Index

K-means algorithm, 238, 265
K-means algorithm (Section 8.4), 228
K-means clustering (Example **??**), 21
z-scores, 287
$\sqrt{\text{Gini}}$, **121**, 131, 133
0-1 loss, **55**
0-norm, **216**
1-norm, **216**, 220
2-norm, **216**, 220

A data set describing properties of learning models (Example 1.7), 34
abstraction, **280**
accuracy, **16**, **47**
Accuracy as a weighted average (Example 2.1), 48
active learning, **109**
adjacent violators, **67**
affine transformation, **180**
agglomerative merging, **285**
Agglomerative merging using χ^2 (Example 10.7), 286
AggloMerge(S, f, Q) (Algorithm 10.2), 286
aggregation, **280**
Aleph, *see* ILP systems
All splitting criteria are equal, but some are more equal than others... (Example 5.3), 130
analysis of variance, **323**
ANOVA, *see* analysis of variance
anti-unification, **110**
Apriori, *see* association rule algorithms
AQ, *see* rule learning systems
Assessing and visualising ranking performance (Section 2.2), 55
Assessing classification performance (Section 2.1), 47
Assessing probability estimates (Section 2.3), 64
association rule, **13**, **169**
association rule algorithms
  Apriori, 177
  Warmr, 178
Association rule discovery (Example 3.12), 89
Association rule mining (Section 6.3), 167
AssociationRules(D, f_0, c_0) (Algorithm 6.7), 171
at least as general as, **94**
attribute, *see* feature
AUC, **58**
average recall, *see* recall, average, 88

backtracking search, **120**
bag of words, **36**
bagging, 141
Bagging and random forests (Section 11.1), 302
Bagging(D, T, $\mathscr{A}$) (Algorithm 11.1), 303
basic linear classifier, xxiv, **18**, 190, 201, 219, 249
Bayes' rule, **23**

Bayes-optimal, **24**, 26, **243**
beam search, **157**
Bernoulli distribution, **252**
    multivariate, **251**
Bernoulli trial, **251**
Bernoulli, Jacob, 39, 252
BestSplit-Class(D, F) (Algorithm 5.2), 124
Beyond binary classification: Summary and further reading (Section 3.4), 90
Beyond conjunctive concepts (Section 4.3), 105
bias, **83**
bias and variance, 308
Bias, variance, and margins (Section 11.3), 308
bias-variance dilemma, **82**
Big Lebowski, The, 13
bigram, 295
bin, **283**
Binary classification and related tasks: Summary and further reading (Section 2.4), 69
binomial distribution, **252**
bit vector, **251**
Bivariate Gaussian mixture (Example 9.3), 246
Bivariate linear regression in matrix notation (Example 7.3), 186
Bonferroni-Dunn test, **325**
Boosted rule learning (Section 11.2), 307
Boosting (Section 11.2), 304
boosting, 55
Boosting(D, T, $\mathscr{A}$) (Algorithm 11.3), 305
bootstrap sample, **302**
breadth-first search, 168
Brier score, **65**

C4.5, *see* tree learning systems
Calculating impurity (Example 5.1), 122
Calculations on features (Section 10.1), 274
calibrating classifier scores, 206, 289
calibration, 203
    isotonic, **68**, 207, 262, **291**
    logistic, xxviii, 262, **289**
calibration loss, **66**
calibration map, **67**
Calibration of categorical features (Example 10.8), 288
CART, *see* tree learning systems
Cartesian product, **44**
categorical distribution, **252**
Categorical, ordinal and real-valued features (Section 10.1), 278
central limit theorem, **203**, 318
central moment, **278**
centre around zero, **20**, **182**, **185**, 187, 296
centre of mass, **20**
centroid, **86**, **219**
characteristic function, **44**
Chebyshev distance, **216**
Chebyshev's inequality, **274**
Chervonenkis, Alexey, 112
chicken-and-egg problem, 263
cityblock distance, *see* Manhattan distance
class
    label, **46**
Class imbalance (Example 2.4), 58
Classification (Section 2.1), 46
classification
    binary, **46**
    multi-class, **12**, 71
classifier, **46**
clause, **94**
Closed concepts (Section 4.2), 105
clustering, **12**
    agglomerative, **235**, 283
    descriptive, **15**
    predictive, **15**
    stationary point, **229**
Clustering around medoids (Section 8.4), 229
Clustering machine learning methods (Example 8.4), 228
Clustering trees (Section 5.3), 137
clustering trees, 233
CN2, *see* rule learning systems, *see* rule learning systems, *see* rule learning systems

