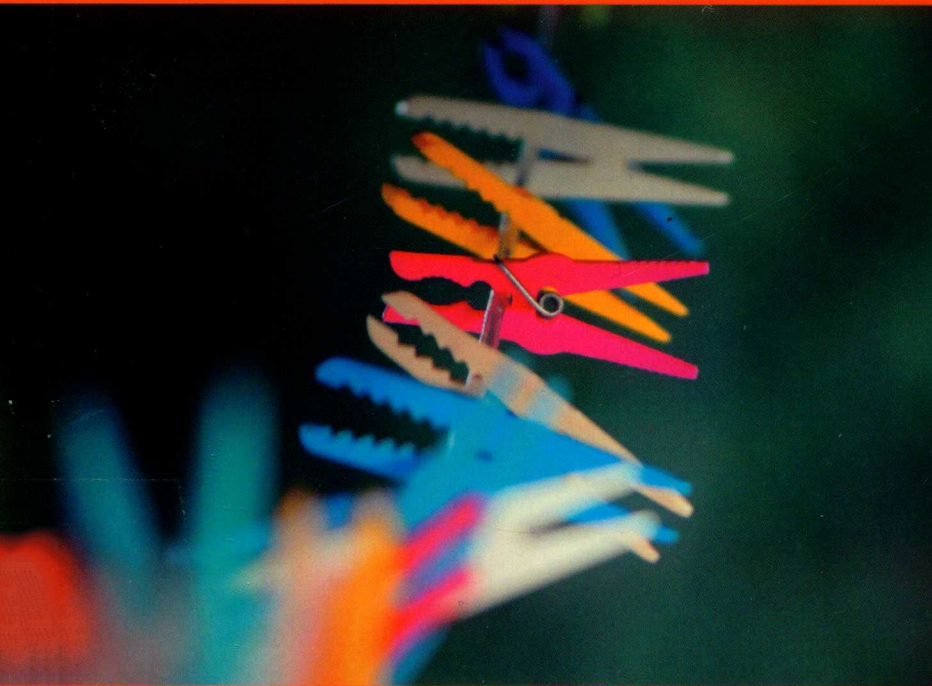


Machine Learning Algorithms, 2nd Edition

机器学习算法

第2版（影印版）

Giuseppe Bonaccorso 著



Packt>

www.packtpub.com



东南大学出版社
SOUTHEAST UNIVERSITY PRESS

机器学习算法 第2版(影印版)
Machine Learning Algorithms,
2nd Edition

Giuseppe Bonaccorso 著

南京 东南大学出版社

图书在版编目(CIP)数据

机器学习算法:英文/(意)朱塞佩·博纳科尔索
(Giuseppe Bonaccorso)著. —2 版(影印本). —南京:东南大学出版社,2019.3

书名原文:Machine Learning Algorithms, 2nd Edition
ISBN 978-7-5641-8291-5

I. ①机… II. ①朱… III. ①机器学习-算法-英文
IV. ①TP181

中国版本图书馆 CIP 数据核字(2019)第 025733 号
图字:10-2018-489 号

© 2018 by PACKT Publishing Ltd.

Reprint of the English Edition, jointly published by PACKT Publishing Ltd and Southeast University Press, 2019.
Authorized reprint of the original English edition, 2018 PACKT Publishing Ltd, the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 PACKT Publishing Ltd 出版 2018。

英文影印版由东南大学出版社出版 2019。此影印版的出版和销售得到出版权和销售权的所有者——PACKT Publishing Ltd 的许可。

版权所有,未得书面许可,本书的任何部分和全部不得以任何形式重制。

机器学习算法 第2版(影印版)

出版发行:东南大学出版社

地 址:南京四牌楼2号 邮编:210096

出 版 人:江建中

网 址: <http://www.seupress.com>

电子邮件: press@seupress.com

印 刷:常州市武进第三印刷有限公司

开 本:787毫米×980毫米 16开本

印 张:32.5

字 数:636千字

版 次:2019年3月第1版

印 次:2019年3月第1次印刷

书 号:ISBN 978-7-5641-8291-5

定 价:108.00元

To my family and to all the people who always believed in me and encouraged me in this long journey!

