A
activation functions (artificial neural networks), 94, 95–96
advertising software, 6–7
Aggregator, SpringXD and, 192
algorithms
assignments, 165–166
association rules learning, 123–124
decision trees, 47–48
Forgy method of initialization, 165
initialization, 165
k-means, 164–168
random partition method of initialization, 165
updating, 166
anonymity, data and, 26–27
Apache Spark. See Spark
Apriori algorithm, 123–124
Arff, converting to from CSV, 114
.arff files, LibSVM library, 154–155
artificial neural networks, 91
activation functions, 94, 95–96
back propagation, 98–99
connections, removing, 108
credit applications, 93
data center management, 93
data preparation, 99–100
HFT (high-frequency trading), 92–93
learning rate, 99
medical monitoring, 93–94
nodes, 108
perceptrons, 94–98, 103–105
robotics, 93
test data, increasing size, 108–109
Weka and, 100–109
association rules learning, 119–120
algorithms, 123–124
beer and diapers, 118–119
certainty, 121–122
correlation, 122
lift, 122
Mahout, 124–131
process, 122–123
support, 121
uses, 117–118
web usage mining, 118
attributes, decision trees, 55
axons, 92

B
back propagation (artificial neural networks), 98–99
batch processing
EMR (Elastic Map Reduce), 226–227
frequency and, 224–225
Hadoop, 225–226
  walk through, 227–233
Mahout, 226
MapReduce, 233–234
Pig, 226
process method, 225
quantity of data, 225
scheduling jobs, 273–274
Sqoop, 226
volume and, 224–225
walkthroughs, 227
Bayes’ Theorem, 73–75
Bayesian Networks, 69–70, 75–76
  base graph, 84
coding, 81–90
domain experts, 78–79
graph theory and, 70–71
Java APIs, 79
JavaBayes library, 82–83
network testing, 87–90
nodes, 78, 80, 85–86
planning, 79–81
probabilities
  assigning, 76–77, 86–87
  planning and, 80–81
probability theory, 72–73
project creation, 81–90
results calculation, 77–78
Beer and Diapers legend, 118–119
bias-variance dilemma, 3
Big Data, 223
  resources, 368
  Target stores and, 27–28
binary classification, support vector machines, 140–142
blogs, 370
Britney dilemma, 30–33
classification, support vector machines
  binary, 140–142
  confidence, 143
  linear classifiers, 142–144
  multiclass, 140–142
  Weka, 148–154
classifications, support vector machines
  linear classifiers, 144–146
  non-linear classifiers, 146–147
Clojure, 11
cloud-based services, data processing, 24–25
cloud-based storage, 25
clustering, 161–168
command-line method for clustering (Weka), 174–178
conditional probability, 72
country names, 33–34
credit applications, neural networks and, 93
creepy line of data privacy, 27–28
cross-validation method, calculating
  cluster datasets, 168
CSV (comma separated variables), 36–37
  converting to Arff, 114
  .csv files, LibSVM library and, 154–155
cultural norms, data and, 25–26
cycle of machine learning, 17–18

D
data
  downloading, Mahout, 124–125
firehose, 187
input data, 36–41
output data, 42
planning and, 19–20
real-time system, 188–189
data capture, 187
data center management, neural networks and, 93
data cleaning, 30–36
data files, Mahout, 126–129
data preparation (artificial neural networks), 99–100
data privacy, 25–28
data processing, 24–25
data quality, 28–30
data repositories
  Infochimps, 14
  Kaggle, 15
  UC Irvine Machine Learning Repository, 14
data science, resources, 368
data storage, 25
data team, 22–23
databases, 41
datasets
  clusters, 166–168
  resources, 369–370
  Weka, 100–102
dates/times, 35
decision making, resources, 369
decision trees, 46–60
dendrites, 92
development portion of machine learning, 21
domain knowledge, data team, 23
domains, Bayesian Networks, 78–79

E

e-commerce software, 7–8
elbow method, calculating cluster datasets, 167
Emacs text editor, 364–365
EMR (Elastic Map Reduce). See also MapReduce
  batch processing and, 226–227
  233–234
error handling, LibSVM, 153–154
ETL (extract, transform, load), existing data and, 247–250
experimentation, 42

F
File, SpringXD and, 191
Filters, SpringXD and, 192
firehose of data, 187
Forgy method of algorithm initialization, 165
format checks, 30
formats, date/time, 35
FP-Growth (Frequent Pattern Growth) algorithm, 124

G
gaming analytics software, 8–9
Gemfire, SpringXD and, 191
Gemfire Server, SpringXD and, 192
generational expectations, data and, 26
graph theory, 70–71
graphic design, data team, 23