CNF, *see* conjunctive normal form
Combining transformations (Example **??**), 19
comparable, **44**
Comparing Laplace-corrected precision and average recall (Example 6.6), 164
complement, 163
complete, **102**
component, **246**
Compression-based models (Section 9.5), 268
computational learning theory, 111
concavity, **67**
concept, 143
  closed, **105**, 169
  conjunctive, 94
concept learning, **46**, **93**
Concept learning: Summary and further reading (Section 4.5), 113
conditional likelihood, **260**
conditional random field, 271
confidence, **169**
Confidence interval (Example 12.5), 320
confidence interval, **320**
confusion matrix, **47**
conjugate prior, **243**
conjunction, **29**
conjunction $\wedge$, **94**
conjunctive normal form, **94**, 105
conjunctively separable, **103**
consistent, **102**
constructive induction, **119**
contingency table, **47**
Contingency tables for clustering (Example 3.10), 87
continuous feature, *see* feature, real-valued
convex, **196**, **221**
  hull, **67**
  loss function, **55**
  ROC curve, **67**, 125
  set, **102**, **168**
convex hull
  lower, **283**
correlation, 137
correlation coefficient, **39**, **245**, 295
cosine similarity, **238**
cost ratio, **63**
count vector, **251**
counter-example, **108**
covariance, **39**, 182
covariance matrix, **185**, 186, 188, 219, 245, 247, 248
coverage counts, **76**
Coverage counts as scores (Example 3.4), 76
coverage curve, **58**
coverage plot, **50**
covering algorithm, **149**
covers, **94**, **168**
critical difference, **324**
critical value, **322**
Cross-validation (Example 12.4), 319
cross-validation, **17**, **318**
  internal, **326**
  stratified, **319**
curse of dimensionality, **224**

d-prime, **289**
data mining, 167
data set characteristics, 310
Data that is not conjunctively separable (Example 4.4), 103
De Morgan laws, **94**, 119
decile, **276**
decision boundary, 4, 12
decision list, **29**, 177
decision rule, **22**
decision stump, **309**
decision tree, **27**, 46
decision tree learning, 89
decision tree training algorithm, 287
Decision trees (Section 5.1), 121
decoding, **73**
  loss-based, **75**
deduction, **16**

default rule, **30**, **147**
degree of freedom, **321**
degrees of freedom, **50**
Dendrogram (Definition 8.4), 234
dendrogram, **234**
descriptive clustering, **84**
descriptive model, **14**
Descriptive rule learning (Section 6.3), 161
dimensionality reduction, 298
Dirichlet prior, **66**
discretisation, 140
  agglomerative, **283**
  bottom-up, **283**
  divisive, **283**
  equal-frequency, **283**
  equal-width, **283**
  top-down, **283**
Discriminative learning by optimising conditional likelihood (Section 9.3), 259
disjunction, **29**
disjunction $\vee$, **94**
disjunctive normal form, **94**
dissimilarity, 85, **137**
  cluster, **137**
  split, **138**
distance, **20**
  Euclidean, **21**, 279
  Manhattan, **21**
Distance metric (Definition 8.2), 216
distance metric, **216**, 279
distance weighting, **225**
Distance-based clustering (Section 8.4), 225
Distance-based models: Summary and further reading (Section 8.7), 239
divide-and-conquer, **30**, **120**, 125, 147
DKM, **141**
DNF, *see* disjunctive normal form
dominate, **51**
DualPerceptron(D) (Algorithm 7.2), 192

