



A

algorithms
 bagged decision trees, 4
 base learners, 211–212
 boosted decision trees, 4
 bootstrap aggregation, 226–236
 choosing, 11–13
 comparison, 6
 ensemble methods, 1
 linear, compared to nonlinear, 87–88
 logistic regression, 4
 multiclass classification problems, 314–315
 nonlinear, compared to linear, 87–88
 penalized linear regression
 methods, 1
 Random Forests, 4
 ANNs (artificial neural nets), 4
 argmin, 110–111
 attributes. *See also* features;
 independent variables; inputs;
 predictors
 categorical variables, 26, 77
 statistical characterization, 37
 cross plots, 42–43
 factor variables, 26, 77

features, 25
 function approximation and, 76
 increase, 5
 labels, relationship visualization, 42–49
 numeric variables, 26, 77
 predictions and, 3
 real-valued, 62–68, 77
 squares of, 197
 targets, correlation, 44–47
 times residuals, 197
 AUC (area under the curve), 88

B

bagging, 11, 212, 226–236, 270–275
 bias *versus* variance, 229–231
 decision trees, 235–236
 multivariable problems and, 231–235
 random forests and, 247–250
 base learners, 9, 211–212
 basis expansion, 19
 linear methods/nonlinear problems, 156–158
 best subset selection, 103
 bias *versus* variance, 229–231

- binary classification problems, 78
 - ensemble methods, 284–302
 - penalized linear regression methods and, 181–191
- binary decision trees, 9–10, 212–213
 - bagging, 11
 - categorical features, 225–226
 - classification features, 225–226
 - overfitting, 221–225
 - predictions and, 213–214
 - training, 214–217
 - tree training, 218–221
- boosting, 212
- bootstrap aggregating. *See* bagging
- box and whisker plots, 54–55
 - normalization and, 55

C

- categorical variables, 19, 26
 - binary decision trees, 225–226
 - classification problems, 27
 - statistical characterization, 37
- chapter content and dependencies, 18–20
- chirped signals, 28
- chirped waveform, 151
- class imbalances, 305–307
- classification problems
 - algorithms and, 2–3
 - binary, penalized linear regression and, 181–191
 - binary decision trees, 225–226
 - categorical variables, 27
 - chirped signals, 28
 - class imbalances, 305–307
 - converting to regression, 152–154
 - multiclass, 68–73, 204–209
 - ensemble methods, 302–314
 - multiple outcomes, 155–156
 - penalized linear regression methods, 151–155
- coefficient estimation
 - Lasso penalty and, 129–131
 - penalized linear regression and, 122
- coefficient penalized regression, 111

- complex models, compared to simple models, 82–86
- complexity
 - balancing, 102–103
 - simple problems *versus* complex problems, 80–82
- complexity parameter, 110
- confusion matrix, 91
- contingency tables, 91
- correlations
 - heat map and, 49–50
 - regression problems, 60–62
 - Pearson’s, 47–49
 - targets and attributes, 44–47
- cross plots, 42–43
- cross-validation
 - out-of-sample error, 168–172
 - regression, 182–183

D

- data frames, 37–38
- data sets
 - examples, 24
 - instances, 24
 - items to check, 27–28
 - labels, 25
 - observations, 24
 - points, 7–8
 - problems, 24–28
 - shape, 29–32
 - size, 29–32
 - statistical summaries, 32–35
 - unique ID, 25
 - user ID, 25
- deciles, 34
- decision trees, binary, 9–10
 - bagging, 11
- degree of freedom, 86–87
- dependencies, chapters in book, 18–20
- dependent variables, 26

