Index

0-1 loss, see loss function, 0-1

.632 bootstrap, 225–226, 306

absolute loss, see loss function, absolute
accuracy, 191, 196
AdaBoost, 435–437
    instance weighting for, 435
    model weighting for, 435
    multiclass, 438–439
AdaBoost.MH, 439
adaptive boosting, see AdaBoost
add-one smoothing, see probability,
    Laplace estimate of
agglomerative clustering, see hierarchical
    clustering, agglomerative aggregation, 506
AHC, see hierarchical clustering,
    agglomerative algorithm randomization, 412–413
        anomaly detection, 163
        by clustering, 17, 352
ANOVA, see F-test
AODE, see averaged one-dependence estimators
apparent disutility, 570
apparent utility, 570
attribute, xxii, 5
    continuous, 6, 25
        discretization of, see discretization
discrete, 36
        aggregation of, see aggregation
        binary encoding of, see attribute encoding
hidden, 16
input, see input attribute
nominal, 5
numeric, see attribute, continuous
observable, 16
ordinal, 6
target, see target attribute
attribute encoding, 154, 250, 508–509
attribute sampling, 410
attribute selection
    motivation for, 559–560
    target algorithm for, 560
    target task for, 560
attribute selection filter, 561, 568
    consistency-based, see correlation-based filter
    correlation-based, see correlation-based filter
random forest, see random forest, for
    attribute selection
    simple statistical, see simple statistical filter
attribute selection search, 562, 563
    backward, see backward elimination
    filter-driven, 586–587
    forward, see forward selection
greedy, 566
attribute selection search (continued)
  initial state for, 563
  operator for, 564
  preference criteria for, 565
  stop criteria for, 565
attribute selection wrapper, 561, 588, 591
  search strategy for, 591
  subset evaluation for, 588–589
attribute transformation
  as modeling, see attribute transformation, modeling
  for ensemble modeling, 411–412
  model, see transformation model modeling, 500–501
  nonmodeling, 503
  target attribute for, see target attribute, for attribute transformation
  target task for, 499
  training set for, see training set, for attribute transformation
attribute value vector, 137, 239, 240, 455
AUC, see ROC, curve, area under
average linkage, see linkage, average averaged one-dependence estimators, 131
backward elimination, 564, 566
bagging, 431–432
  base models for, 431
  model aggregation for, 431
  probabilistic, 171
barplot, 64
base models, 405, 406
  aggregation of, see model aggregation
  combining of, see model aggregation creation of
    by algorithm randomization, see algorithm randomization
    by attribute sampling, see attribute sampling
    by attribute transformation, see attribute transformation, for ensemble modeling
    by different algorithms, 412
  by different parameter setups, 412
  by instance replication, see instance replication
  by instance sampling, see instance sampling, for ensemble modeling
  by instance weighting, see instance weighting, for ensemble modeling
diversity of, 406, 418
basic statistics
  as modeling, 24–25
  for distribution description, 23, 25
  for relationship detection, 23, 47, 568
Bayes rule, 118–119
  for conditional class probability, 121
  inverse conditional probability for, 119
  posterior probability for, 119
  prior probability for, 119
Bayesian inference, 118, 120
  for class probabilities, 120
  for model probabilities, 120
Bayesian network, 130
Bernoulli trial, 42
boosting, 433–434
  adaptive, see AdaBoost
  base models for, 433–434
  gradient, see gradient boosting
  by instance weighting, 433, 435
  by residual prediction, 434, 439
bootstrap sample, 43, 305, 406
bootstrapping, 171, 305
  for model evaluation, 223–226, 306
bottom-up discretization, see discretization, bottom-up
boxplot, 62
break evaluation
  by class distribution dissimilarity, 544–545
  by class impurity, 542–544
  by dominating class count, 539–540
  by instance count, 537–538
  by misclassification count, 541–542
INDEX 669

