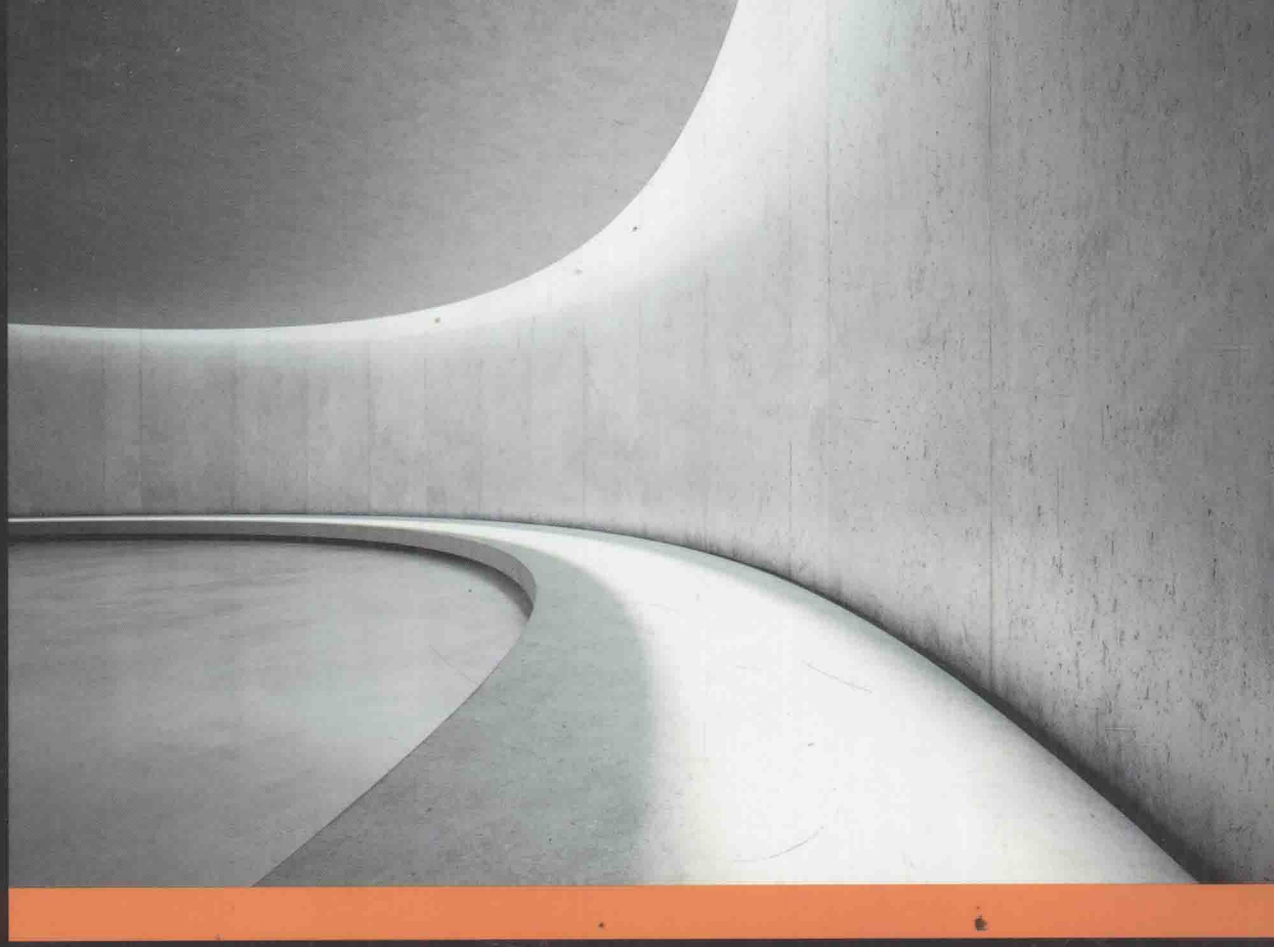

Python语言构建机器学习系统 第2版（影印版）

Building Machine Learning Systems with Python

Second Edition

Luis Pedro Coelho, Willi Richert 著

[PACKT] PUBLISHING

运用Python语言创建实用的机器学习系统，从数据中获得更多价值

东南大学出版社 SOUTHEAST UNIVERSITY PRESS

Python 语言构建机器学习系统

第 2 版

（影印版）

Luis Pedro Coelho，*Willi Richert* 著

南京　东南大学出版社

图书在版编目(CIP)数据

Python 语言构建机器学习系统:第 2 版:英文/(美)科埃略(Coelho,L.P.),(美)里克特(Richert,W.)著.—影印本. —南京:东南大学出版社,2016.1

