Index

A

abstraction, 6, 8, 62, 63

accuracy, of linear regression, 230–232

activation functions

  hyperbolic tangent function, 278–279

  identity function, 276

  rectified linear unit function, 277

  sigmoid function. *See* sigmoid function

  threshold/step function, 276

active learning

  heuristics, 305–306

  query strategies, 306

AdaBoost. *See* adaptive boosting

ADALINE network model. *See* adaptive linear neural element network model

adaptive boosting, 86, 311

adaptive linear neural element (ADALINE) network model, 285–286 agglomerative hierarchical clustering, 258–259

AI. *See* artificial intelligence (AI)

Aibo, 10

Alpha error, 141

AlphaGo program, 2, 29

alternate hypothesis, 141

ANN. *See* artificial neural network (ANN)

anomaly checking, clustering, 244

anti-monotone property of support measure, 265

Apriori algorithm, for association rule learning, 264–265, 309

Apriori principle rules,
265–268

area of property, 217, 227

area under curve (AUC) value, 80–81

artificial intelligence (AI), 1, 2, 243

artificial neural network (ANN), 273

- adaptive linear neural element network model, 285–286

- backpropagation, 292–296

- competitive network, 289

direction of signal flow, 291

McCulloch–Pitts neural model, 279–281

multi-layer feed forward network, 288–289

number of layers, 290–291

number of nodes in layers, 291

recurrent neural networks, 289–290

Rosenblatt’s perceptron. *See* Rosenblatt’s perceptron

single-layer feed forward network, 287–288

structure of, 275

weight of interconnection between neurons, 291–292

artificial neurons, 273, 274–275, 287

association analysis, 16, 242, 261

application of, 261

itemset, 262

support count, 262

association rule, 262–264. *See also* association analysis

Apriori algorithm for, 264–265

frequent itemset, 265, 267

strengths and weaknesses, 268

strong rules, 265, 267

association rule learning algorithm, 308–309

attributes, 35, 36–37

AUC value. *See* area under curve value

Auto MPG data set, 36, 326, 366

box plot of, 43

histogram, 46–47

‘horsepower’, 38–39

mean versus median for, 38

scatter plot, 51

autoencoders, 304–305

axon, 274

B

backpropagation algorithm, 292–296

backpropagation networks, 278, 294

backward phase, epochs, 292

backward stepwise selection, 232

bagging. *See* bootstrap aggregation

banking industry, machine learning in, 20

Bayes optimal classifier, 156–157

Bayes rule. *See* Bayes' theorem

Bayes' theorem, 2, 121–122, 150–151, 158

concept learning and, 154–157

likelihood, 152

posterior probability, 151–152

prior knowledge, 151

Bayesian Belief network, 165–166, 171

conditional independence, 166–170

independence, 166–170

in machine learning, 170

Bayesian classifiers, 149

Bayesian concept learning. *See also* Bayes' theorem; Bayesian Belief network

brute-force Bayesian algorithm, 154–156

consistent learners, 156

methods, 148, 149–150

Naïve Bayes classifier. *See* Naïve Bayes classifiers

optimal classification, 156–157

Bayesian interpretation, 119

Bernoulli distributions, 127

best linear unbiased estimator (BLUE), 229

best subset selection, 232

Beta error, 141

bias, 63

bias-variance trade-off, 73–74

big data, 117

binary sigmoid function, 277–278

binomial distribution, 127–128

bioinformatics, 170, 209

biological neural network, 273, 274. *See also* artificial neural network (ANN)