C index, 388
Canberra dissimilarity, see distance, Canberra
center linkage, see linkage, center
centroid, 334
Cestnik estimate, see probability, 
m-estimate of
CFS, see correlation-based filter
Chebyshev dissimilarity, see distance, Chebyshev
\( \chi^2 \) discretization, see break evaluation, by
class distribution dissimilarity
\( \chi^2 \) statistic, see \( \chi^2 \) test, statistic for
\( \chi^2 \) test, 52
\( p \)-value for, 52
statistic for, 52
for discretization breaks, 544
ChiMerge, 545
class, 10
negative, 195
positive, 195
class encoding, see multiclass encoding
class imbalance, 162
class label, see class
classification algorithm, 13
cost-sensitive, 13, 164
weight-sensitive, 13, 164
classification margin
functional, 457
geometric, 457
soft, 469
hyperplane, see margin hyperplane
instance lying on, 458, 461
instance lying outside, 461
instance lying within, 468
maximization of, 460
dual form of, 464–466
primal form of, 460–461
soft, 468–470
classification model, 11
discrete, 12
mixing of, see operating point, interpolation
probabilistic, 12
scoring, 11–12, 200
classification performance, see predictive performance, for classification
classification performance measures, 190
classification task, xxii, 9–10
binary, see classification task, two-class
two-class, 10, 137, 162, 165, 170
classifier, see classification model
Cluster center, 328
Cluster formation, 16
for \( k \)-centers clustering, 331–332
Cluster modeling, 16
for \( k \)-centers clustering, 332
explicit, 343–344
implicit, 343
Clustering algorithm, 19
conceptual, 19
dissimilarity-based, 19
probabilistic, 19
similarity-based, see clustering algorithm, dissimilarity-based
Clustering model, 18
crisp, 18
descriptive, 19
flat, 18
fuzzy, see clustering model, soft
hierarchical, 18
predictive, 19, 329, 369
soft, 18
Clustering performance, see predictive performance, for clustering
Clustering quality
measures of, see clustering quality measures
Clustering quality measures, 373, 385
external, 393
per-cluster, 376
relationship detection-based, 396
using of, 397–398
Clustering task, xxii, 16
hierarchical, 18
motivation for, 16–17
coefficient of determination, 300
weighted, 301
coefficient of variation, 33
Complete linkage, see linkage, Complete
Complexity parameter, 99, 277
concept, 10
  example of, 10
  multiclass, 10
  negative example of, 10
  single, 10, 195
confidence interval, 41
confidence level, 41
confusion matrix, 194
  multiclass, 198
    1-vs-1 analysis of, 198
    1-vs-rest analysis of, 198
  two-class, 196
    weighted, 199
consistency-based filter, 575–576
constraints, 509
correlation
  as a regression performance measure, 301
  as a similarity measure, see similarity measure, correlation-based linear, 50
    p-value for, 50
  Pearson’s, see correlation, linear rank, 51
    p-value for, 51
  Spearman’s, see correlation, rank correlation-based filter, 571
  relationship measure for, 571–572
  subset evaluation for, 572–573
cost-complexity pruning
  for decision trees, 99–100
  for regression trees, 277
  minimum-error, 100
  one-standard-deviation, 100
cost-sensitive classification, 159
  algorithm for, see classification algorithm, cost-sensitive
  by instance relabeling, see instance relabeling
  by instance resampling, see instance resampling, for cost-sensitive classification
  by instance weighting, see instance weighting, for cost-sensitive classification
  by the minimum-cost rule, see minimum-cost rule
  wrappers for, 164
cross-validation, 219–220, 277, 304
  for attribute selection wrappers, 589
cutoff
  for a scoring classifier, 200
  for attribute selection, 585
  for cost-sensitive classification, 171
  for operating point shifting, 205
  for ROC analysis, 201
data, see dataset
data mining, xxi, xxii
  modeling view of, xxi, 3
  roots of, xxiii
  tasks, xxii
data preprocessing, 499, 532, 560
dataset, xxii, 5, 7
  Boston Housing, 237, 296, 403, 455, 558
  Census Income, 603
  Communities and Crime, 603
  Cover Type, 603
  Glass, 329, 350, 498, 524
  HouseVotes84, 403
  incomplete, 20, 507
  Iris, 329, 350, 374
  linearly inseparable, see linear separability
  linearly separable, see linear separability
  noisy, 20
  Pima Indians Diabetes, 135, 455
  Soybean, 190, 558
  Vehicle Silhouettes, 159, 498, 524, 558
  weather, 10–11
  weather.c, 11
  weathercl, 17
  weatherr, 15
Davies–Bouldin index, 382
  average, 387
  for a cluster, 383
  for a cluster pair, 382
decision boundary, 138, 139
  distance to, see separating hyperplane, distance to
decision stump, 437
decision tree
as a classification model, 74
as a nonparametric model, 136
as a probabilistic classifier, see
decision tree, prediction for,
probabilistic
branch of, 72
class distribution calculation for, 77, 78
class label assignment for, 77, 79
growing of, see decision tree growing
leaf of, 72
subset corresponding to, 74
node of, 72
closed, 77
open, 77
subset corresponding to, 73, 74
post-pruning of, see decision tree pruning
prediction for, 104
