
Index

- Accuracy, 119, 252
- AdaBoost, 217–223
- Akaike information criterion, 70
- Arrays, 21–23
 - arithmetic, 22
 - creating, 21
 - recycling, 22
 - sub-setting, 22
- Artificial neural networks, *see* Neural networks
- AUC, 252

- Bioconductor, 233
 - installing, 235
 - installing packages, 239
- Blocks of instructions, 31
- Boosting, 217
- Box percentile plots, 50
- Box plot rule, 173
- Box plots, 47
 - conditioned, 49
- Built-in data sets, 30

- Candlestick graphics, 110
- Classes and Methods, 33–34
 - classes, 33
 - generic functions, 34
 - inheritance, 34
 - methods, 33
 - polymorphism, 34
 - slots, 33
 - the @ operator, 33
- Classification tasks, 118
- Clustering analysis, 184
- Clustering methods
 - dendrograms, 205
 - hierarchical agglomerative, 205
- Conditioned plots, 49
- Confusion matrices, 120, 191, 264
- Continuous variables
 - discretizing, 51
- Correlation, 56
 - calculating, 56
 - nominal variables, 59
- Cross validation, 81–91
- Cross-tabulation, *see* Frequency tables
- Cumulative recall charts, 192

- Data frames, 26–30
 - creating, 26
 - entering data, 29
 - extending, 28
 - indexing, 49
 - naming columns, 29
 - number of columns, 29
 - number of rows, 29
 - querying, 27
 - sub-setting, 27
- Data streams, 122
- Decision stumps, 219
- Descriptive data mining, 184
- Distance metrics, 184, 255
 - calculating, 182
 - Euclidean, 61, 256
 - mixed mode, 201
- Dummy variables, 65, 201

- Ensemble learning, 88, 217, 250
- Error rate, 119, 252
- Errors scatter plot, 79
- Excess return, 132

ExpressionSet objects, 235

F measure, 120

Factors, 11–13, 49

- count occurrences, 12
- creating, 11
- levels, 11
- re-ordering levels, 60

Feature selection, 112, 241–251

- ANOVA test, 244
- feature clustering ensembles, 248
- filters, 112, 241
- random forests, 114, 246
- RELIEF, 243
- wrappers, 112, 241

Frequency tables, 12, 13

Function composition, 12, 17, 47

Functions

- creating, 30
- default values of parameters, 31
- ... parameter, 82
- returned value, 31

Future markets, 131

- buy limit orders, 131
- buy stop orders, 131
- long positions, 131
- sell limit orders, 132
- sell stop orders, 132
- short positions, 131

Growing window, 122

Histograms, 44, 239

- conditioned, 59

Hold out experiments, 194–195

Imbalanced classes, 185, 209–211

- over-sampling, 210
- SMOTE, 210
- under-sampling, 210

Incremental learners, 122

Index vectors

- character, 18
- empty, 18

integer, 17

logical, 16

negative indexes, 18

Interactive identification of cases, 80

K nearest neighbors, 255–257

Kernel density estimate, 46

Kolmogorov-Smirnov tests, 180

Leave one out CV, 253–254, 258–265

Lift charts, 191

Linear regression, 63–71

- adjusted r =Adjusted R^2 , 66
- anova tests, 67
- diagnostic information, 66
- graphical diagnostics, 66
- model simplification, 67, 69
- model updating, 67
- nominal variables, 65
- obtaining, 64
- predictions, 78
- proportion of variance, 66
- $r=R^2$, 66
- summary, 65
- `summary()`, 67, 70

Lists, 23–26

- components, 23
- concatenating, 25
- creating, 23
- extending, 25
- named components, 24
- number of components, 25
- removing components, 25
- subsetting, 23
- unflattening, 26