biological neuron, 273–274

biplot function, 342

bipolar sigmoid function, 278

bivariate random variables, 134–135

BLUE. *See* best linear unbiased estimator (BLUE) boosting, 86, 310–311

bootstrap aggregation, 86, 310

bootstrap sampling, 70, 71, 335, 375–376

box and whisker plot. *See* box plots

box plots, 41–43, 369–370

  Auto MPG attributes, 43

  cylinders, 43–44

  data exploration, 329–330

  displacement, 44–45

  model year, 45

  origin, 44

branch and bound search, decision tree, 190–191

branch node, 187

brute-force approach, 266

brute-force Bayesian learning algorithm, 154–156

btissue data set, 343, 344

C

candidate hypothesis, 149, 152

capping, 54

caret package, 334, 347

categorical data, 33

exploring, 47–49

nominal data, 33–34

ordinal data, 34

categorical distribution, 129

cdf. *See* cumulative distribution function (cdf)

central limit theorem, 132, 138

central nervous system (CNS), 273–274

central tendency, 37–39

centroids, 249

chain rule, 120, 166

chi-square test, 234

class, 14, 178

class package, 346

classification algorithms, 180

decision tree. *See* decision tree

k-nearest neighbour, 181–186

random forest model, 199–201

support vector machines, 201–209

classification learning steps, 179–180

algorithm selection, 180

data pre-processing, 180

definition of training data set, 180

evaluation with test data set, 180

identification of required data, 179

problem identification, 179

training, 180

classification model, 66, 177–178, 182

classification phase, bootstrap aggregation, 310

classification problem, 12

cluster centroids, recomputing, 250–254

cluster package, 338, 349

clustering, 16, 242, 255, 377

anomaly checking, 244

customer segmentation, 243

data mining, 244

- of data set, 251
- different methods, 246–247
- external evaluation, 84
- initial centroids, 252
- internal evaluation, 82–84
- as machine learning task, 244–246
- partitioning-based. *See* partitioning-based clustering
- text data mining, 243

CNS. *See* central nervous system (CNS)

competitive network, 289

Comprehensive R Archive Network (CRAN), 315

computational complexity, 306

concept learning, 150, 154–157

conditional distributions, 136–137

conditional independence, 166–170

conditional probability, 120–121, 165

confusion matrix, 76

confusionMatrix function, 336

consistent learners, 156

construct frequency table, 162

contains (), 324

contingency table. *See* two-way cross-tabulations

continuous numeric features, 164–165

continuous random variables, 125–126

  mean and variance, 126

  uniform distribution, 130–131

converging connection, 169–170

convex hull, 206

correlation, 137–138

correlation-based similarity measure, 106

cosine similarity, 109, 110

cost function, 64

covariance, 137–138

CPython, 21

CRAN. *See* Comprehensive R Archive Network (CRAN)

cross-tab. *See* two-way cross-tabulations

cross-validation, 71

cumulative distribution function (cdf), 123, 124, 126 cumulative probability, 161