Eddington, Arthur, 313
edit distance, **216**
eigendecomposition, **297**
Einstein, Albert, 26, 313
Elliptical distance (Example 8.1), 218
EM, *see* Expectation-Maximisation
empirical probability, **66**, 121, 122, 125
entropy, 121, 123, 124, 131, 133, 145, **269**
equivalence class, **44**
equivalence oracle, **107**
equivalence relation, **44**
error rate, **48**
estimate, **39**
Euclidean distance, **216**
European Conference on Machine Learning, 2
European Conference on Principles and Practice of Knowledge Discovery in Databases, 2
evaluation measures, 314
example, **44**
excess kurtosis, *see* kurtosis
exemplar, **85**, **86**, **219**
Expectation-Maximisation, **265**
Expectation-Maximisation (Section 9.4), 264
expectation-maximisation, 85
Expectation-Maximisation algorithm, 294
Expected accuracy and *AUC* (Example 12.3), 316
Expected accuracy for unknown class distributions (Example 12.1), 314
expected value, **39**, 245
experiment, **313**
experimental objective, **314**
explanation, **31**
explanatory variable, *see* feature
exponential loss, **55**, 307
extension, **94**

F-measure, **87**, **275**, 316
  insensitivity to true negatives, 316
false alarm rate, **48**
false negative, **48**
false negative rate, **48**

false positive, **48**
false positive rate, **48**
feature, **11**, **45**, 241
  binarisation, **281**
  Boolean, **278**
  calibration, **288**
  categorical, **278**
  construction, **36**, 45
  decorrelation, 186, 219, 248, 249
  discretisation, **36**, **283**
  discretisation, supervised, **283**
  discretisation, unsupervised, **283**
  domain, **33**, **45**
  normalisation, 186, 219, 248, 249, **287**
  ordinal, **278**
  real-valued, **278**
  structured, **280**
  thresholding, **282**
  thresholding, supervised, **282**
  thresholding, unsupervised, **282**
  transformation, **281**
  two uses of, 35
  unordering, **281**
feature calibration, 254
Feature construction and selection (Section 10.3), 295
feature list, **28**
feature selection, 224
  backward elimination, **296**
  filter, **295**
  forward selection, **296**
  Relief, **295**
  wrapper, **296**
feature space, **207**, **208**
Feature transformations (Section 10.2), 281
Feature tree (Definition 5.1), 119
feature tree, **27**, **119**, 141
  complete, **28**
Features: Summary and further reading (Section 10.4), 299
Features: the workhorses of machine learning (Section 1.3), 33
finding the point that minimises sum of squared Euclidean distances to a set of points, 267
first-order logic, **110**
First-order rule learning (Section 6.4), 172
FOIL, *see* ILP systems
forecasting theory, **65**
frequency, *see* support
frequent item sets, **168**
FrequentItems(D, f_0) (Algorithm 6.6), 170
Friedman test, **323**
Friedman test (Example 12.8), 324
From kernels to distances (Section 8.6), 237
function estimator, **80**
functor, **175**

Gauss, Carl Friedrich, 90, 181
Gaussian distribution, **245**
Gaussian kernel, **209**
  bandwidth, **209**
Gaussian mixture model, 246, **265**
  relation to K-means, 268
Gaussian mixture models (Section 9.4), 265
general, **40**
generalised linear model, 271
generality ordering, **94**
generative model, **25**
geometric median, **219**
Gini coefficient, 121
Gini index, 121–124, 131, 133, 134, 145
Gini, Corrado, 121
glb, *see* greatest lower bound
Godfather, The, 13
Going beyond linearity with kernel methods (Section 7.5), 207
Golem, *see* ILP systems
Gosset, William Sealy, 321
gradient, **196**, 219
grading model, 45, 81
Gram matrix, **185**, 193, 198, 297

greatest lower bound, **95**
greedy algorithm, **120**
grouping model, 45, 81
Growing a tree (Example 5.2), 126
GrowTree(D, F) (Algorithm 5.1), 120
Guinness, 321

HAC(D, Dis) (Algorithm 8.4), 235
Hamming distance, **73**, **216**, 279
Handling more than two classes (Section 3.1), 71
harmonic mean, 87
hidden variable, **13**, **264**
hierarchical agglomerative clustering, 287
Hierarchical clustering (Section 8.5), 233
Hierarchical clustering of machine learning methods (Example 8.6), 233
hinge loss, **55**, **200**
Histogram (Example 10.2), 277
histogram, **277**
homogeneous coordinates, 4, **20**, 180, 184
Horn clause, **94**, 105, **172**
Horn theory, 105
Horn(Mb, Eq) (Algorithm 4.5), 108
Horn, Alfred, 94
How to interpret it (Section 12.3), 319
How to measure it (Section 12.2), 317
Hume, David, 16
hyperplane, **18**
hypothesis space, **95**, 172