– Giuseppe Bonaccorso



mapt.io

Mapt is an online digital library that gives you full access to over 5,000 books and videos, as well as industry leading tools to help you plan your personal development and advance your career. For more information, please visit our website.

Why subscribe?

- Spend less time learning and more time coding with practical eBooks and Videos from over 4,000 industry professionals
- Improve your learning with Skill Plans built especially for you
- Get a free eBook or video every month
- Mapt is fully searchable
- Copy and paste, print, and bookmark content

PacktPub.com

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.PacktPub.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at service@packtpub.com for more details.

At www.PacktPub.com, you can also read a collection of free technical articles, sign up for a range of free newsletters, and receive exclusive discounts and offers on Packt books and eBooks.

Contributors

About the author

Giuseppe Bonaccorso is an experienced team leader/manager in AI, machine/deep learning solution design, management, and delivery. He got his MScEng in electronics in 2005 from the University of Catania, Italy, and continued his studies at the University of Rome Tor Vergata and the University of Essex, UK. His main interests include machine/deep learning, reinforcement learning, big data, bio-inspired adaptive systems, cryptocurrencies, and NLP.

I want to thank the people who have been close to me and have supported me, especially my parents, who never stopped encouraging me.

About the reviewer

Doug Ortiz is an experienced enterprise cloud, big data, data analytics, and solutions architect who has architected, designed, developed, re-engineered, and integrated enterprise solutions. Other expertise includes Amazon Web Services, Azure, Google Cloud, business intelligence, Hadoop, Spark, NoSQL databases, and SharePoint, to name a few.

He is the founder of Illustris, LLC and is reachable at dougortiz@illustris.org.

Huge thanks to my wonderful wife, Milla, Maria, Nikolay, and our children for all their support.

Packt is searching for authors like you

If you're interested in becoming an author for Packt, please visit authors.packtpub.com and apply today. We have worked with thousands of developers and tech professionals, just like you, to help them share their insight with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Table of Contents

Preface	1
Chapter 1: A Gentle Introduction to Machine Learning	7
Introduction – classic and adaptive machines	8
Descriptive analysis	11
Predictive analysis	12
Only learning matters	13
Supervised learning	14
Unsupervised learning	17
Semi-supervised learning	19
Reinforcement learning	21
Computational neuroscience	23
Beyond machine learning – deep learning and bio-inspired adaptive systems	24
Machine learning and big data	26
Summary	27
Chapter 2: Important Elements in Machine Learning	29
Data formats	29
Multiclass strategies	33
One-vs-all	33
One-vs-one	34
Learnability	34
Underfitting and overfitting	36
Error measures and cost functions	39
PAC learning	43
Introduction to statistical learning concepts	44
MAP learning	45
Maximum likelihood learning	46
Class balancing	51
Resampling with replacement	52
SMOTE resampling	54
Elements of information theory	57
Entropy	57
Cross-entropy and mutual information	59
Divergence measures between two probability distributions	61
Summary	62
Chapter 3: Feature Selection and Feature Engineering	65
scikit-learn toy datasets	66

Creating training and test sets	67
Managing categorical data	69
Managing missing features	72
Data scaling and normalization	74
Whitening	76
Feature selection and filtering	78
Principal Component Analysis	81
Non-Negative Matrix Factorization	88
Sparse PCA	90
Kernel PCA	92
Independent Component Analysis	95
Atom extraction and dictionary learning	99
Visualizing high-dimensional datasets using t-SNE	102
Summary	104
Chapter 4: Regression Algorithms	105
Linear models for regression	105
A bidimensional example	107
Linear regression with scikit-learn and higher dimensionality	112
R2 score	116
Explained variance	117
Regressor analytic expression	118
Ridge, Lasso, and ElasticNet	119
Ridge	119
Lasso	122
ElasticNet	124
Robust regression	125
RANSAC	126
Huber regression	128
Bayesian regression	130
Polynomial regression	134
Isotonic regression	138
Summary	141
Chapter 5: Linear Classification Algorithms	143
Linear classification	144
Logistic regression	147
Implementation and optimizations	150
Stochastic gradient descent algorithms	153
Passive-aggressive algorithms	157
Passive-aggressive regression	163
Finding the optimal hyperparameters through a grid search	167
Classification metrics	170
Confusion matrix	172