curve linear negative slope, 220–221

curve linear positive slope, 219–220

customer segmentation, clustering, 243

D

data, 32, 35–37

  categorical, 32

  collection, errors in, 53

  dictionary, 35

  interval, 34

  nominal, 33–34

  ordinal, 34

  qualitative, 33–34

  quantitative, 34–35

  ratio, 34

data dispersion, 39–40

data exploration

  data pre-processing, 332–334

  plots for, 329–332, 368–371

statistical functions for, 326–329, 365–368

data frame, 319

data handling commands, 323

data holdout, 374

data input, 6, 62

data manipulation commands, 324–325

data matrix, 101

data mining, 244

data pre-processing, 56, 180, 332–334, 372

capping of values, 373

dimensionality reduction, 56

feature subset selection, 56–57

imputing standard values, 373

outliers and missing values, 372–373

data quality, 53

data remediation, 53

handling outliers, 54

missing values, 54–55

data set, 32, 150

Auto MPG, 36

features, 92

five-dimensional, 92

data spread, 39

data dispersion, 39–40

data value position, 40–41

data types

mathematical operations on, 322–323

Python, 357–358

R language, 318–319

data value position, 40–41

datasets, 369

DBSCAN, 260–261

decision node, 187

decision theory, 140

decision tree, 14, 186–187

algorithm for, 197

application, 198–199

avoiding overfitting in, 197–198

branch and bound search, 190–191

building, 188–190

entropy of, 191–196

example, 188

exhaustive search, 190

information gain, 192–197

output variable, 187

post-pruning, 197

pre-pruning, 197, 198

pruning of, 197

strengths, 198

structure, 187

types of nodes, 187

weaknesses, 198

decision tree classifier, 347, 387–388

delta rule, 286

dendrites, 274

dendrogram, 258

density-based clustering, 260–261

dependent variable, 216–217, 222, 227–229, 234

descriptive model, 16, 66–67

digital neurons, 273

dimensionality reduction, 56, 232

discrete bivariate random variable, 135, 136

discrete distribution, 129

discrete random variable, 123–125

distance-based clustering, 16

distance-based similarity measure, 106–110

distribution function, 123

diverging connection, 169

divisive hierarchical clustering, 258–259

document-term matrix, 98

double-sided exponential distribution, 134

dplyr package, 324, 339

dummy code categorical variables, 339–340, 379

dummy encoding, 129

E

e1071 package, 345, 346

eager learning, 71

Eclat algorithm, 309

eigenvalues, 101, 102

eigenvectors, 101, 102

elastic net, 311

elbow method, 249

  initial centroids, 249–250

  recomputing cluster centroids, 250–254

elbow point, 249, 250

embedded approach, 112

encoding categorical (nominal) variables, 95–97

encoding categorical (ordinal) variables, 97, 340, 380–381

ends_with (), 324

ensemble learning algorithms, 309–311

ensembling, 85, 86, 199

entropy, of decision tree, 191–196

epochs, 292

  backward phase, 292

  forward phase, 292

error(s)

in data collection, 53

due to bias, 73

due to variance, 73–75

error function. *See* cost function

error rate, 77

Euclidean distance, 106, 183, 250–251, 307

Euclidean space, 100

evaluation criterion, 110

exclusive-OR (XOR) circuit, 279

exhaustive search, decision tree, 190

expected error reduction, 306

expected model change, 306

expert system, 11

F

factor, 319

feature, 92

distance measures between, 108

entropy, 106

n-dimensional data set, 107

feature construction, 93, 94–95

dummy coding categorical (nominal) variables, 339–340

encoding categorical (nominal) variables, 95–97

encoding categorical (ordinal) variables, 97, 340

text-specific feature construction, 97–99

transforming numeric (continuous) features, 97, 341

feature discovery, 93

feature engineering, 93

feature extraction. *See* feature extraction

feature subset selection. *See* feature subset selection

feature extraction, 93, 99

linear discriminant analysis, 102, 343–344

principal component analysis, 100–101, 341–342

singular value decomposition, 101–102, 342–343

feature redundancy, 105–110

feature relevance, 104–106

feature selection. *See* feature subset selection

feature subset selection, 56–57, 93, 102, 344–345 approaches, 111–112

feature redundancy, 105–110

feature relevance, 104–106

high-dimensional data, 103–104

process, 110–111

feature transformation, 93

feature construction. *See* feature construction

feature extraction. *See* feature extraction

feature vectors, 100

feed forward, 287

filter approach, 111

F-measure, 79–80

for loop, 320

forward phase, epochs, 292

forward stepwise selection, 232

for–while loops, Python, 358–359

foundation rules, 119–120

fraud detection, 29

frequency table, 161, 162

frequent itemset, 265

frequentist interpretation, 119

FSelector package, 344

full batch gradient descent, 295

G

Gaussian (normal) distribution, 131–133

Gaussian function, 307

Gaussian radial filter, 308

Gaussian RBF kernel, 208

Gauss–Markov theorem, 229

GBM. *See* gradient boosting machines (GBM)