ID3, *see* tree learning systems
ILP, *see* inductive logic programming
ILP systems
  Aleph, 178
  FOIL, 177
  Golem, 178
  Progol, 31, 178
IMDb movie database, 14
implication $\rightarrow$, **94**
impurity
  relative, **131**
impurity measure, 269
imputation, **294**
incomparable, **44**
incomplete, **30**
Incomplete features (Section 10.2), 294
inconsistent, **30**
independent variable, *see* feature
indicator function, **47**
induction, **16**
  problem of, **16**
inductive bias, **119**
inductive logic programming, **172**, 281
information content, **269**
information gain, **123**, 284, 295
information retrieval, 87, 275, 298, 315
Information-based classification (Example 9.7), 269
input space, **208**
instance
  labelled, **44**
instance space, **17**, 33, 35, **43**
  segment, **27**, 44, 93, **119**
instances, **43**
intercept, **180**, 182
Internal disjunction (Example 4.3), 100
Internal disjunction (Section 4.1), 100
internal disjunction, **100**, 148
Interpretation of results over multiple data sets (Section 12.3), 322
interquartile range, **276**, 287
isometric
  $\sqrt{\text{Gini}}$, 131
  accuracy, 53, 61, 67, 105
  average recall, **53**, 63
  entropy, 131
  Gini index, 131
  impurity, 145
  precision, 153
  precision (Laplace-corrected), 157
  splitting criteria, 131
Isotonic calibration of two features (Example 10.11),

293
Isotonic feature calibration (Example 10.10), 292
item set, **168**
closed, **169**

Jaccard coefficient, **12**
jackknife, 318

K-means
K-means
relation to Gaussian mixture model, 268
K-means, **21**, 84, 85, 283
K-means problem, **226**
K-medoids, 283
K-medoids algorithm, **229**
k-nearest neighbour, **225**
Karush-Kuhn-Tucker conditions, **196**
kernel, **38**, 295
kernel perceptron, **208**
kernel trick, **38**
Kernel-KMeans(D, K) (Algorithm 8.5), 238
KernelPerceptron(D, κ) (Algorithm 7.4), 209
Kinds of features (Section 10.1), 273
KKT, *see* Karush-Kuhn-Tucker conditions
KMeans(D, K) (Algorithm 8.1), 228
KMedoids(D, K, Dis) (Algorithm 8.2), 230
kurtosis, **278**

L_0 norm, *see* 0-norm
label space, **43**
Labelling a feature tree (Example 1.5), 27
Lagrange multiplier, **196**
landmarking, 311
Langley, Pat, 327
Laplace correction, **66**, 125, 133, 157, 243, 252, 256, 263
lasso, **189**
latent semantic indexing, **298**
latent variable, *see* hidden variable
latent variables, *see* hidden variable
lattice, **95**, 168, 172
law of large numbers, **39**
Learnability (Section 4.4), 111
learnability, **111**
Learning a clustering tree (Example 5.5), 138
Learning a clustering tree with Euclidean distance (Example 5.6), 139
Learning a Horn theory (Example 4.5), 108
Learning a quadratic decision boundary (Example 7.8), 207
Learning a regression tree (Example 5.4), 134
Learning a rule list (Example 6.1), 145
Learning a rule set for one class (Example 6.3), 153
Learning conjunctive concepts (Example 4.1), 94
learning from entailment, **114**
learning from interpretations, **114**
learning model, **111**
Learning ordered rule lists (Section 6.1), 144
learning rate, **191**
Learning unordered rule sets (Section 6.2), 152
LearnRule, 175
LearnRule(D) (Algorithm 6.2), 149
LearnRuleForClass(D, C_i) (Algorithm 6.4), 157
LearnRuleList, 175
LearnRuleList(D) (Algorithm 6.1), 149
LearnRuleSet(D) (Algorithm 6.3), 156
Least general generalisation (Section 4.1), 95
least general generalisation, **95**, 100, 101, 103, 105, 119
least upper bound, **95**
least-squares classifier, **190**
least-squares method, **181**
least-squares solution to a linear regression problem, 250
leave one out, **318**
level-wise search, 168
Levenshtein distance, **216**
LGG, *see* least general generalisation
LGG-Conj(x, y) (Algorithm 4.2), 97
LGG-Conj-ID(x, y) (Algorithm 4.3), 100
LGG-Set(D) (Algorithm 4.1), 97