Precision	176
Recall	176
F-Beta	177
Cohen's Kappa	178
Global classification report	180
Learning curve	180
ROC curve	182
Summary	186
Chapter 6: Naive Bayes and Discriminant Analysis	187
Bayes' theorem	188
Naive Bayes classifiers	190
Naive Bayes in scikit-learn	191
Bernoulli Naive Bayes	191
Multinomial Naive Bayes	194
An example of Multinomial Naive Bayes for text classification	196
Gaussian Naive Bayes	199
Discriminant analysis	203
Summary	208
Chapter 7: Support Vector Machines	209
Linear SVM	209
SVMs with scikit-learn	214
Linear classification	215
Kernel-based classification	217
Radial Basis Function	218
Polynomial kernel	219
Sigmoid kernel	219
Custom kernels	219
Non-linear examples	220
v-Support Vector Machines	225
Support Vector Regression	228
An example of SVR with the Airfoil Self-Noise dataset	232
Introducing semi-supervised Support Vector Machines (S3VM)	236
Summary	243
Chapter 8: Decision Trees and Ensemble Learning	245
Binary Decision Trees	246
Binary decisions	247
Impurity measures	250
Gini impurity index	250
Cross-entropy impurity index	250
Misclassification impurity index	252
Feature importance	252
Decision Tree classification with scikit-learn	252
Decision Tree regression	260

Example of Decision Tree regression with the Concrete Compressive Strength dataset	261
Introduction to Ensemble Learning	267
Random Forests	268
Feature importance in Random Forests	271
AdaBoost	273
Gradient Tree Boosting	277
Voting classifier	280
Summary	284
Chapter 9: Clustering Fundamentals	285
Clustering basics	285
k-NN	288
Gaussian mixture	294
Finding the optimal number of components	298
K-means	301
Finding the optimal number of clusters	308
Optimizing the inertia	308
Silhouette score	310
Calinski-Harabasz index	314
Cluster instability	316
Evaluation methods based on the ground truth	319
Homogeneity	319
Completeness	320
Adjusted Rand Index	321
Summary	322
Chapter 10: Advanced Clustering	323
DBSCAN	324
Spectral Clustering	328
Online Clustering	331
Mini-batch K-means	332
BIRCH	334
Biclustering	337
Summary	340
Chapter 11: Hierarchical Clustering	343
Hierarchical strategies	343
Agglomerative Clustering	344
Dendrograms	346
Agglomerative Clustering in scikit-learn	349
Connectivity constraints	354
Summary	359
Chapter 12: Introducing Recommendation Systems	361
Naive user-based systems	362

Implementing a user-based system with scikit-learn	363
Content-based systems	365
Model-free (or memory-based) collaborative filtering	367
Model-based collaborative filtering	370
Singular value decomposition strategy	371
Alternating least squares strategy	373
ALS with Apache Spark MLlib	374
Summary	378
Chapter 13: Introducing Natural Language Processing	379
NLTK and built-in corpora	380
Corpora examples	381
The Bag-of-Words strategy	382
Tokenizing	384
Sentence tokenizing	384
Word tokenizing	385
Stopword removal	386
Language detection	387
Stemming	388
Vectorizing	389
Count vectorizing	389
N-grams	391
TF-IDF vectorizing	391
Part-of-Speech	393
Named Entity Recognition	395
A sample text classifier based on the Reuters corpus	396
Summary	397
Chapter 14: Topic Modeling and Sentiment Analysis in NLP	399
Topic modeling	399
Latent Semantic Analysis	400
Probabilistic Latent Semantic Analysis	407
Latent Dirichlet Allocation	413
Introducing Word2vec with Gensim	418
Sentiment analysis	422
VADER sentiment analysis with NLTK	426
Summary	427
Chapter 15: Introducing Neural Networks	429
Deep learning at a glance	429
Artificial neural networks	430
MLPs with Keras	434
Interfacing Keras to scikit-learn	443
Summary	445
Chapter 16: Advanced Deep Learning Models	447
Deep model layers	447