generalization, 6, 9, 62

generation versus recognition, 303

ggplot2 library functions, 329

glial cells, 274

Go board game, 2

Google, 2, 29

Google Brain, 2, 29

GPU. *See* graphics processing unit (GPU) gradient boosting machines (GBM), 311

gradient descent, 292

graphics processing unit (GPU), 296

H

Hamming distance, 107

healthcare, machine learning in, 21

heteroskedasticity, 229–230

hierarchical clustering, 258

agglomerative, 258–259

dendrogram representation, 260

distance measure, 259–260

divisive, 258–259

high-dimensional data set, 103–104

histogram, 45–47, 331, 370–371

Auto MPG attributes, 46–47

box plot and, 45

shapes, 46

holdout method, 67–68, 334

homogeneous group, 246

horsepower attribute, 38–39, 55, 328

human detection, 62

human learning, 4, 7, 62

  under expert guidance, 4–5

  knowledge gained from experts, 5

  by self, 5

  types of, 4–5

hybrid approach, 112

hybrid recommender system, 163–164

hyperbolic tangent function, 278–279

hyperplane, 201, 202–203

hypothesis testing, 140–142

I

IBM, 1, 2, 22, 28, 177

ICA. *See* independent component analysis (ICA)

ICU. *See* intensive care unit (ICU)

identification of required data, 179

identity function, 276

if-else statement, 320–321, 359

imputation, 54, 55

incorrect sample set selection, 53

incremental gradient descent, 295

independence, 166–170. *See also* conditional independence

independent component analysis (ICA), 303–304

independent variables, 216, 227–230

information gain, of decision tree, 192–197

initial centroids, 249–250

instance-based learning, 306–308

insurance industry, machine learning in, 20

intensive care unit (ICU), 176

intercept, interpretation of, 224–225

interdependent, 118

internal node, 187

interval data, 34

iris data set, 15, 329–330, 342, 369–371

irrelevant variables, 231

itemset, 262, 265–266

J

Jaccard distance, 107

Jaccard index/coefficient, 107–108

joint cumulative distribution function, 135

joint distribution, 120, 135

joint probability, 120, 165, 166, 167

joint probability density functions, 136

joint probability mass functions, 135–136

Julia programming language, 23

K

Kappa value, 77

kernel trick, 207–208

kernels, 207

k-fold cross-validation method, 68–70, 335, 374–375

k-means algorithm, 67, 247–255, 349

appropriate number of clusters, 249

elbow method, 249–254

strengths and weaknesses, 248

k-medoids algorithm, 255–257

k-nearest neighbour (*k*NN) algorithm, 14

application, 186

category of lazy learner, 185–186

Euclidean distance, 183

strengths, 186

student data set, 181–183

weaknesses, 186

k-nearest neighbour (*k*NN) classifier, 346, 386–387

*k*NN algorithm. *See* *k*-nearest neighbour (*k*NN) algorithm

knowledge, 118

knowledge discovery, 16

L

L_1 norm, 107

L_2 norm, 107

lab schedule, machine learning in, 353–354

label, 12, 68, 176

labelled input data, 182

labelled training data, 12, 176

Laplace distribution, 134

lasso regression, 231

layers, neural network, 290–291

lazy learning, 71

LDA. *See* linear discriminant analysis (LDA)

leaf node, 187

learning algorithm, 180

learning process of machines, 61–62

learning rate, 296

least mean square (LMS), 286

least squares method, 2

leave-one-out cross-validation (LOOCV), 68, 70

level of significance, 141

likelihood, 152, 170

linear discriminant analysis (LDA), 102, 343–344, 384–385

linear kernel, 208

linear negative slope, 220

linear positive slope, 219

linear regression model, improving accuracy of, 230

- dimensionality reduction, 232

shrinkage (regularization) approach, 231

subset selection, 231–232

linearly separable data, 206

list, 319

LMS. *See* least mean square (LMS)

logistic regression, 233–236

logit regression. *See* logistic regression

LOOCV. *See* leave-one-out cross-validation (LOOCV)