lift, **171**
likelihood function, **23**, 279
likelihood ratio, **24**
Line fitting example (Example 3.8), 80
linear
  approximation, **180**
  combination, **180**
  function, **180**
  model, **179**
  transformation, **180**
Linear classification (Example 1), 2
linear discriminants, 18
Linear models: Summary and further reading (Section 7.6), 210
linear regression, 81, 137
linear, piecewise, **180**
linearly separable, 191
Linkage function (Definition 8.5), 234
linkage function, **234**
  monotonicity, **236**
Linkage matters (Example 8.7), 235
literal, **94**
Lloyd's algorithm, 228
local variables, **175**, **280**
log-likelihood, 249
log-linear models, **206**
log-odds space, 254, **289**
Logistic calibration of a linear classifier (Example 7.7), 205
Logistic calibration of two features (Example 10.9), 290
logistic function, **204**
logistic regression, 206, **259**
loss function, **55**, 81
Loss-based decoding (Example 3.3), 75
L_p norm, *see* p-norm
LSA, *see* latent semantic indexing
lub, *see* least upper bound

m-estimate, **66**, **256**
Mach, Ernst, 26
machine learning
  definition of, 4
  univariate, **46**
Machine learning experiments: Summary and further reading (Section 12.5), 326
Mahalanobis distance, **219**, 249
majority class, **28**, 30, **46**, **49**
Manhattan distance, **216**
manifold, **224**
MAP, *see* maximum a posteriori
Mapping the ensemble landscape (Section 11.3), 308
margin, **19**, 194
  of a classifer, 55
  of a decision boundary, **195**
  of an example, **55**, 195, 309
margin error, **200**
marginal, **47**, **171**
marginal likelihood, **25**
market basket analysis, 89
matrix
  diagonal, 186
  inverse, 245
  rank, **298**
matrix completion, **298**
matrix decomposition, 14, 85, 296–299
  Boolean, 298
  non-negative, 300
  with constraints, 297
Matrix transformations and decompositions (Section 10.3), 296
maximum a posteriori, **24**
maximum likelihood, **24**
maximum-likelihood estimate, 184
maximum-likelihood estimation, **249**, 264
mean, 245, **274**
  arithmetic, **275**
  geometric, **275**
  harmonic, **275**
mean squared error, **64**
Measuring similarity (Example 1.1), 12

median, 245, **274**
medoid, **219**
membership oracle, **107**
Meta-learning (Section 11.3), 310
meta-model, **309**
MGConsistent(C, N) (Algorithm 4.4), 104
midrange point, **275**
Minimum description length principle (Definition 9.1), 270
Minkowski distance, **215**
Minkowski distance (Definition 8.1), 215
Missing values (Example 1.2), 22
mixture model, **246**
ML, *see* maximum likelihood
mode, 245, **274**
model, **11**, **43**
  declarative, **31**
  geometric, **17**
  grading, **31**
  grouping, **31**
  logical, 27
  parametric, **179**
  probabilistic, 22
  univariate, **35**
model ensemble, **301**
Model ensembles: Summary and further reading (Section 11.4), 310
model selection, **244**
model tree, **137**
Models: the output of machine learning (Section 1.2), 17
monotonic, **168**, **279**
Monty Python's Flying Circus, 1
more general than, **94**
Most general consistent hypotheses (Section 4.2), 104
MSE, *see* mean squared error
Multi-class AUC (Example 3.5), 76
Multi-class classification (Section 3.1), 71
Multi-class probabilities from coverage counts (Example 3.7), 79
Multi-class scores and probabilities (Section 3.1), 75
multinomial distribution, **252**
multivariate linear regression, 254
multivariate naive Bayes
  decomposition into univariate models, **27**
multivariate normal distribution, 265
multivariate regression
  decomposition into univariate regression, 187, 332

