

# Machine Learning Algorithms

A reference guide to popular algorithms for data science and machine learning



**Packt**>

[www.packt.com](http://www.packt.com)

By Giuseppe Bonaccorso

# Table of Contents

<b>Preface</b>	1
<b>Chapter 1: A Gentle Introduction to Machine Learning</b>	7
<b>Introduction - classic and adaptive machines</b>	7
<b>Only learning matters</b>	10
Supervised learning	11
Unsupervised learning	13
Reinforcement learning	15
<b>Beyond machine learning - deep learning and bio-inspired adaptive systems</b>	16
<b>Machine learning and big data</b>	18
<b>Further reading</b>	19
<b>Summary</b>	20
<b>Chapter 2: Important Elements in Machine Learning</b>	21
<b>Data formats</b>	21
Multiclass strategies	24
One-vs-all	24
One-vs-one	24
<b>Learnability</b>	25
Underfitting and overfitting	28
Error measures	29
PAC learning	31
<b>Statistical learning approaches</b>	33
MAP learning	35
Maximum-likelihood learning	35
<b>Elements of information theory</b>	40
<b>References</b>	43
<b>Summary</b>	43
<b>Chapter 3: Feature Selection and Feature Engineering</b>	45
<b>scikit-learn toy datasets</b>	45
<b>Creating training and test sets</b>	46
<b>Managing categorical data</b>	48
<b>Managing missing features</b>	51
<b>Data scaling and normalization</b>	52

<b>Feature selection and filtering</b>	55
<b>Principal component analysis</b>	57
Non-negative matrix factorization	63
Sparse PCA	65
Kernel PCA	66
<b>Atom extraction and dictionary learning</b>	69
<b>References</b>	71
<b>Summary</b>	71
<b>Chapter 4: Linear Regression</b>	73
<hr/>	
<b>Linear models</b>	73
<b>A bidimensional example</b>	74
<b>Linear regression with scikit-learn and higher dimensionality</b>	76
Regressor analytic expression	80
<b>Ridge, Lasso, and ElasticNet</b>	81
<b>Robust regression with random sample consensus</b>	87
<b>Polynomial regression</b>	88
<b>Isotonic regression</b>	92
<b>References</b>	94
<b>Summary</b>	94
<b>Chapter 5: Logistic Regression</b>	95
<hr/>	
<b>Linear classification</b>	96
<b>Logistic regression</b>	98
<b>Implementation and optimizations</b>	100
<b>Stochastic gradient descent algorithms</b>	104
<b>Finding the optimal hyperparameters through grid search</b>	108
<b>Classification metrics</b>	111
<b>ROC curve</b>	116
<b>Summary</b>	120
<b>Chapter 6: Naive Bayes</b>	121
<hr/>	
<b>Bayes' theorem</b>	121
<b>Naive Bayes classifiers</b>	123
<b>Naive Bayes in scikit-learn</b>	124
Bernoulli naive Bayes	124
Multinomial naive Bayes	127
Gaussian naive Bayes	129
<b>References</b>	132
<b>Summary</b>	133
<b>Chapter 7: Support Vector Machines</b>	135
<hr/>	

<b>Linear support vector machines</b>	135
<b>scikit-learn implementation</b>	140
Linear classification	140
Kernel-based classification	143
Radial Basis Function	144
Polynomial kernel	144
Sigmoid kernel	145
Custom kernels	145
Non-linear examples	145
<b>Controlled support vector machines</b>	151
<b>Support vector regression</b>	153
<b>References</b>	155
<b>Summary</b>	155
<b>Chapter 8: Decision Trees and Ensemble Learning</b>	157
<hr/>	
<b>Binary decision trees</b>	158
Binary decisions	159
Impurity measures	161
Gini impurity index	162
Cross-entropy impurity index	162
Misclassification impurity index	163
Feature importance	163
<b>Decision tree classification with scikit-learn</b>	163
<b>Ensemble learning</b>	170
Random forests	170
Feature importance in random forests	173
AdaBoost	174
Gradient tree boosting	177
Voting classifier	179
<b>References</b>	183
<b>Summary</b>	183
<b>Chapter 9: Clustering Fundamentals</b>	185
<hr/>	
<b>Clustering basics</b>	185
K-means	187
Finding the optimal number of clusters	192
Optimizing the inertia	192
Silhouette score	194
Calinski-Harabasz index	198
Cluster instability	200
DBSCAN	203
Spectral clustering	206
<b>Evaluation methods based on the ground truth</b>	208

Homogeneity	208
Completeness	209
Adjusted rand index	209
<b>References</b>	210
<b>Summary</b>	211
<b>Chapter 10: Hierarchical Clustering</b>	<b>213</b>
<b>Hierarchical strategies</b>	213
<b>Agglomerative clustering</b>	214
Dendrograms	217
Agglomerative clustering in scikit-learn	219
Connectivity constraints	223
<b>References</b>	225
<b>Summary</b>	226
<b>Chapter 11: Introduction to Recommendation Systems</b>	<b>227</b>
<b>Naive user-based systems</b>	227
User-based system implementation with scikit-learn	228
<b>Content-based systems</b>	231
<b>Model-free (or memory-based) collaborative filtering</b>	234
<b>Model-based collaborative filtering</b>	237
Singular Value Decomposition strategy	238
Alternating least squares strategy	240
Alternating least squares with Apache Spark MLlib	241
<b>References</b>	245
<b>Summary</b>	246
<b>Chapter 12: Introduction to Natural Language Processing</b>	<b>247</b>
<b>NLTK and built-in corpora</b>	247
Corpora examples	249
<b>The bag-of-words strategy</b>	250
Tokenizing	252
Sentence tokenizing	252
Word tokenizing	253
Stopword removal	254
Language detection	255
Stemming	256
Vectorizing	257
Count vectorizing	257
N-grams	259
Tf-idf vectorizing	260
<b>A sample text classifier based on the Reuters corpus</b>	262

<b>References</b>	264
<b>Summary</b>	264
<b>Chapter 13: Topic Modeling and Sentiment Analysis in NLP</b>	267
<b>Topic modeling</b>	267
Latent semantic analysis	268
Probabilistic latent semantic analysis	275
Latent Dirichlet Allocation	281
<b>Sentiment analysis</b>	288
VADER sentiment analysis with NLTK	292
<b>References</b>	293
<b>Summary</b>	293
<b>Chapter 14: A Brief Introduction to Deep Learning and TensorFlow</b>	295
<b>Deep learning at a glance</b>	295
Artificial neural networks	296
Deep architectures	300
Fully connected layers	300
Convolutional layers	301
Dropout layers	303
Recurrent neural networks	303
<b>A brief introduction to TensorFlow</b>	304
Computing gradients	306
Logistic regression	309
Classification with a multi-layer perceptron	313
Image convolution	317
<b>A quick glimpse inside Keras</b>	320
<b>References</b>	326
<b>Summary</b>	326
<b>Chapter 15: Creating a Machine Learning Architecture</b>	327
<b>Machine learning architectures</b>	327
Data collection	329
Normalization	330
Dimensionality reduction	330
Data augmentation	331
Data conversion	331
Modeling/Grid search/Cross-validation	332
Visualization	332
<b>scikit-learn tools for machine learning architectures</b>	332
Pipelines	333

Feature unions	337
<b>References</b>	<b>338</b>
<b>Summary</b>	<b>338</b>
<b>Index</b>	<b>339</b>

---