loops, 320–321

loss function, 64

M

machine learning (ML), 1, 7, 29

abstraction, 6, 8

activities, 30–32

algorithms, 14

applications of, 20–21

in banking industry, 20

data. *See* data

data input, 6, 8

definition, 5–6

evolution of, 2–3

formalism, 9

foundation of, 2

generalization, 6, 9

in healthcare, 21

in insurance industry, 21

issues, 23

languages/tools, 21–23

problem solving, 9–10

problems not using, 20

process, 6

reinforcement learning, 17–18, 19

supervised learning, 11–15, 19

types, 11–20

unsupervised learning, 16–17, 19

Manhattan distance, 107

MAP hypothesis. *See* maximum a posteriori (MAP) hypothesis

margin, 203

marginal distribution, 120

market basket analysis, 261

Markov chain Monte Carlo (MCMC), 142

MASS package, 343

matches (), 324

mathematical operations on data types, 322–323

MATLAB (matrix laboratory), 22

matplotlib, 368

matrix, 318–319

maximum a posteriori (MAP) hypothesis, 152, 156, 171

maximum likelihood estimation (MLE), 236

maximum likelihood (ML) hypothesis, 152

maximum margin hyperplane (MMH), 205

  linearly separable data, 206

  non-linearly separable data, 207

  support vectors, 206

maximum point of curves, 226–227

McCulloch–Pitts neural model, 279–281

MCMC. *See* Markov chain Monte Carlo

mean, 37, 38

mean of random variable, 126, 128, 131

median, 37, 38

memory-based learning, 306–308

merger points, clusters, 258

minimum marginal hyperplane (MMH), 306

minimum point of curves, 226–227

Minkowski distance, 107

missing values, 54

estimating, 55

imputing, 55

records elimination, 54

mixed bivariate random variable, 135

ML. *See* machine learning (ML)

MLE. *See* maximum likelihood estimation (MLE)

MMH. *See* maximum margin hyperplane (MMH); minimum marginal hyperplane (MMH)

mode, 34, 49, 55

model, 8

classification, 75–76

definition, 63

descriptive, 66–67

evaluating performance of, 75–84

improving performance of, 85–86

predictive, 65–66

representation and interpretability, 72–75

selecting, 63–67

sensitivity of, 78

specificity of, 78–79

training, 67–72

model accuracy, 76

model parameter tuning, 85

model training, 63

bootstrap sampling, 335

classification, supervised learning, 336–337

clustering, 338–339

holdout, 334

k-fold cross-validation, 335

regression, supervised learning, 337–338

Monte Carlo approximation, 142

Monte Carlo integration, 142

multi layer feed forward network, 394

multicollinearity, 229

multi-layer feed forward network, 288–289, 292

multi-layer feedforward neural network, 350, 352

multi-layer perceptron, 284–285

multinomial distribution, 128–129

multinoulli distribution, 128–129

multiple linear regression, 227–228, 349

heteroskedasticity, 229–230

multicollinearity, 229

multiple random variables

bivariate random variables, 134–135

conditional distributions, 136–137

covariance and correlation, 137–138

joint distribution functions, 135

joint probability density functions, 136

joint probability mass functions, 135–136

mutate function, 339

mutual information, 105

N

Naïve Bayes classifiers, 171, 346, 386

applications, 163–164

assumption, 167

benefits, 159

continuous numeric features, 164–165

parameter estimation
for, 158

principles, 158

problem with, 161–163

steps, 161

strengths and weaknesses, 159, 160

training data for, 160

naiveBayes function, 346

n-dimensional data set, 92

nerve cell, 273

nervous system, 273

nesting functions, 324

network security, 149

neural network, 302, 392–395. *See also* artificial neural network (ANN)