n-gram, **295**
naive Bayes, **26**, 187
naive Bayes assumption, **251**
naive Bayes classifier, 295
Nearest-neighbour classification (Section 8.3), 223
nearest-neighbour classifier, **21**, **223**
nearest-neighbour retrieval, **224**
negation ¬, **94**
Negative examples (Example 4.2), 97
neighbour, **219**
Neighbours and exemplars (Section 8.2), 219
Nemenyi test, **324**
neural network, 191
Newton, Isaac, 26
no free lunch theorem, **16**, 310
noise, **45**
  instance, **45**
  label, **45**
nominal feature, *see* feature, categorical
normal distribution, 203, **245**, 279
  multivariate, **245**
  multivariate standard, **245**
  standard, **245**, 248
normal vector, **180**
normalisation, **182**
  row, **79**
Normalisation and calibration (Section 10.2), 287
null hypothesis, **321**

objective function, **55**, **196**

Obtaining probabilities from linear classifiers (Section 7.4), 202
Occam's razor, **26**
one-versus-one, **72**
One-versus-one voting (Example 3.2), 74
one-versus-rest, **72**
operating conditions, **63**
operating context, **314**
optimisation
  constrained, 195, **196**
  dual, **196**, 197
  multi-criterion, **51**
  primal, **196**
  quadratic, 195, 196
Opus, *see* rule learning systems
ordinal features, 215
ordinal, **274**
Other descriptive models (Section 3.3), 87
Other ensemble methods (Section 11.3), 309
outlier, **183**, 220, **276**
output code, **72**
output space, **43**
Overfitting (Example 2), 5
overfitting, xxiii, 16, 28, 45, 80, 81, 85, 119, 137, 180, 194, 262, 295
Overlapping rules (Example 1.6), 29

p-norm, **215**
p-value, **321**
PAC, *see* probably approximately correct
Paired t-test (Example 12.6), 321
paired t-test, **321**
PAM(D, K, Dis) (Algorithm 8.3), 231
Pareto front, **51**
partial order, **44**
partition, **44**
partition matrix, **85**
partitioning around medoids, **229**, 283
Paths through the hypothesis space (Section 4.2), 101
PCA, *see* principal component analysis
Pearson, Karl, 274
percentile, **275**
Percentile plot (Example 10.1), 276
percentile plot, **276**
perceptron, **191**
  online, **191**
Perceptron(D) (Algorithm 7.1), 191
PerceptronRegression(D) (Algorithm 7.3), 194
Performance evaluation (Section 1.1), 15
Performance of multi-class classifiers (Example 3.1), 72
piecewise linear, *see* linear, piecewise
population mean, **39**
post-hoc test, **324**
post-processing, **171**
Posterior odds (Example 1.3), 24
posterior odds, **24**
posterior probability, **22**, 241
powerset, **44**
Príncipe, 313
precision, **49**, 87, **153**, 172, 275
  Laplace-corrected, 157
Precision and recall as evaluation measures (Example 12.2), 315
predicates, *see* first-order logic
predicted positive rate, **316**
Prediction using a naive Bayes model (Example 9.4), 253
Predictive and descriptive clustering (Section 3.3), 84
predictive clustering, **84**
predictive model, **14**
predictor variable, *see* feature
principal component analysis, 20, 224, **296**
prior odds, **24**
prior probability, **23**
probabilistic model
  discriminative, **241**
  generative, **241**
Probabilistic models for categorical data (Section 9.2), 251

Probabilistic models with hidden variables (Section 9.4), 263
Probabilistic models: Summary and further reading (Section 9.6), 270
probability distribution
cumulative, **277**
right-skewed, **277**
Probability estimation (Section 2.3), 63
probability estimation tree, 241
probability estimator, **63**
probability smoothing, **66**
probability space, **290**
probably approximately correct, **111**, 301
Progol, *see* ILP systems, *see* ILP systems
projection, 202
Prolog, *see* query languages, *see* query languages, *see* query languages
propositional logic, **110**
propositionalisation, **280**
prototype, **21**
PruneTree(T, D) (Algorithm 5.3), 130
pruning, **28**, **130**
pruning set, **130**
pseudo-count, **252**, **256**
pseudo-counts, xxii, **66**, 158
pseudo-metric, **217**
pure, **121**
purity, **30**, 144

quantile, **276**
quartile, **276**
query, **280**
query languages
Prolog, 172, 175–177, 280
SQL, 280