Fully connected layers	448
Convolutional layers	449
Dropout layers	450
Batch normalization layers	451
Recurrent Neural Networks	451
An example of a deep convolutional network with Keras	452
An example of an LSTM network with Keras	456
A brief introduction to TensorFlow	462
Computing gradients	464
Logistic regression	467
Classification with a multilayer perceptron	471
Image convolution	474
Summary	476
Chapter 17: Creating a Machine Learning Architecture	477
Machine learning architectures	477
Data collection	479
Normalization and regularization	480
Dimensionality reduction	480
Data augmentation	481
Data conversion	483
Modeling/grid search/cross-validation	483
Visualization	484
GPU support	484
A brief introduction to distributed architectures	488
Scikit-learn tools for machine learning architectures	491
Pipelines	491
Feature unions	495
Summary	496
Other Books You May Enjoy	497
Index	501

Preface

This book is an introduction to the world of machine learning, a topic that is becoming more and more important, not only for IT professionals and analysts but also for all the data scientists and engineers who want to exploit the enormous power of techniques such as predictive analysis, classification, clustering, and natural language processing. In order to facilitate the learning process, all theoretical elements are followed by concrete examples based on Python.

A basic but solid understanding of this topic requires a foundation in mathematics, which is not only necessary to explain the algorithms, but also to let the reader understand how it's possible to tune up the hyperparameters in order to attain the best possible accuracy. Of course, it's impossible to cover all the details with the appropriate precision. For this reason, some topics are only briefly described, limiting the theory to the results without providing any of the workings. In this way, the user has the double opportunity to focus on the fundamental concepts (without too many mathematical complications) and, through the references, examine in depth all the elements that generate interest.

The chapters can be read in no particular order, skipping the topics that you already know. Whenever necessary, there are references to the chapters where some concepts are explained. I apologize in advance for any imprecision, typos or mistakes, and I'd like to thank all the Packt editors for their collaboration and constant attention.

Who this book is for

This book is for machine learning engineers, data engineers, and data scientists who want to build a strong foundation in the field of predictive analytics and machine learning. Familiarity with Python would be an added advantage and will enable you to get the most out of this book.

What this book covers

Chapter 1, *A Gentle Introduction to Machine Learning*, introduces the world of machine learning, explaining the fundamental concepts of the most important approaches to creating intelligent applications and focusing on the different kinds of learning methods.

Chapter 2, *Important Elements in Machine Learning*, explains the mathematical concepts regarding the most common machine learning problems, including the concept of learnability and some important elements of information theory. This chapter contains theoretical elements, but it's extremely helpful if you are learning this topic from scratch because it provides an insight into the most important mathematical tools employed in the majority of algorithms.

Chapter 3, *Feature Selection and Feature Engineering*, describes the most important techniques for preprocessing a dataset, selecting the most informative features, and reducing the original dimensionality.

Chapter 4, *Regression Algorithms*, describes the linear regression algorithm and its optimizations: Ridge, Lasso, and ElasticNet. It continues with more advanced models that can be employed to solve non-linear regression problems or to mitigate the effect of outliers.

Chapter 5, *Linear Classification Algorithms*, introduces the concept of linear classification, focusing on logistic regression, perceptrons, stochastic gradient descent algorithms, and passive-aggressive algorithms. The second part of the chapter covers the most important evaluation metrics, which are used to measure the performance of a model and find the optimal hyperparameter set.

Chapter 6, *Naive Bayes and Discriminant Analysis*, explains the Bayes probability theory and describes the structure of the most diffused Naive Bayes classifiers. In the second part, linear and quadratic discriminant analysis is analyzed with some concrete examples.

Chapter 7, *Support Vector Machines*, introduces the SVM family of algorithms, focusing on both linear and non-linear classification problems thanks to the employment of the kernel trick. The last part of the chapter covers support vector regression and more complex classification models.