- multi-layer feedforward, 350, 352
- single-layer feedforward, 350, 351

neuralnet function, 350

neurolab, 392

neurons, 273, 274

‘No Free Lunch’ theorem, 65

nodes in layers, 291

noise-free training data, 156

nominal data, 33–34

non-linearly separable data, 207

normal random variable, 131–133

null hypothesis, 141

numerical data, 34

box plots, 41–45

central tendency, 37–39

data dispersion, 39–40

data spread, 39

data value position, 40–41

exploring, 37–41

histogram, 45–47

interval data, 34

plotting, 41–47

ratio data, 34–35

numpy library, 367

O

objective function, 64

OLS. *See* ordinary least squares (OLS)

one-hot encoding, 129

one_of (), 324

online sentiment analysis, 164

OOB error. *See* out-of-bag (OOB) error

ordinal data, 34

ordinary least squares (OLS), 223, 226

outliers, 53, 54

out-of-bag (OOB) error, 200

overfitting, 73, 197–198

P

PAM algorithm. *See* partitioning around medoids (PAM) algorithm

pandas library, 361, 378

partial regression coefficients, 227

partitioning around medoids (PAM) algorithm, 256–257

partitioning-based clustering, 247

- *k*-means algorithm, 247–255
- *k*-medoids algorithm, 255–257

pattern discovery, 16

patterns, 15

PCA. *See* principal component analysis (PCA)

pdf. *See* probability density function (pdf)

Pearson correlation coefficient, 106

peripheral nervous system, 273

piping, 324

plyr package, 324

pmf. *See* probability mass function (pmf)