probabilistic, 105
with fractional instances, see
fractional instances,
decision tree prediction with
prepruning of, 90
pruning randomized, 412
split application for, 77, 86
with fractional instances, see
fractional instances,
split application with,
for decision trees
split evaluation for, 82–83
with fractional instances, see
fractional instances,
split evaluation with,
for decision trees
split for, 72
binary, 76
equality-based, 75
inequality-based, 75
interval-based, 75
membership-based, 75
multivariate, 74
subset-based, 75
univariate, 74
value-based, 74
split selection for, 77, 82
for continuous attributes, 86
by impurity minimization, 83
with two classes, 85
stop criteria for, 77
relaxed, 81
strict, 80–81
structure of, 72
test for, see decision tree split for
top-down induction of, 76
with missing values, see missing value handling, for decision trees
with weighted instances, 105
decision tree growing, 76
algorithm scheme for, 76–77
algorithm steps for, 77
recursive, 76
with fractional instances, see
fractional instances,
decision tree growing with
decision tree pruning, 90
best-first, 101
bottom-up, 101
by node removal, 91
by subtree cutoff, 91
control strategy for, 90, 100
cost-complexity, see cost-complexity pruning, for decision trees
criterion for, 90
minimum error, see minimum error pruning, for decision trees
operators for, 90, 91
pessimistic, see pessimistic pruning, for decision trees
reduced error, see reduced error pruning, for decision trees
top-down, 101
delta rule, 145, 243–244
batch, 244
for generalized linear
representation, 253
for linear logit classification, 148
for linear threshold classification, 146, 152
delta rule (continued)
  incremental, 243
  linear, 244
dendrogram, 364
descriptive statistics, see basic statistics, for distribution description
diameter, 376
discretization, 525
  as a modeling transformation, 527–528
  as a nonmodeling transformation, 532
  bin for, see discretization, interval for bottom-up, 535
  algorithm scheme for, 535
  initialization for, 536
  merge criterion for, 537
  stop criteria for, 545–546
  break evaluation for, see break evaluation
  break for, 527
  equal-frequency, 531–532
  equal-width, 530
  interval for, 527
  model for, 528
  motivation for, 525–526
  pure-class, 533–534
  requirements for, 529
  supervised, 526, 533
  target attribute for, see target attribute, for discretization
  target task for, 526
  top-down, 546
    algorithm scheme for, 546–547
    cut criterion for, 549–550
    initialization for, 548–549
    recursive, 547
    stop criteria for, 550–551
    training set for, see training set, for discretization
  unsupervised, 526, 530
dispersion, 31
dissimilarity measure
  Canberra, see distance, Canberra
  Chebyshev, see distance, Chebyshev
correlation based, see similarity measure, correlation-based
difference-based, 314
  weighted, 320
  with discrete attributes, 315
  Euclidean, see distance, Euclidean
  Gower, see Gower’s coefficient
  Hamming, see distance, Hamming
  Manhattan, see distance, Manhattan
  Minkowski, see distance, Minkowski
distance
  Canberra, 316
    weighted, 320
  Chebyshev, 317
    weighted, 320
  chessboard, see distance, Chebyshev
  Euclidean, 314
    weighted, 320
  Hamming, 318
    weighted, 320
  Manhattan, 316
    weighted, 320
  maximum, see distance, Chebyshev
  Minkowski, 315
    weighted, 320
distribution
  binomial, 42
  $\chi^2$, 52
  $F$, 58
  normal, 34, 42
  $t$, 56
divisive clustering, see hierarchical clustering, divisive
domain, xxi, 5
  decomposition of
    by clustering, 17, 352
    by a decision tree, 72
    by a regression tree, 262
dot product, 137, 241, 455
  for SVM, 466
  for SVR, 477
dummy coding, 509
dummy variable, 509
  Dunn index, 386
ensemble modeling
  base models for, see base models
  justification of, 404–405
model aggregation for, see model aggregation
predictive performance of, 448
entropy, 39
as a consistency measure, 575
conditional, 54
for decision tree splits, 83
for discretization breaks, 543
entropy-based discretization, see break evaluation, by class impurity
equal-frequency discretization, see
discretization, equal-frequency
equal-width discretization, see
discretization, equal-width
error-correcting code, 515, 516
estimator, 24, 25, 41
Euclidean dissimilarity, see distance, Euclidean
evaluation overfitting, 213, 216
evaluation procedure, 190, 213
choosing of, 227, 307
for regression models, 303
exponential loss, see loss function, exponential
false negative, 196
false positive, 196
rate, 196, 197
$F$-measure, 198
forward selection, 564, 566
fractional instances, 106, 279
decision tree growing with, 107
decision tree prediction with, 111–112
for decision trees, 106–107
for regression trees, 279
regression tree growing with, 279
regression tree prediction with, 283
split application with
for decision trees, 107
for regression trees, 280
split evaluation with
for decision trees, 107
for regression trees, 279
fraud detection, 163