Rand index, **87**
random forest, **303**
random forests, 309
random variable, **39**
RandomForest(D, T, d) (Algorithm 11.2), 304
range, **275**
ranking, **56**
Ranking accuracy (Example 2.3), 56
ranking accuracy, **56**
Ranking and probability estimation trees (Section 5.2), 125
ranking error, **56**
ranking error rate, **56**
Ranking example (Example 2.2), 56
recalibrated likelihood decision rule, **255**
recall, **50**, 87, 275
average, **52**, **164**
receiver operating characteristic, **53**
RecPart(S, f, Q) (Algorithm 10.1), 284
recursive partitioning, **284**
Recursive partitioning using information gain (Example 10.6), 284
reduced-error pruning, **130**, 133, 137
incremental, 177
Reducing scatter by partitioning data (Example 8.3), 226
refinement, **60**
refinement loss, **66**
Regression (Section 3.2), 80
regression, **12**, 55
isotonic, 68
multivariate, **186**
univariate, **181**
regression coefficient, **182**
Regression trees (Section 5.3), 134
regressor, **80**
regularisation, **188**, 200
regularisation term in ridge regression, 270
Regularised regression (Section 7.1), 188
reject, **73**
relation, **44**
antisymmetric, **44**
equivalence, *see* equivalence relation
reflexive, **44**
symmetric, **44**
total, **44**
transitive, **44**

Representing clusterings (Example 3.9), 86
residual, **81**, **181**
Reweighting multi-class scores (Example 3.6), 77
ridge regression, **189**
Ripper, *see* rule learning systems, *see* rule learning systems
ROC curve, **58**
ROC heaven, **61**, 132
ROC plot, **53**
Rocchio classifier, **221**
rotation, **20**
rule, **94**
  body, **143**
  head, **143**
  list, **143**
  set, **143**
Rule learning for subgroup discovery (Section 6.3), 162
rule learning systems
  AQ, 177
  CN2, xxix, 177, 325
  Opus, 177
  Ripper, 177, 311
  Slipper, 311
  Tertius, 178
rule lists, 125
Rule lists as rankers (Example 6.2), 150
Rule models: Summary and further reading (Section 6.5), 177
Rule sets as rankers (Example 6.4), 158
Rule tree (Example 6.5), 160
rule tree, **160**

sample complexity, **111**
sample covariance, **39**
sample mean, **39**
sample variance, **39**
scale, **274**
  reciprocal, **275**
scaling, **20**
  uniform, **20**
scaling matrix, 186
Scatter (Definition 8.3), 226
scatter, **85**, **226**
  within-cluster, **85**
scatter matrix, **185**, **226**, 297
  between-cluster, **226**
  within-cluster, **226**
Scoring and ranking (Section 2.2), 53
scoring classifier, **53**
SE, *see* squared error
search heuristic, **55**
seed example, **156**
segment, **31**
semi-supervised learning, **15**
sensitivity, **48**
Sensitivity to skewed class distributions (Section 5.2), 130
separate-and-conquer, **30**, **147**, 148
sequential minimal optimisation, **202**
set, **44**
  cardinality, **44**
  complement, **44**
  difference, **44**
  disjoint, **44**
  intersection, **44**
  subset, **44**
  union, **44**
  universe, **44**
Shannon, Claude, 269
shatter, **112**
Shattering a set of instances (Example 4.7), 112
Shawshank Redemption, The, 13
shrinkage, **188**
sigmoid, **204**
signal detection theory, **53**, 289
significance test, **321**
silhouette, **232**
Silhouettes (Section 8.4), 232
similarity, 63
singular value decomposition, **296**