Poisson distribution, 129

polynomial kernel, 208

polynomial regression model, 232–233

posterior probability, 151–152, 154, 155, 156, 158, 171

prcomp function, 341

precision, 79

prediction, 230

predictive models, 65–66

predictors, 216

preparation, machine learning system, 30

price of property, 217, 227

principal component analysis (PCA), 100–101, 303, 341–342, 381–383

prior knowledge, 151, 165

prior probability, 158, 170

probabilistic classifications, 158

probabilistic inference process, 170

probability

  posterior, 151–152, 154–156

  prior, 151, 158, 170

  revised, 168

  rules, 148

  unconditional, 168

probability density function (pdf), 125, 126

probability mass function (pmf), 123, 124, 127

probability rule, 151, 162

probability theory, 117

  Bayes rule, 121–122

  Bayesian interpretation, 119

  central limit theorem, 138

  chain rule, 120

  concept, 118–119

  conditional, 120–121

  of correct decisions, 142

foundation rules, 119–120

frequentist interpretation, 119

hypothesis testing, 140–142

joint, 120

Monte Carlo approximation, 142

random variables. *See* random variables

sampling distributions, 138–140

sum rule, 120, 125

type I and type II errors, 141

union of two events, 120

problem identification, 179

product rule, 120

pruning of decision tree, 197

purity, cluster algorithms, 84

Python, 21–22

basic commands, 355–357

bootstrap sampling, 375–376

classification model, 376–377

clustering, 377

data exploration. *See* data exploration

data handling commands, 361–365

data holdout, 374

data pre-processing, 372–373

data types, 357–358

feature construction, 378–381

feature extraction, 381–385

feature subset selection, 385

for–while loops, 358–359

if–else statement, 359

k-fold cross-validation, 374–375

machine learning lab using, 396–397

mathematical operations, 360–361

model training, 376

neural network, 392–395

purity, 377

regression model, 377

scripts, 356

sklearn framework, 376

supervised learning. *See* supervised learning

variables, 357–358

writing functions, 359–360

Python Anaconda, 355

Python Software Foundation, 21

Q

qualitative data. *See* categorical data

quantitative data. *See* numerical data

query by committee, 306

R

R language, 22

basic commands, 317–318

boxplot, 329–330

cluster, 338–339

data exploration. *See* data exploration

data types, 318–319

histogram, 331

installation, 315

loops, 320–321

mathematical operations on data types, 322–323

model training. *See* model training

modelling and evaluation, 334

scatterplot, 331–332

scripts management, 316

writing code in, 316

writing functions, 321

radial basis function (RBF), 307–308

radial basis function network (RFFN), 307

radial function, 308

random forest classifier, 347–348, 388

random forest model, 85, 199

application, 201

out-of-bag error in, 200

strengths, 200–201

weaknesses, 201

random numbers, 67

random sample, 139

random variables, 122

  Bernoulli, 127

  binomial, 127–128

  bivariate, 134–135

  continuous, 125–126

  discrete, 123–125

  domain of, 122

  multinomial and multinoulli, 128–129

  multiple. *See* multiple random variables

  normal, 131–133

  Poisson, 129

  standard normal, 132, 133, 138

  uniform, 130–131

randomForest function, 347

ratio data, 34

RBF. *See* radial basis function

recall, 79

receiver operating characteristic (ROC) curve, 80–81 recognition, generation versus, 303

record, 32

rectified linear unit (ReLU) function, 277

recurrent neural networks, 289–290

recursive partitioning, 188

regression, 12, 14–15, 377

  assumptions, 228–229

  common algorithms, 217

  example of, 216

  logistic, 233–235

  maximum likelihood estimation, 236

  multiple linear, 227–230

  polynomial regression model, 232–233

  simple linear. *See* simple linear regression

  supervised learning, 81–82

regularization algorithms, 311

reinforcement learning, 17–18, 19

ReLU function. *See* rectified linear unit (ReLU) function

remove outliers, 54

repeated holdout, 68

representation learning, 301–302

  active learning. *See* Active learning

  association rule learning algorithm, 308–309

  autoencoders, 304–305

  clustering forms, 305

  ensemble learning algorithms, 309–311

  generation versus recognition, 303

  independent component analysis, 303–304

  instance-based learning, 306–308

  multilayer perceptron, 303

  regularization algorithms, 311

  supervised neural networks, 303

  triangle types, 302

residual, 82, 222

revised probability, 168

RFFN. *See* radial basis function network (RFFN)

ridge regression, 231

risk prediction, 29

ROC curve. *See* receiver operating characteristic (ROC) curve

root node, 187

Rosenblatt's perceptron, 281–282

  class assignment, 283

  class separability, 284

  classification by decision boundary, 283

  classification with two decision lines, 285

  decision boundary, 282

  multi-layer perceptron, 284–285

rpart package, 347

R-squared, 82, 377

rule of total probability, 120

S

sampling distributions, 138–140

  mean and variance, 140

  with replacement, 139

  without replacement, 139

sampling theory, 70

SAS. *See* Statistical Analysis System (SAS) scatter plot, 49–51, 331–332, 371

scikit-learn, 22, 373, 374

scripts management, in R language, 316

semi-supervised learning, 176, 305

sensitivity of model, 78

serial connection, 169

set.seed function, 335

Shannon’s formula, 106

shrinkage (regularization) approach, 231

Sibyl, 29

sigmoid function, 277

- binary, 277–278

- bipolar, 278

sigmoid kernel, 208

signal flow direction, neural network, 291

silhouette coefficient, 83

silhouette width, 83–84, 378, 390

simple hypothesis, 141

- simple linear regression, 217–218, 349
  - error in, 221
  - example, 221–225
  - maximum and minimum point of curves, 226–227
  - no relationship graph, 221
  - ordinary least squares algorithm, 226
  - slopes, 218–221
- simple matching coefficient (SMC), 108–109
- Simple Random Sampling with Replacement (SRSWR), 70
- single-layer feed forward network, 287–288
- single-layer feedforward neural network, 350, 351
- single-valued real function, 122
- singular value decomposition (SVD), 101–102, 342–343, 383–384
- slopes, linear regression model, 218–219
  - curve linear negative slope, 220–221
  - curve linear positive slope, 219–220
  - linear negative slope, 220
  - linear positive slope, 219