$F$-test, 57
$p$-value for, 58
statistic for, 57
$G$ test, see loglikelihood ratio test
generalization, 8
for classification, 13
for clustering, 19
for regression, 15
generalized linear model, 253
Gini index, 40
GLM, see generalized linear model
Gower's coefficient, 319
gradient, 243
gradient ascent, 148, 149
gradient boosting, 439–441
model weighting for, 440, 441
with regression trees, 441
pseudoresidual for, 440
shrinkage for, 440
stochastic, 439
gradient descent, 145, 242, 245
batch, 245
for linear classification, 149
incremental, 245
with randomized instance order, 246
online, see gradient descent, incremental
stochastic, see gradient descent, incremental
stop criteria for, 246
Hamming dissimilarity, see distance, Hamming
heuristic function, 562
hierarchical clustering
agglomerative, 349, 353
algorithm scheme for, 353–355
cluster merging for, 354
cutting of, 366
dissimilarity measures for, 349
divisive, 349, 361
algorithm scheme for, 361
cluster partitioning for, 361
stop criteria for, 362
hierarchical clustering (continued)
leaf of, 351
model representation for, 351–352
motivation for, 351–352
node of, 351
prediction for, 366, 368
visualization of, see dendrogram

histogram, 63
hold-out, 217, 304

impurity, 39, 82, 575
imputation, 507
with means, 507
with medians, 507
with modes, 507
inconsistency rate, 575
inductive bias, 8
preference, 9
representation, 9
inductive learning, 3, 5
algorithm for, see modeling algorithm
as search, 9
inference, 4
deductive, 4
inductive, 3, 4
information gain, 83
input attribute, xxii, 6
instance, xxii, 5
labeled, 10
negative, 195
positive, 195
instance relabeling, 174
instance replication, 167, 408
instance resampling, 162
for cost-sensitive
classification, 167–168
for ensemble modeling, see instance sampling, for ensemble modeling
instance sampling
bootstrap, see bootstrap sample
for ensemble modeling, 406–407
instance undersampling, 167
instance weighting, 8, 191, 199, 301
for boosting, see boosting, by instance weighting
for classification, 13

for cost-sensitive
classification, 164–166
for decision trees, see decision tree,
with weighted instances
for ensemble modeling, 408
per-class, 13
intercept, 137, 240, 455
for SVM, 465
soft-margin, 469
for SVR, 477
interquartile range, 35
interval estimation, 41
bootsrapping, 43
parametric, 41–42
for binomial distribution, 42
isolated cluster, see isolation
isolation, 378

k-centers clustering
adaptive, 343
algorithm scheme for, 330
center adjustment for, 332
choice of k for, 343, 397
convergence of, 331
dissimilarity measures for, 329
initialization for, 331–332
instance assignment for, 331, 332
multiple runs of, 343
operation principle of, 328
prediction for, 332
stop criteria for, 331–332
k-means, 334
center adjustment for, 335
convergence of, 335
dissimilarity measures for, 335
dissimilarity minimization by, 338
with discrete attributes, 335
with the Euclidean dissimilarity, 336
k-medians, 338
dissimilarity measures for, 338
dissimilarity minimization by, 338
k-medoids, 340
center adjustment for, 340
dissimilarity measures for, 340
dissimilarity minimization by, 340
INDEX 675