Local outlier factors (LOF), 205–208

Marginal frequencies, *see* Frequency tables

Matrices, 19–21

- creating, 19
- naming, 21
- sub-setting, 19

- Matrix algebra, 23
 - Maximum drawdown, 133
 - Mean Absolute Deviation, *see* Mean Absolute Error
 - Mean Absolute Error, 77
 - Mean Squared Error, 78, 116
 - Model formulas, 64
 - Model selection criteria, 77
 - Monte Carlo estimates, 142–156
 - Multivariate adaptive regression splines, 129–130
 - MySQL
 - creating a database, 36
 - creating a table, 37
 - inserting records, 37
 - listing records, 37
 - logging into the server, 36
 - quitting, 38
 - using a database, 36
 - NA value, 8, 42
 - Naive Bayes, 211–217
 - Neural networks, 123–126
 - Non-stationary time series, 121
 - Normal distribution, 16, 44
 - QQ plots, 45
 - Normalization, *see* Standardization
 - Normalized distance to typical price, 193
 - Normalized Mean Squared Error, 78
 - ORh, 205
 - Outlier detection, 184
 - Outliers, 46, 48
 - identification, 48
 - Overfitting, 72, 81
 - Packages
 - adabag, 218
 - ALL, 235
 - Biobase, 235, 241
 - car, 45
 - class, 256
 - cluster, 201
 - DBI, 105, 107
 - dprep, 201, 243
 - e1071, 212
 - earth, 129
 - ff, 107
 - genefilter, 239, 243
 - Hmisc, 167, 250
 - installing, 4
 - kernlab, 127
 - klaR, 212
 - lattice, 49, 248
 - loading, 49
 - mda, 129
 - nnet, 124
 - PerformanceAnalytics, 132, 158
 - quantmod, 103, 108
 - randomForest, 255
 - RMySQL, 105, 107
 - ROCR, 189, 252
 - RODBC, 105
 - RWeka, 218
 - tseries, 102
 - TTR, 112
 - updating, 5
 - xts, 98
 - zoo, 98, 102
 - Precision, 120, 188
 - Precision/recall curves, 188
 - Probabilistic classifiers, 186
 - Profit/loss, 132
- ## R
- command prompt, 3
 - entering commands, 3
 - executing commands in a file, 35
 - help mailing lists, 1
 - help system, 5
 - installing, 3
 - installing add-on packages, *see* Packages, installing
 - loading objects, 35
 - quitting, 4
 - running, 4
 - saving objects, 35

- saving the workspace, 35
- R objects
 - attributes, 198
 - class, 33
 - listing, 7
 - methods, 33
 - removing, 7
 - slots, 33, 92
 - structures, 198, 265
 - valid names, 7
- R operators
 - @, 33, 92
 - %in%, 171
 - arithmetic, 6
 - assignment, 6
 - logical, 17
 - logical negation, 49
 - sequence, 14
- Random forests, 88, 114, 255
- Random sequences, *see* Sequences, random
- Reading data from a text file, 42
- Recall, 120, 188
- Recycling rule, 10, 14, 17
- Regime shift, 121
- Regression tasks, 117
- Regression trees, 63, 71–77
 - cost complexity pruning, 74
 - graphical representation, 72
 - model interpretation, 72
 - obtaining, 71
 - overfitting, 72
 - predictions, 78
 - pruning, 72
 - 1-SE rule, 75
 - interactive, 76
 - set of sub-trees, 74
 - stopping criteria, 74
 - summary, 72
- Relative frequencies, *see* Frequency tables
- Reliable performance estimates, 81
- ROC analysis, 252
- Self-training, 223–229

- Semi-supervised learning, 186
- Sequences, 14
 - of factors, 15
 - of integers, 14
 - of reals, 14
 - random, 16
 - with repetitions, 15
- Sharpe ratio, 133
- Shorth, 239
- Similarity, 184, 255
- Sliding window, 122
- Standardization, 62, 124, 182, 245
- Statistics of centrality, 55, 179, 239, 257
- Stratified samples, 194
- Strip plots, 51
- Student *t* distribution, 16
- Summary statistics, 43
- Supervised learning, 184, 185
- Support vector machines, 127–129, 187, 254
- T indicator, 109
- Technical indicators, 112
- Time classes, 99
 - POSIXt, 99
 - Date, 99
- Time series, 97
- Trading simulator, 133
- Trading strategies, 131
- Type coercion, 8
- Unknown values, 53–63
 - imputation strategies, 53
- Unsupervised learning, 184
- Vectorization, 10–11
- Vectors, 7–9
 - adding elements, 9
 - arithmetic, 10
 - creating, 8
 - empty vector, 9
 - indexing, 8
 - length, 7
 - naming elements, 18
 - recycling, 10

removing elements, 9
sub-setting, 16

Web sites

CRAN mirrors, 4
MySQL, 2, 36
R, 3
R mailing lists, 1
this book, 2

Weka, 218

Wilcoxon tests, 89

Working directory

changing, 35
checking, 35

xts objects, 98

creating, 98
indexing, 99