service call, 135–6, 135–6
churn proportion, 630, 632–3
exploratory analysis
capital gains/losses, 627, 627
capital gains/losses, 627, 627
contingency tables, 628, 628
overall error rate, 628–9
performance enhancement, 625
processing steps, 626, 626
target variable, 166–7, 169
unsupervised methods, 160–161
user-defined composites
definition, 118
definition, 118
houses data set, 118–19
measurement error, 118
measurement error, 118
summed scales, 117
variable selection method
all-possible-regression, 269–70
backward elimination, 268
class proportion, 630, 632–3
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
class proportion, 630, 633
backward elimination, 268
class proportion, 630, 633
variable selection method
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
class proportion, 630, 633
variable selection method
all-possible-regression, 269–70
backward elimination, 268
best subsets method, 269
forward selection, 268
gas mileage data set (see gas mileage prediction)
partial F-test, 266–7, 267
stepwise regression, 268–9
Waikato Environment for Knowledge Analysis (WEKA)
Bayesian belief networks
Explorer Panel, 441, 442
prior probabilities, 442, 443
test set predictions, 442–4, 443
Waikato Environment for Knowledge Analysis (WEKA) (Continued)
explorer panel, 400, 400
genetic search algorithm
AttributeSelectedClassifier, 689, 689–90, 690
class distribution, 684, 687
initial population characteristics, 690, 691
Preprocess tab, 684, 687
WrapperSubsetEval, 688, 689, 690

Naïve Bayes
ARFF, 432, 433
conditional probabilities, 434, 435
Explorer Panel, 432, 434
load training file, 432
test set predictions, 434–6, 435
RATING field, 399
regression coefficients, 400–401, 401
test set prediction, 401–2, 402
training file, 398–9, 399