kernel function, 483
  Gaussian, see kernel function, radial
  linear, 485
  polynomial, 485
  prediction using, 487
  radial, 485
  RBF, see kernel function, radial
  sigmoid, 486
kernel matrix, 483
kernel trick, 482–483
  advantages of, 484
  for SVM, 489
  for SVR, 492
knowledge, 3, 4
  in inference, 4
  representation of, 5
Kruskal–Wallis test, 60
  p-value for, 61
  statistic for, 60
labeling function, 12
Lagrange multiplier, 464
Laplace smoothing, see probability,
  Laplace estimate of
least squares, 153, 248–249
  for generalized linear
    representation, 254
  for linear threshold classification, 153
leave-one-out, 221–222, 305
likelihood
  for classification, 147, 211
    two-class, 211
linear classification
  by boundary modeling, 138
  by probability modeling, 138
  logit
    model representation for, 142, 253
    parameter estimation
      for, 147–149
    vs. the naïve Bayes classifier, 127
  prediction for, 136, 138
  threshold
    least squares for, see least
      squares, for linear
      threshold classification
model representation for, 139, 253
  parameter estimation
    for, 146–147
  with discrete attributes, see
    parametric classification,
    with discrete attributes, 250
linear regression
  model representation for, see linear
    representation
  prediction for, 240
  with discrete attributes, see
    parametric regression, with
    discrete attributes
linear representation, 136, 240
  advantages of, 251–252
  enhanced, 241, 255, 454
  generalized, 241, 252
  logit, see linear classification, logit,
    model representation for
  piecewise, 257–258
  piecewise-linear, 285
  polynomial, 256
  randomized, 255
  threshold, see linear classification,
    threshold, model
    representation for
linear separability, 139, 152, 460, 461, 468
link function, 252
  inverse, 252
linkage, 354, 356
  average, 358
  center, 359
  choosing, 360
  complete, 357
  monotonicity of, 359, 361, 364
  single, 357
LMS rule, see delta rule, linear
location, 25
log-loss, see loss function, logarithmic
logarithmic loss, see loss function,
  logarithmic
logistic function, see logit, inverse
logistic regression, see linear classification,
  logit
  logit, 142
    inverse, 142
INDEX

loglikelihood
- for classification, 148, 212, 391
  - two-class, 212
- for clustering, 390–392

loglikelihood ratio test, 53
- \( p \)-value for, 53
- statistic for, 53

loss function, 191, 296, 302
- \( 0-1 \), 191
- absolute, 297, 302
- asymmetric, 303
- \( \epsilon \)-insensitive, 303
- exponential, 437
- logarithmic, 148, 212
- quadratic, 242, 298, 302

machine learning, xxiii, 3

MAE, see mean absolute error

Manhattan dissimilarity, see distance, Manhattan

Mann-Whitney test, see
- Mann-Whitney–Wilcoxon test

Mann-Whitney–Wilcoxon test, 58
- \( p \)-value for, 60
- statistic for, 59

margin hyperplane, 458
- instance lying on, see classification
- margin, instance lying on instance lying outside, see
- classification margin, instance lying outside
- instance lying within, see
- classification margin, instance lying within

maximum weighted spanning tree, 130

maximum-probability rule, 12, 142
- mean, 25
  - \( m \)-estimated, see mean, \( m \)-estimate of
  - \( m \)-estimate of, 46, 276
- weighted, 26

mean absolute error, 297
- weighted, 301

mean misclassification cost, 165, 192, 193
- weighted, 301

mean square error, 242, 297
- weighted, 301

median, 26
- weighted, 27

median absolute deviation, 34

medoid, 339

\( m \)-estimation, 43, 276
- of mean, see mean, \( m \)-estimate of
- of probability, see probability, \( m \)-estimate of
- of variance, see variance, \( m \)-estimate of
- priors for, 47

miclassification error
- for clustering, 393

minimum error pruning
- for decision trees, 96–97
- for regression trees, 276

minimum-cost rule, 169–170
- two-class, 170–171

Minkowski dissimilarity, see distance, Minkowski

misclassification costs, 192
- expected, 170
- experimental procedure for, 180
- function, 164
- incorporation of, see cost-sensitive classification
- instance-specific, 163, 165, 170
- matrix, 161, 192
- objective, 161
- per-class, see misclassification costs, vector
- per-instance, see misclassification costs, instance-specific
- subjective, 161
- vector, 162, 165, 192