H
Hadoop, 13
  batch processing and, 225–233
  coffee shop case, 256–272
downloading, 351–352
hashtags, 235–236
HDFS filesystem, 352
installation, 351–352
Mahout and, 132–133, 250–256
MapReduce, 236–247
process list, 353
R and, 342–347
resources, 368–369
SpringXD support, 235
Sqoop, 247–250
starting/stopping, 353
hash values, 27
hashtags
  Hadoop, 235–236
  MapReduce class, 236–247
HDFS, SpringXD and, 192
healthcare, software, 6
HFT (high-frequency trading), neural networks and, 92–93
HTTP, SpringXD and, 190
hyperplane, 142

I
ID3 (Iterative Dichotomiser 3) algorithm, 47
IDE (integrated development environment), 14
Infochimps, 14
input data
  CSV (comma separated variables), 36–37
  databases, 41
  images, 41
JSON (JavaScript Object Notation), 37–39
  raw text, 36
  spreadsheets, 40–41
XML (extensible markup language), 39–40
YAML (YAML Ain’t Markup Language), 39
input sources (SpringXD), 190–191
Internet of things, 9–10

J
Java
  APIs, Bayesian Networks, 79
  LibSVM library, 154–159
  neural networks, 109–115
  Spark and, 276, 291–294
  version, 11
JavaBayes, 79
Jayes, 79
JDBC, SpringXD and, 191
JMS, SpringXD and, 191
JSON (JavaScript Object Notation), 37–39
  field Extractor, SpringXD and, 192
  field value, SpringXD and, 192
JVM (Java Virtual Machine), languages and, 10

K
Kaggle, 15
k-means algorithm
  assignments, 165–166
  clustering and, 164–166
  Weka, 168–186
  initialization, 165
  updating, 166

L
languages
  Clojure, 11
  Matlab, 10
  Python, 10
  R, 10
  Ruby, 11
  Scala, 10–11
learning rate, 99
LibSVM library
  .arff files and, 154–155
  .csv files, 154–155
  error handling, 153–154
  installation, 147–148
  Java, 154–159
  predicting with data, 158–159
  project setup, 155–158
  training with data, 158–159
linear classifiers, support vector machines, 142–144, 146–147
Log, SpringXD and, 191
log file analysis, 7

M
machine clusters, data processing, 24
machine learning
  algorithm types, 3–4
cycle, 17–18
description, 2
history, 1–2
humans and, 4
resources, 367
supervised learning, 3
unsupervised learning, 3–4
uses, 4–10
Machine Learning (Mitchell), 2
Mahout, 12
association rules learning, 124–131
batch processing and, 226
Hadoop and, 132–133, 250–256
results, 133–135
standalone mode, 131–132
Mail, SpringXD and, 190, 191
main method, clustering in Weka, 180
MapReduce
batch processing and, 233–234
file testing, 242–245
jar file, 242
job configuration, 241–242
mapper class, 237–240
project creation, 236–237
reducer class, 240–241
required fields, 237
Spark comparison, 285–288
SpringXD configuration, 245–246
streaming data testing, 246–247
marketing, Beer and Diapers legend, 119
MARS (multivariate adaptive regression splines) algorithm, 48
mathematics, data team, 22–23
Matlab, 10
medical monitoring, neural networks and, 93
medicine, software, 6
Mitchell, Tom M., 2
Machine Learning, 2
MLib (Machine Learning Library), 311–313
MQTT, SpringXD and, 191, 192
multiclass classification, support vector machines, 140–142

N-O
Nano text editor, 364
Netica, 79
network training, artificial neural networks, 105–107
neural networks, 91
Java, 109–115
neurons, 91–92
nodes
artificial neural networks, 108
Bayesian Networks, 78
decision trees, 48–49
non-linear classifiers, support vector machines, 146–147
output data, 42

P
perceptrons (artificial neural networks), 94–95
multilayer, 96–98
Weka, 103–105
physical storage, 25
Pig
batch processing and, 226
sales data mining, 263–272
planning aspect of machine learning, 19–20
presence checks, 28–29
probabilities, Bayesian Networks, 76–77
process of machine learning, 19–22
processors
sentiment analysis and, 217–221
SpringXD, 206–215
processors (SpringXD), 192
production portion of machine learning, 22
programming, data team, 23
project setup, LibSVM library, 155–158
Python, 11
Spark and, 276
Q-R
question, planning and, 18