E

- ElasticNet package, 128–129, 131–132, 181–191

ensemble methods, 1, 20, 211–212

bagged decision trees, 4

base learners, 9–11

binary decision trees, 9–10

bagging, 11

boosted decision trees, 4

multiclass classification problems,
302–314

penalized linear regression methods
and, 124

penalized linear regression methods
comparison, 11–13

Random Forests, 4

speed, 11

ensemble models

binary classification problems,
284–302

non-numeric attributes

coded variables, 278, 282–284

gradient boosting regression,
278–282

random forest regression,
275–278

ensemble packages, 255–256

random forest model, 256–270

errors, out-of-sample, 80

F

factor variables, 26

predictions, 50–62

false negatives, 92

false positives, 92

feature engineering, 7, 17–18, 76

feature extraction, 17–18

feature selection, 7

features, 25

function approximation and, 76

forward stepwise regression, 102

LARS and, 132–144

overfitting and, 103–108

function approximation, 1, 76,
124–125

performance, 78–79

training data, 76–78

G

Glmnet, 132, 144–145

initialization, 146–151

iterating, 146–151

LARS comparison, 145–146

gradient boosting, 236–239, 256–262,
291–298

classifier performance, 298–302,
307–311

multivariable problems and,
244–246

parameter settings, 239

performance, 240–243

predictive models and, 240

random forest model base learners,
311–314

GradientBoostingRegressor, 263–267

model performance, 269–270

regression model implementation,
267–269

H

heat map, correlations, 49–50

regression problems, 60–62

I

importance, 138

independent variables, 26

inputs, 11, 26

K

KNNs (k nearest neighbors), 4

L

labels, 16. *See also* dependent variables;

outcomes; responses; targets

attributes, relationship visualization,
42–49

categorical, classification problems,
27

data sets, 25

function approximation and, 76

numeric, regression problems, 27

LARS (least-angle regression), 132

- forward stepwise regression and, 132–144
- Glmnet comparison, 145–146
- model selection, 139–142
 - cross-validation in Python Code, 142–143
 - errors on cross-validation fold, 143
 - practical considerations, 143–144
- Lasso penalty, 129–131
- lasso training, data sets, 173–176
- linear algorithms *versus* nonlinear, 87–88
- linear methods
 - nonlinear problems and, 156–158
 - non-numeric attributes, 158–163
- linear models, penalized linear regression and, 124
- linear regression, 1
 - model training, 126–132
 - numeric input and, classification problems, 151–155
 - penalized linear regression methods, 1, 124–132
- logistic regression, 1, 4, 155

M

- MACD (moving average convergence divergence), 17
- machine learning, problem
 - formulation, 15–17
- MAE (mean absolute error), 78–79, 88
- mean, Pandas, 39
- misclassification errors, 96
- mixture model, 81
- models
 - inputs, 11
 - LARS and, 136–138
- MSE (mean squared error), 78–79, 88
- multiclass classification problems, 68–73, 78, 204–209
 - algorithm comparison, 314–315
 - class imbalances, 305–307
 - ensemble methods, 302–314
- multivariable regression, 167–168

- bagging and, 231–235
- gradient boosting and, 244–246
- model building, 168–172
- testing model, 168–172

N

- n-fold cross-validation, 100
- nonlinear algorithms, *versus* linear, 87–88
- nonlinear problems, linear methods and, 156–158
- non-numeric attributes, linear methods and, 158–163
- normalization, box plots and, 55
- notation, predictors, 77
- numeric values, assigning to binary labels, 152–154
- numeric variables, 26, 77
 - regression problems, 27

O

- OLS (ordinary least squares), 7, 101, 121
 - coefficient penalties, 127–128
 - L1 norm, 129
 - Manhattan length, 129
- outcomes, 26
 - function approximation and, 76
- outliers, quantile-quantile plot, 35–37
- out-of-sample errors, 80
 - cross-validation and, 168–172
- overfitting
 - binary decision trees, 221–225
 - forward stepwise regression and, 103–108
 - ridge regression and, 110–119

P

- packages
 - ElasticNet, 181–191
 - penalized linear regression methods, 166–167
- Pandas, 37–39