misclassification error, 12, 191, 196
- weighted, 164, 192

missing value handling
- by imputation, see imputation
- for decision trees, 106
- using fractional instances, see
- fractional instances, for decision trees
- using surrogate splits, see
- surrogate splits, for decision trees
for dissimilarity measures, 324
for regression trees, 279
  using fractional instances, see fractional instances, for regression trees
  using surrogate splits, see surrogate splits, for regression trees
  for similarity measures, see missing value handling, for dissimilarity measures
  for the naïve Bayes classifier, 128–129
modal value, see mode
model, xxi, xxii, 7
  weighted, 37
model aggregation, 420
  by averaging, 420–421
  weighted, 424–425
  by probability averaging, 422–423
  by using as attributes, 427–428
  by voting, 420–421
  weighted, 424–425
model ensemble, see ensemble modeling
model evaluation
  bias of, 214, 303
  bias vs. variance, 214, 228, 307
  final, 190, 215
  for temporal data, 230–231
  intermediate, 190, 215
  intermediate vs. final, 214
  procedure for, see evaluation procedure
  variance of, 214, 304
model parameter vector, 137, 239, 240, 455
  canonical form of, 461
  for SVM, 465
  for SVR, 476
  vector of, see model parameter vector
model parameters, 136, 239
  estimation of, see parameter estimation
  vector of, see model parameter vector
model selection, 191
model tree, 285
  growing of, see model tree growing
  linear models for, 285
  attribute preselection for, 285
  prediction for, 290–291
  with smoothing, 290
  pruning of, see model tree pruning
  split selection for, 286–287
  stop criteria for, 286
  target function dispersion for, 286
  model tree growing, 285
  model tree pruning, 289–290
modeling algorithm, 8
  randomization of, see algorithm randomization
  stable, 407
  unstable, 406, 431
  weight-sensitive, 8
modeling procedure, 214, 303
MSE, see mean square error
multiclass decomposition, see multiclass encoding
multiclass encoding, 439, 511
  1-of-k, 514
  as ensemble modeling, 512
  binary models for, 511
  codeword for, 511
  decoding function for, 511
  encoding function for, 511
  inverse, 511
  error-correcting, 515–518
multiconcept, see concept, multiclass mutual information, 54
  conditional, 130
  vs. conditional entropy, 54
  vs. loglikelihood ratio, 54
m-variance, see variance, m-estimate of naïve Bayes classifier
  as a linear classifier, 127
  as a parametric model, 136
  augmented, 130
  conditional attribute value probability for, 122
  conditional joint probability for, 122
  independence assumption for, 122
  logarithmic form of, 126, 127
  model representation for, 123
INDEX