R language, 10
Apriori algorithm, 333–336
data frames, 321
data loading, 323–324
Hadoop and, 342–347
installation, 315–316
Java access, 337–342
linear regression, 329–331
lists, 320–321
matrices, 319–320
packages, 322–323
plotting data, 324–327
R-Studio, installation, 317–318
sentiment analysis, 331–333
shell, 316
statistics, 327–328
variables, 318–319
vectors, 318–319
RabbitMQ, SpringXD and, 191, 192
random partition method of algorithm
initialization, 165
range checks, 30
raw text input, 36
real-time data system, 188
uses, 188–189
refining portion of machine learning, 22
reporting portion of machine learning, 21–22
resources
Big Data, 368
blogs, 370
data science, 368
datasets, 369–370
decision making, 369
Hadoop, 368–369
machine learning, 367
statistics, 368
tools, 370
visualizaton, 369
websites, 370
retail software, 7–8
robotics, neural networks and, 93
robotics software, 6
Ruby, 11
rule of thumb method, calculating cluster datasets, 167

S
salt values, 27
Samuel, Arthur, 2
Scala, 10–11
classes, 278
data types, 277–278
function calls, 278–279
if statements, 280
installation, 276–277
for loops, 279
operators, 279
packages, 277
Spark and, 276, 288–291
while loops, 279
scheduling, batch jobs, 273–274
sentiment analysis, 215–217
processor creation, 217–221
Sigmoid function, 95–96
silhouette method, calculating cluster datasets, 168
SimpleKMeans class, 168
sinks (SpringXD), 191–192
software
advertising, 6–7
e-commerce, 7–8
gaming analytics, 8–9
Hadoop, 13
healthcare, 6
IDE (integrated development environment), 14
Internet of things, 9–10
Java, version, 11
Mahout, 12
medicine, 6
retail, 7–8
robotics, 6
spam detection, 4–5
SpringXD, 13
stock trading, 5–6
voice recognition, 5
Weka toolkit, 12
spam detection software, 4–5
Spark, 275
data sources, 282
downloading, 280
installation, 280
Java and, 276, 291–294
Machine Learning Libraries, 311–313
MapReduce comparison, 285–288
monitor, 284–285
Python and, 276
Scala and, 276, 288–291
shell, starting, 281–282
standalone programs, 288–295
streaming, 305–311
testing, 282–284
SparkSQL, 295–305
Split, SpringXD and, 192
Splunk Server, SpringXD and, 192
spreadsheets, 40–41
SpringXD, 13, 187
application context, 211–212
code writing, 210–211
Hadoop support, 235
input sources, 190–191
installation, manual, 349
jar files, 212–214
Maven, 209–210
overview, 189
processors, 192, 206–215
project creation, 208–209
project deployment, 214
sample data, 198
sinks, 191–192
startup, 349
stream creation, 350
streams, 190, 199–202
taps, 221–222
Twitter data and, 193–198, 202–205
Twitter key, 350
xd-shell script, 198–199
Sqoop, 226, 247–250
statistics
data team, 22–23
resources, 368
stock trading software, 5–6
streaming, Spark and, 305–311
supervised learning, 3
support vector machines, 139–154

T
TAI (Temps Atomique International), 35
Tail, SpringXD and, 191
Target stores, 7
Big Data and, 27–28
TCP, SpringXD and, 190, 191
Tesco Clubcard, 7, 28
test data, artificial neural networks, increasing size, 108–109
testing portion of machine learning, 21
text editors for Unix, 363–365
time, SpringXD and, 191
times. See dates/times tools, 370
Transform, SpringXD and, 192
Turing, Alan, 1–2
Twitter, SpringXD, 193–196
stream creation, 203–205
Twitter credentials, 202–203
Twitter API Developer Application, configuration, 194–196
Twitter Search, SpringXD and, 191
Twitter Stream, SpringXD and, 191
type checks, 29

U
UC Irvine Machine Learning Repository, 14
Unix commands
| (pipe symbol), 363
cat, 356–357
find, 362
grep, 357–358
head, 361
sort, 360
text editors, 363–365
uniq, 360–361
wc, 361
unsupervised learning, 3–4

V
variables, R, 318–319
vectors, R, 318–319
Vi text editor, 363–364
Vim text editor, 363–364
visualization, resources, 369
voice recognition software, 5

W
web usage mining, 118
websites, 370
Weka toolkit, 12
   artificial neural networks, 102–109
classification, 60–66
clustering, k-means algorithm, 168–186
coded method for clustering, 178–186
command-line method for clustering, 177–178
decision trees, 53–60
LibSVM, 147–148, 153–154
support vector machines, 147–154
workbench method for clustering, 169

X-Y-Z
xd-shell script, SpringXD, 198–199
XML (extensible markup language), 39–40
YAML (YAML Ain’t Markup Language), 39
YARN (Yet Another Resource Locator), 275