Chapter 8, *Decision Trees and Ensemble Learning*, explains the concept of a hierarchical decision process and describes the concepts of decision tree classification, random forests, bootstrapped and bagged trees, and voting classifiers.

Chapter 9, *Clustering Fundamentals*, introduces the concept of clustering, describing the Gaussian mixture, K-Nearest Neighbors, and K-means algorithms. The last part of the chapter covers different approaches to determining the optimal number of clusters and measuring the performance of a model.

Chapter 10, *Advanced Clustering*, introduces more complex clustering techniques (DBSCAN, Spectral Clustering, and Biclustering) that can be employed when the dataset structure is non-convex. In the second part of the chapter, two online clustering algorithms (mini-batch K-means and BIRCH) are introduced.

Chapter 11, *Hierarchical Clustering*, continues the explanation of more complex clustering algorithms started in the previous chapter and introduces the concepts of agglomerative clustering and dendrograms.

Chapter 12, *Introducing Recommendation Systems*, explains the most diffused algorithms employed in recommender systems: content- and user-based strategies, collaborative filtering, and alternating least square. A complete example based on Apache Spark shows how to process very large datasets using the ALS algorithm.

Chapter 13, *Introduction to Natural Language Processing*, explains the concept of the Bag-of-Words strategy and introduces the most important techniques required to efficiently process natural language datasets (tokenizing, stemming, stop-word removal, tagging, and vectorizing). An example of a classifier based on the Reuters dataset is also discussed in the last part of the chapter.

Chapter 14, *Topic Modeling and Sentiment Analysis in NLP*, introduces the concept of topic modeling and describes the most important algorithms, such as latent semantic analysis (both deterministic and probabilistic) and latent Dirichlet allocation. The second part of the chapter covers the problem of word embedding and sentiment analysis, explaining the most diffused approaches to address it.

Chapter 15, *Introducing Neural Networks*, introduces the world of deep learning, explaining the concept of neural networks and computational graphs. In the second part of the chapter, the high-level deep learning framework Keras is presented with a concrete example of a Multi-layer Perceptron.

Chapter 16, *Advanced Deep Learning Models*, explains the basic functionalities of the most important deep learning layers, with Keras examples of deep convolutional networks and recurrent (LSTM) networks for time-series processing. In the second part of the chapter, the TensorFlow framework is briefly introduced, along with some examples that expose some of its basic functionalities.

Chapter 17, *Creating a Machine Learning Architecture*, explains how to define a complete machine learning pipeline, focusing on the peculiarities and drawbacks of each step.

To get the most out of this book

To fully understand all the algorithms in this book, it's important to have a basic knowledge of linear algebra, probability theory, and calculus.

All practical examples are written in Python and use the scikit-learn machine learning framework, **Natural Language Toolkit (NLTK)**, Crab, langdetect, Spark (PySpark), Gensim, Keras, and TensorFlow (deep learning frameworks). These are available for Linux, macOS X, and Windows, with Python 2.7 and 3.3+. When a particular framework is employed for a specific task, detailed instructions and references will be provided. All the examples from chapters 1 to 14 can be executed using Python 2.7 (while TensorFlow requires Python 3.5+); however, I highly suggest using a Python 3.5+ distribution. The most common choice for data science and machine learning is Anaconda (<https://www.anaconda.com/download/>), which already contains all the most important packages.

Download the example code files

You can download the example code files for this book from your account at www.packtpub.com. If you purchased this book elsewhere, you can visit www.packtpub.com/support and register to have the files emailed directly to you.

You can download the code files by following these steps:

1. Log in or register at www.packtpub.com.
2. Select the **SUPPORT** tab.
3. Click on **Code Downloads & Errata**.
4. Enter the name of the book in the **Search** box and follow the onscreen instructions.

Once the file is downloaded, please make sure that you unzip or extract the folder using the latest version of:

- WinRAR/7-Zip for Windows
- Zipeg/iZip/UnRarX for Mac
- 7-Zip/PeaZip for Linux