naïve Bayes classifier (continued)
  naïvety of, see naïve Bayes classifier,
  independence assumption
  for
  not-so-naïve, 129
  prediction for, 124
  prior class probability for, 121
  small probabilities for, 126
  tree-augmented, 130
  with continuous attributes, 127–128
  with missing values, see missing value handling, for the naïve Bayes classifier
  zero probabilities for, 125
nonlinear regression, see nonlinear representation
nonlinear representation, 241
nonparametric representation, 136, 239
normalization, 505
  as a modeling transformation, 505
  for dissimilarity calculation, 321
  for k-centers clustering, 329
observation, xxiii
  Ockham’s razor 9
    for regression trees, 268, 277
    for decision trees, 81, 82, 90
OLS, see least squares
OOB, see out-of-bag
operating point, 12, 200, 201
  default, 12, 204
  interpolation of, 206
  shifting of, 205
order statistic, 29
ordinary least squares, see least squares
out-of-bag, 224, 306, 445
outlier, 35, 63
overfitting, 8
  for classification, 13
  for clustering, 374
  for gradient boosting, 440
  for model evaluation, see evaluation overfitting
  for regression, 15
  for regression trees, 268, 285
  resistance to
    for AdaBoost, 437
for bagging, 431
for random forest, 443
for SVM, 460, 470
for SVR, 474
for the naïve Bayes classifier
oversearching, 9
PAM, see partitioning around medoids
  parameter estimation, 145, 242
    by gradient descent, see gradient descent
    by least squares, see least squares
    delta rule for, see delta rule
    for linear threshold classification, see linear classification, threshold, parameter estimation
    for linear threshold logit, see linear classification, logit, parameter estimation
    for SVM, see classification margin, maximization of
    for SVR, see regression flatness, maximization of
parametric classification, 137
  logit, 142
  threshold, 139
  with discrete attributes, 154
parametric regression
  model representation for, see parametric representation
  prediction for, 239
parametric representation, 136, 239
partitioning around medoids, 341
perceptron, 147
performance, see predictive performance
pessimistic pruning
  for decision trees, 95–96
piecewise-constant regression, 262
piecewise-linear regression, see linear representation, piecewise polynomial regression, see linear representation, polynomial
population, xxiii
precision, 196, 197
prediction, xxii, 7
for decision trees, see decision tree, prediction for
for ensemble modeling, see model aggregation
for hierarchical clustering, see hierarchical clustering, prediction for
for k-centers clustering, see k-centers clustering, prediction for
for linear classification, see linear classification, prediction for
for linear regression, see linear regression, prediction for
for parametric regression, see parametric regression, prediction for
for regression trees, see regression tree, prediction for
for SVM, see SVM, prediction for
for SVR, see SVR, prediction for
for the naïve Bayes classifier, see naïve Bayes classifier, prediction for
of class probabilities, 171
predictive performance, 7, 189, 295
dataset, 189, 295
evaluation of, see model evaluation
for classification, 12
measures of, see classification performance measures
for clustering, 18
measures of, see clustering quality measures
for regression, 15
measures of, see regression performance measures
training, 7, 189, 295, 374
true, 7, 189, 295
probability, 37
conditional, 37
Laplace estimate of, 45, 96, 126
m-estimated, see probability, m-estimate of
m-estimate of, 44, 95, 96, 125
weighted, 38
pure-class discretization, see discretization, pure-class
quadratic loss, see loss function, quadratic
quadratic programming, 461, 464, 475, 476
quantile, 29
R type 3, 30
R type 7, 30
R type 8, 30
R type 9, 30
quartile, 30
quartile dispersion coefficient, 35
R package
cluster, 350, 362, 374–376, 640
datasets, 350
datasets, 329
digest, 576
e1071, 159, 403, 447, 498, 524
ipred, 159, 175
kernlab, 455, 461
lattice, 72, 135, 455
Matrix, 461
mlbench, 135, 159, 190, 237, 296, 350, 403, 498, 524, 558
quadprog, 455, 461
randomForest, 584, 605, 631, 640
rpart, 90, 100, 114, 159, 190, 261, 263, 296, 329, 344, 403, 498, 524, 558, 593, 605, 631, 640
rpart.plot, 72, 329, 344, 605, 631, 640
stats, 23, 355, 362, 403
R-squared, see coefficient of determination
R², see coefficient of determination
RAE, see relative absolute error
Rand index, 394–395
random forest, 443
attribute utility estimation by, 446
base models for, 443
for attribute selection, 584
instance proximity by, 445–446
model aggregation for, 443
out-of-bag evaluation for, 445
random naïve Bayes, 446–447
base models for, 447
model aggregation for, 447
rank, 28
c ompetition, 28
INDEX

rank (continued)
dense, 28
fractional, 28
ordinal, 28, 29
ranking, see rank
recall, 196
receiver operating characteristic, see ROC
reduced error pruning
for decision trees, 92–93
for regression trees, 275–276
regression algorithm, 16
regression flatness, 474
maximization of, 474
dual form of, 475–477
primal form of, 475
regression model, 15
regression performance, see predictive performance, for regression
regression performance measures, 296
regression task, xxii, 14
regression tree
as a nonparametric model, 239
as a regression model, 262
branch of, 262
growing of, see regression tree growing
leaf of, 262
node of, 262
closed, 264
open, 264
prediction for, 277
with fractional instances, see fractional instances, regression tree prediction with
pruning of, see regression tree pruning
randomized, 412
split application for, 271
with fractional instances, see fractional instances, split application with, for regression trees
split evaluation for, 269
with fractional instances, see fractional instances, split evaluation with, for regression trees
split for, 262
stop criteria for, 267–268
structure of, 262
target function dispersion for, 265, 269
target function location for, 265, 266
target function statistics for, 265–266
target value assignment for, 266
test for, see regression tree, split for
top-down induction of, 263
with weighted instances, 278
regression tree growing, 263–264
with fractional instances, see fractional instances, regression tree growing with
regression tree pruning, 274
control strategy for, 275, 277
cost-complexity, see cost-complexity pruning, for regression trees
criterion for, 275
minimum error, see minimum error pruning, for regression trees
operators for, 275
reduced error, see reduced error pruning, for regression trees
regression tube, 474
instance lying on, 474
instance lying outside, 474
instance lying within, 474
relative absolute error, 299
weighted, 301
relative square error, 277, 300
relative standard deviation, 34
RELIEF, 577
algorithm scheme for, 577–578
attribute utility estimation by, 578
dissimilarity measure for, 578
with missing values, 580–581
multiclass, 581
with multiclass encoding, 581
for regression, 582–583
representation function, 136, 239
inner, 137, 252
linear, 137
outer, 137
residual, 296, 297, 299
RMSE, see root mean square error
ROC analysis, 200
  weighted, 209
curve, 201
  area under, 209
  for a random model, 202
plane, 200
point, 201
root mean square error, 299
  weighted, 301
RSE, see relative square error
rule set
decision tree converted to, see decision tree, conversion to rules of
  pruning of, see rule set pruning
rule set pruning, 101
  by condition removal, 101
  by rule removal, 101
SAMME, 439
sample, xxiii
scoring function, 12, 201
search, 562
  cost function for, 562
  evaluation function for, 562
  final state for, 562
  for attribute selection, see attribute selection search
  initial state for, 562
  operator for, 562
  state space for, 562
search strategy, 562
  blind, 562
  heuristic, 562
  informed, see search strategy, heuristic
  uninformed, see search strategy, blind
sensitivity, 196
separating hyperplane
  maximum-margin, 460
separating hyperplane, 139, 152
distance to, 152, 457
  minimization of, 152
  signed, 152
separation, 377
significance test, 48
false negative for, 49
false positive for, 49
hypothesis for, 48
  alternative, 48
null, 48
  multiple, 62
  nonparametric, 58
  parametric, 58
p-value for, 49
  significance level for, 49
  statistic for, 48
  type I error for, see significance test, false positive for
type II error for, see significance test, false negative for
  unsatisfied assumptions of, 61
  vs. relationship strength, 49
silhouette plot, 380
silhouette width, 379
  average, 389
  for a cluster, 380
  for an instance, 379
similarity measure
  correlation-based, 314, 322
  linear, 322
  Pearson’s, see similarity measure,
  correlation-based, linear
  rank, 323
  Spearman’s, see similarity measure,
  correlation-based, rank
  with discrete attributes, 322
  cosine, 323
  Gower, see Gower’s coefficient
simple statistical filter, 568–569
  relationship measure for, 568
  with mixed attributes
  types, 568–569
single linkage, see linkage, single
slack variable, 468
specificity, 196, 197
spread, see dispersion
stacking, 433
  base models for, 433
  model aggregation for, 433
INDEX