skewness, **278**
Skewness and kurtosis (Example 10.3), 278
slack variable, **200**
slack variable term in soft-margin SVMs, 270
Slipper, *see* rule learning systems
slope, **180**
So many roads... (Section 8.1), 213
soft margin, **200**
Soft margin SVM (Section 7.3), 200
Soft margins (Example 7.6), 201
Spam or not? (Example 9.1), 242
SpamAssassin, 1–12, 53, 63
sparse data, **19**
sparsity, **189**
specific, **40**
specificity, **48**
split, **119**
    binary, **35**
SQL, *see* query languages
Squared error (Example 2.6), 65
squared error, **64**
squared Euclidean distance, 219
stacking, **309**
standard deviation, **274**
Stationary points in clustering (Example 8.5), 229
statistic
    of central tendency, **274**
    of dispersion, **274**
    shape, **274**, 277
stop word, **257**
stopping criteria, **149**
stopping criterion, **284**
Structured features (Example 10.4), 280
Structured features (Section 10.1), 279
Student's t distribution, **321**
sub-additivity, 218
subgroup, **88**, 162
    extension, **88**
Subgroup discovery (Example 3.11), 88
subgroup discovery, 14
subspace sampling, **303**
Summary and outlook (Section 1.4), 40
supervised learning, **12**, 14
support, **168**
support vector, **195**
support vector machine, **19**, 55, 195, 279
Support vector machines (Section 7.3), 193
SVD, *see* singular value decomposition
SVM, *see* support vector machine
    complexity parameter, **200**

target variable, **22**, 80, 241
task, **11**
Tasks: the problems that can be solved with machine learning (Section 1.1), 11
terms, *see* first-order logic
Tertius, *see* rule learning systems
test set, **16**, **45**
Texas Instruments TI-58 programmable calculator, 210
text classification, 9, 19
The arithmetic mean minimises squared Euclidean distance (Theorem 8.1), 219
The effect of outliers (Example 7.2), 183
The effect of weighted covering (Example 6.7), 165
The hypothesis space (Section 4.1), 94
The kernel trick (Example 1.9), 37
The least-squares method (Section 7.1), 181
The normal distribution and its geometric interpretations (Section 9.1), 245
The perceptron: a heuristic learning algorithm for linear classifiers (Section 7.2), 191
Thresholding and discretisation (Section 10.2), 282
top-down, **30**
total order, **44**
Training a naive Bayes model (Example 9.5), 256
Training a naive Bayes model (Section 9.2), 256
training set, 12, **44**
transaction, **168**
translation, **20**

Tree learning as variance reduction (Section 5.3), 134
tree learning systems
C4.5, 141
CART, 141
ID3, 141
Tree models: Summary and further reading (Section 5.4), 141
triangle inequality, **218**
trigram, 295
true negative, **48**
true negative rate, **48**
true positive, **48**
true positive rate, **48**, 87
Tuning your spam filter (Example 2.5), 61
Turning rankers into classifiers (Section 2.2), 61
turning rankers into classifiers, 255
Turning rankers into probability estimators (Section 2.3), 67
Two maximum-margin classifiers and their support vectors (Example 7.5), 198
Two neighbours know more than one (Example 8.2), 222
Two uses of features (Example 1.8), 35

underfitting, **180**
unification, **110**
Unification and anti-unification (Example 4.6), 110
unigram, 295
Univariate least-squares classifier (Example 7.4), 189
Univariate linear regression (Example 7.1), 181
Univariate logistic regression (Example 9.6), 261
Univariate mixture model with unequal variances (Example 9.2), 246
unstable, **188**
Unsupervised and descriptive learning (Section 3.3), 83
Unsupervised and supervised thresholding (Example 10.5), 282
unsupervised learning, **12**, 14
Using a naive Bayes model for classification (Section 9.2), 253
Using first-order logic (Section 4.3), 110
Using marginal likelihoods (Example 1.4), 25
Usual Suspects, The, 13

Vapnik, Vladimir, 112
variance, **20**, **39**, **83**, **134**, **135**, 182, **185**, **274**, 278
VC-dimension, **112**
Version Space (Definition 4.1), 102
version space, **102**
Viagra, 6–38
Visualising classification performance (Section 2.1), 50
Voronoi diagram, **86**
Voronoi tesselation, **222**

Warmr, *see* association rule algorithms
weak learnability, **301**
Weight updates in boosting (Example 11.1), 304
weighted covering, **166**
weighted covering algorithm, 308
weighted relative accuracy, **164**
WeightedCovering(D) (Algorithm 6.5), 168
What to measure (Section 12.1), 314
What you'll find in the rest of the book (Section 1.4), 41
Wilcoxon's signed-rank test, **322**
Wilcoxon's signed-rank test (Example 12.7), 322

χ^2 statistic, 89, **286**, 295

z-score, **245**, 289