- parallel coordinates plots, 40–42, 64–66
 - regression problems, 56–60
 - Pearson's correlation, 47–49
 - penalized linear regression methods, 1, 20, 121
 - binary classification, 181–191
 - classification problems, 151–155
 - coefficient estimation, 122
 - coefficient penalized regression, 111
 - ensemble methods and, 124
 - ensemble methods comparison, 11–13
 - evaluation speed, 123
 - function approximation and, 124
 - Glmnet, 144–145
 - initialization, 146–151
 - iterating, 146–151
 - LARS comparison, 145–146
 - linear models and, 124
 - linear regression regulation, 124–132
 - multiclass classification, 204–209
 - OLS (ordinary least squares) and, 7
 - packages, 166–167
 - reliable performance, 123
 - sparse solution, 123
 - speed, 11
 - variable importance information, 122–123
 - percentiles, 34
 - plots
 - box and whisker, 54–55
 - cross plots, 42–43
 - parallel coordinates, 40–42
 - quantile-quantile, 35–37
 - scatter plots, 42
 - points, data sets, 7–8
 - pred() function, 79
 - predictions
 - attributes and, 3
 - binary decision trees, 212–213
 - factor variables and, 50–62
 - real-valued, 62–68
 - wine taste, 168–172
 - predictive models
 - building, 13–18
 - feature engineering, 7, 17–18
 - feature extraction, 17–18
 - feature selection, 7
 - function approximation, 76
 - performance, 78–79
 - training data, 76–78
 - gradient boosting and, 240
 - labels, 16
 - mathematical description, 19
 - performance factors, 86–87
 - performance measures, 88–99
 - targets, 14
 - trained, 25
 - performance evaluation, 18
 - predictors, 25
 - function approximation and, 76
 - notation, 77
 - problem formulation, 15–17
- ## Q
- quantiles, Pandas, 39
 - quantil-quantile plot, 35–37
 - quartiles, 34
 - quintiles, 34
- ## R
- random forest model, 256–270
 - base learners, gradient boosting and, 311–314
 - classification, 302–305
 - classifier performance, 291
 - random forests, 212
 - bagging and, 247–250
 - performance and, 251–252
 - RandomForestRegressor object, 256–262
 - real-valued attributes, 77
 - regression
 - penalized linear regression, 121
 - ridge regression, 121
 - step-wise, 121
 - regression problems
 - correlation heat map, 60–62

- numeric variables, 27
- parallel coordinates, 56–60
- regressors, function approximation
 - and, 76
- relationships
 - attributes/labels, visualization, 42–49
 - variable, 56–60
- reliable performance, 123
- residuals, 137
 - attributes times residuals, 197
- responses, 26
- ridge regression, 102, 121
 - overfitting and, 110–119
- RMSE (root MSE), 88
- ROC (receiver operating curves), 88, 183
- RSI (relative strength index), 17

S

- scatter plots, 42
- scikit-learn packages, 166
- simple models, compared to complex models, 82–86
- sklearn.linear_model, 166
- sparse solution, 123
- squares of attributes, 197
- statistics, data sets, 32–35
- stepwise regression, 121
- stratified sampling, 37, 306
- summaries
 - data sets, 32–35
 - Pandas, 38–39
- supervised learning, 1
- SVMs (support vector machines), 4

T

- targets, 14, 26
 - attributes, correlation, 44–47

- binary classification problem, 78
- function approximation and, 76
- multiclass classification problem, 78
- trained models, 25
 - linear, 126–132
 - performance evaluation, 18
- training
 - binary decision trees, 214–217
 - tree training, 218–221
- training data, 76–78
 - deployment and, 172–181
- tree training, 218–221

U

- user ID, 25

V

- validation, cross-validation, out-of-sample errors, 168–172
- variable importance information, 122–123
- variables
 - categorical, 19, 26
 - classification problems, 27
 - statistical characterization, 37
 - creating from old, 178–181
 - factor, 26
 - numeric, 26
 - regression problems, 27
 - relationships, 56–60
- variance
 - versus* bias, 229–231
 - Pandas, 39
- visualization
 - attributes/labels relationship, 42–49
 - parallel coordinates plots, 40–42
 - variable relationships, 56–60