standard deviation, 33
standardization, 504
   as a modeling transformation, 504
   for dissimilarity calculation, 321
   for k-centers clustering, 329
statistical hypothesis, see significance test, hypothesis
step-size, 146, 243, 245
   for gradient boosting, 440
support vector
   for SVM, 461, 465
   soft-margin, 469
   for SVR, 476
support vector machines, see SVM
support vector regression, see SVR
surrogate splits, 106, 279
   for decision trees, 113–114
   for regression trees, 284
SVM, 460
   cost parameter for, 469
   dual form of, see classification
      margin, maximization of, dual form
   hard-margin, 468
   kernel-based, see kernel trick, for SVM
   prediction for, 466, 482
   primal form of, see classification
      margin, maximization of, primal form of
   soft-margin, see classification margin, maximization of, soft
SVR, 473
   cost parameter for, 476
   dual form of, see regression flatness, maximization of, dual form of
   kernel-based, see kernel trick, for SVR
   prediction for, 477, 483
   primal form of, see regression
      flatness, maximization of,
   symmetric uncertainty, 55
TAN, see naïve Bayes classifier, tree-augmented
   target attribute, xxii, 6
      for attribute transformation, 500
      for classification, see concept
      for discretization, 526
      for regression, see target function
   target function, 14
   target label, see target value
   target value, 14
   test set, 190, 213, 303
tile coding, 255
top-down discretization, see discretization, top-down
total probability law, 119, 121
training information, 3, 5
training set, 6–7
   for attribute transformation, 500
   for attribute selection, 560
      in a broad sense, 7
   for classification, 10
   for clustering, 17
   for discretization, 527
      generalized, 216
   for regression, 14
      in a narrow sense, 7
   transformation model, 501
tree-augmented naïve Bayes classifier, see
   naïve Bayes classifier, tree-augmented
type I error, see false positive
type II error, see false negative
unbalanced classes, see class imbalance
unit step function, 139
validation set, 190, 213, 215, 303
variable, xxiii
variance, 31
  \( m \)-estimated, see variance, \( m \)-estimate of
  \( m \)-estimate of, 46, 276
  pooled, 56
unbiased estimator of, 31
  weighted, 32
Ward linkage, see linkage, Ward
weak learner, 434
weighted instances, see instance weighting
Welch’s test, 56
Widrow-Hoff rule, see delta rule, linear
Wilcoxon test, see
  Mann-Whitney-Wilcoxon test