

# Index

## A

- Agglomerative hierarchical clustering, 89
- API, 33
  - get\_score, 18–22
  - GUI, 17
- ARMA, *see* Autoregressive moving-average (ARMA)
- AR model, *see* Autoregressive (AR) model
- Artificial neural network (ANN), 99
- Autoregressive (AR) model
  - parameters, 134–136
  - time series, 134
- Autoregressive moving-average (ARMA), 137–139
- Average linkage method, 91
- AWS Lambda, 169

## B

- Backpropagation network (BPN)
  - algorithm, 104–105
  - computer systems, 100
  - definition, 100
  - fetch-execute cycle, 100
  - generalized delta rule, 100
  - hidden layer weights, 102–104

- mapping network, 100
- output layer weights, 101, 104
- Basket trading, 97

## C

- Clique, 97
- Cloud Datastore by Google, 168–172, 174, 176–178
- Clustering
  - business owners, 77
  - centroid, radius, and diameter, 97
  - and classification, 78
  - distances
    - edit, 85–86
    - Euclidean, 83–84
    - general, 84
    - properties, 82
    - squared Euclidean, 84
  - document, 78
  - elbow method, 82
  - hierarchical (*see* Hierarchical clustering)
  - K-means, 78–81
  - machine learning algorithm, 98
  - similarity types, 87–88
  - wine-making industry, 77

## INDEX

Collaborative filtering, 52  
Complete linkage method, 91  
Correlogram, 129  
Curve fitting method, 68

## D

Decision tree  
    entropy, 59  
    good weather, 59  
    information gain, 60  
    parameter, 59  
    random forest classifier, 60–61  
Divisive hierarchical  
    clustering, 92  
DynamoDB, 169

## E

Edit distance  
    Levenshtein, 85  
    Needleman–Wunsch  
        algorithm, 86–87  
Elasticsearch (ES)  
    API, 33  
    connection\_class, 31–32  
    Kibana, 31  
    Logstash, 31  
Euclidean distance, 83–84  
Exponential smoothing, 124  
Extract, transform, and load (ETL)  
    API, 34  
    e-mail parsing, 40–42  
    ES (*see* Elasticsearch (ES))

in-memory database (*see*  
    In-memory database)  
MongoDB (*see* MongoDB)  
MySQL (*see* MySQL)  
Neo4j, 34  
Neo4j REST, 35  
topical crawling, 40, 42–48

## F

Fourier Transform, 140

## G

Gaussian distribution data, 127  
Google Cloud Datastore, 168–172,  
    174, 176–178

## H

Hadoop  
    combiner function, 147  
    class diagram, 148  
    interfaces, 158  
    MainBDAS class, 152–155  
    RootBDAS class, 147, 150  
    unit testing class, 157–158  
    WordCounterBDAS utility  
        class, 151–152  
HDFS file system, 159  
MapReduce design pattern  
    filtering, 160  
    joining, 161–163, 165–166  
    summarization, 159–160

MapReduce programming,  
145–146

partitioning function, 146

HDFS file system, 159

Hierarchical clustering

bottom-up approach, 89–90

centroid, radius, and  
diameter, 97

definition, 88

distance between clusters

average linkage method, 91

complete linkage method, 91

single linkage method, 90

graph theoretical approach, 97

top-down approach, 92–96

Holt-Winters model, 124–125

## I, J

Image recognition, 67

In-memory database, 35

Internet of Things (IoT), 179

## K

Kibana, 31

K-means clustering, 78–81

## L

Least square estimation, 68–69

Levenshtein distance, 85

Logistic regression, 69–70

Logstash, 31

## M

MA model, *see* Moving-average  
(MA) model

MapReduce programming,  
145–146

MongoDB

database object, 37

document database, 36

insert data, 38

mongoimport, 36

pandas, 38–39

pymongo, 37

remove data, 38

update data, 38

Moving-average (MA) model,  
131–133

Mutual information (MI), 56

MySQL

COMMIT, 28

database, 24

DELETE, 26–27

INSERT, 24–25

installation, 23–24

READ, 25–26

ROLL-BACK, 28–31

UPDATE, 27–28

## N

Naive Bayes classifier, 61–62

Nearest neighbor classifier, 64

Needleman-Wunsch

algorithm, 86–87

## INDEX

Neo4j, 34

Neo4j REST, 35

Neural networks

BPN (*see* Backpropagation network (BPN))

definition, 99

Hebb's postulate, 106

layers, 99

passenger load, 99

RNN, 113, 115–116, 118–119

TensorFlow, 106, 108–109, 111–112

## O

Object-oriented programming (OOP), 3–9, 11–12

Ordinary least squares (OLS), 68–69

## P, Q

Pearson correlation, 50–52

Permanent component, 125

Principal component analysis, 53–55

Python

API, 17–22

high-performance applications, 2

IoT, 1

microservice, 14–17

NLP, 13–14

OOP, 3–9, 11–12

R, 13

## R

Random forest classifier, 60–61

Recurrent neural network (RNN), 113, 115–116, 118–119

Regression, 68

and classification, 70

least square estimation, 68–69

logistic, 69–70

Resilient distributed data set (RDD), 167

RNN, *see* Recurrent neural network (RNN)

## S

Sample autocorrelation coefficients, 129

Sample autocorrelation function, 129

Seasonality, time series

airline passenger loads, 124

exponential smoothing, 124

Holt-Winters model, 124–125

permanent component, 125

removing

differencing, 126

filtering, 125–126

Semisupervised learning, 58

Sentiment analysis, 65–66

- Single linkage method, 90
  - Spark
    - advantage, 166
    - broadcast variable, 167
    - components, 166
    - lineage, 167
    - message-passing
      - interface, 167
    - partition, 167
    - RDD, 167
    - shared variable, 167
    - Spark Core, 168
    - word count program, 167
  - Squared Euclidean distance, 84
  - Stationary time series
    - autocorrelation and
      - correlogram, 129, 130
    - autocovariance, 129
    - description, 128
    - joint distribution, 128
  - Supervised learning
    - classifications, 57
    - dealing, categorical
      - data, 73–76
    - decision tree, 59–61
    - dimensionality reduction
      - investment banking, 50
      - mutual information (MI), 56
      - Pearson correlation, 50–52
      - principal component
        - analysis, 53–55
      - survey/factor analysis, 49
      - weighted average of
        - instruments, 50
    - image recognition, 67
    - Naive Bayes classifier, 61–62
    - nearest neighbor
      - classifier, 64
    - over-or under-predict
      - intentionally, 71–72
    - regression (*see* Regression)
    - semi, 58
    - sentiment analysis, 65–66
    - support vector
      - machine, 62–63
  - Support vector machine, 62–63
- ## T
- Topical crawling, 40
  - TensorFlow
    - logistic regression, 111–112
    - multilayer linear regression,
      - 108–109, 111
    - simple linear regression,
      - 106, 108
  - Time series
    - ARMA models, 137–139
    - AR model, 133
    - definition, 121
    - exceptional scenario, 141, 143
    - Fourier Transform, 140
    - MA model, 131–133
    - missing data, 143
    - SciPy, 130
    - seasonality, 124–126
    - stationary (*see* Stationary time series)

## INDEX

### Time series (*cont.*)

#### transformation

cyclic variation, 127

data distribution

normal, 127

irregular fluctuations, 128

seasonal effect additive, 127

variance stabilization, 126

trends, 121–122

curve fitting, 122

removing, 123–124

variation, 121

Topical crawling, 42–48

## **U, V, W, X, Y, Z**

Unsupervised learning, *see*

Clustering