SMC. *See* simple matching coefficient (SMC)

soma, 274

spam filtering, 163

specificity of model, 78–79

spine.csv, 338

spinem.csv, 338

split, clusters, 258

SPSS. *See* Statistical Package for the Social Sciences (SPSS)

Spyder (Scientific PYthon Development EnviRonment), 355

squares of the errors

  (SSE), 223

squashing function. *See* threshold activation function

SRSWR. *See* Simple Random Sampling with Replacement (SRSWR)

state space, 124

Statistical Analysis System (SAS), 22

Statistical Package for the Social Sciences (SPSS), 22–23

stats package, 341, 342

step function, 276

stepwise subset selection method, 232

stochastic gradient descent, 295

stopping criterion, 111

strong rules, 265

subset generation, 110

subset selection, linear regression model, 231–232, 385

  backward stepwise, 232

  best, 232

  forward stepwise, 232

sum of squared error (SSE), 248, 249, 251, 256

sum rule, 120, 125

summary commands, 326

summary function, 337

summation junction, 275

supervised learning, 11, 19, 29, 176. *See also* unsupervised learning

  bootstrap sampling, 70–71

  classification, 13–14, 75–81, 336–337, 345

  classification algorithms. *See* classification algorithms

  classification learning steps. *See* classification learning steps classification model, 177–178, 376–377, 386–389

decision tree classifier, 347

example of, 12, 176–177

holdout method, 67–68

k-fold cross-validation method, 68–70

kNN classifier, 346

lazy versus eager learner, 71

Naïve Bayes classifier, 346

random forest classifier, 347–348

regression, 14–15, 81–82, 337–338, 349, 389–390

SVM classifier, 348

unsupervised learning versus, 176, 242

support count, 262

support vector machines (SVM), 201

application, 209

classification using hyperplanes, 201–203

generalization error, 202

hard margin, 202

hyperplane, 201, 202–203

identifying correct hyperplane, 203–205

kernel trick, 207–208

margin, 203

maximum margin hyperplane, 205–207

strengths, 208

weaknesses, 208

support vectors, 202, 206

support-based pruning, 265

SVD. *See* singular value decomposition (SVD)

svd function, 342

SVM. *See* support vector machines (SVM)

SVM classifier, 348, 389

synapse, 274, 275

T

target function, 64

10-fold cross-validation, 68

term-document matrix, 98

test data, 13, 180, 182

text data mining, 243

text-based classification, 149, 163

text-specific feature construction, 97–99

threshold activation function, 275

threshold function, 276

TID list. *See* Transaction IDs (TID list)

total probability rule, 120

train function, 347

training data, 12, 151, 154, 176, 180, 182

‘training data is labelled’, 176

training, learning algorithm, 180

training phase, bootstrap aggregation, 310

Transaction IDs (TID list), 309

transforming numeric (continuous) features, 97, 341

triangle types, 302

two-way cross-tabulations, 51–52

type I error, 141

type II error, 141

U

uncertainty, 118

uncertainty sampling, 306

unconditional probability, 168

underfitting, 72

uniform distribution, 124, 125, 130–131

unstructured data, 92

unsupervised learning, 15–17, 19, 82–84, 105, 241, 349–350. *See also* supervised learning

application of, 242–243

clustering, 338–339, 377–378, 390–392

supervised learning versus, 242

V

validation, 111

validation data, 68

variable reduction, linear regression model, 232

variables, exploring relationship between, 49–52

variance, errors due to, 73–75

variance inflation factor (VIF), 229

variance of random variable, 126, 128, 131 variance reduction, 306

vector, 318

vector spaces, 100

vectorization process, 98

VIF. *See* variance inflation factor (VIF)

Voronoi diagram, 251

W

Waymo, 29

weight of interconnection between neurons, 291–292

while loop, 320

wrapper approach, 111, 112

writing functions, 321

X

XOR circuit. *See* exclusive-OR (XOR) circuit