书名原文:Building Machine Learning Systems with Python, Second Edition

ISBN 978-7-5641-6062-3

Ⅰ.①P… Ⅱ.①科… ②里… Ⅲ.①软件工具—程序设计—英文 Ⅳ.①TP311.56

中国版本图书馆 CIP 数据核字(2015)第 241940 号

Python 语言构建机器学习系统 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:20.25
字　　数:397 千字
版　　次:2016 年 1 月第 1 版
印　　次:2016 年 1 月第 1 次印刷
书　　号:ISBN 978-7-5641-6062-3
定　　价:68.00 元

本社图书若有印装质量问题,请直接与营销部联系。电话(传真):025-83791830

Credits

Authors
Luis Pedro Coelho
Willi Richert

Reviewers
Matthieu Brucher
Maurice HT Ling
Radim Řehůřek

Commissioning Editor
Kartikey Pandey

Acquisition Editors
Greg Wild
Richard Harvey
Kartikey Pandey

Content Development Editor
Arun Nadar

Technical Editor
Pankaj Kadam

Copy Editors
Relin Hedly
Sameen Siddiqui
Laxmi Subramanian

Project Coordinator
Nikhil Nair

Proofreaders
Simran Bhogal
Lawrence A. Herman
Linda Morris
Paul Hindle

Indexer
Hemangini Bari

Graphics
Sheetal Aute
Abhinash Sahu

Production Coordinator
Arvindkumar Gupta

Cover Work
Arvindkumar Gupta

About the Authors

Luis Pedro Coelho is a computational biologist: someone who uses computers as a tool to understand biological systems. In particular, Luis analyzes DNA from microbial communities to characterize their behavior. Luis has also worked extensively in bioimage informatics—the application of machine learning techniques for the analysis of images of biological specimens. His main focus is on the processing and integration of large-scale datasets.

Luis has a PhD from Carnegie Mellon University, one of the leading universities in the world in the area of machine learning. He is the author of several scientific publications.

Luis started developing open source software in 1998 as a way to apply real code to what he was learning in his computer science courses at the Technical University of Lisbon. In 2004, he started developing in Python and has contributed to several open source libraries in this language. He is the lead developer on the popular computer vision package for Python and mahotas, as well as the contributor of several machine learning codes.

Luis currently divides his time between Luxembourg and Heidelberg.

I thank my wife, Rita, for all her love and support and my daughter, Anna, for being the best thing ever.

Willi Richert has a PhD in machine learning/robotics, where he used reinforcement learning, hidden Markov models, and Bayesian networks to let heterogeneous robots learn by imitation. Currently, he works for Microsoft in the Core Relevance Team of Bing, where he is involved in a variety of ML areas such as active learning, statistical machine translation, and growing decision trees.

This book would not have been possible without the support of my wife, Natalie, and my sons, Linus and Moritz. I am especially grateful for the many fruitful discussions with my current or previous managers, Andreas Bode, Clemens Marschner, Hongyan Zhou, and Eric Crestan, as well as my colleagues and friends, Tomasz Marciniak, Cristian Eigel, Oliver Niehoerster, and Philipp Adelt. The interesting ideas are most likely from them; the bugs belong to me.

About the Reviewers

Matthieu Brucher holds an engineering degree from the Ecole Supérieure d'Electricité (Information, Signals, Measures), France and has a PhD in unsupervised manifold learning from the Université de Strasbourg, France. He currently holds an HPC software developer position in an oil company and is working on the next generation reservoir simulation.

Maurice HT Ling has been programming in Python since 2003. Having completed his PhD in Bioinformatics and BSc (Hons.) in Molecular and Cell Biology from The University of Melbourne, he is currently a Research Fellow at Nanyang Technological University, Singapore, and an Honorary Fellow at The University of Melbourne, Australia. Maurice is the Chief Editor for *Computational and Mathematical Biology*, and co-editor for *The Python Papers*. Recently, Maurice cofounded the first synthetic biology start-up in Singapore, AdvanceSyn Pte. Ltd., as the Director and Chief Technology Officer. His research interests lies in life—biological life, artificial life, and artificial intelligence—using computer science and statistics as tools to understand life and its numerous aspects. In his free time, Maurice likes to read, enjoy a cup of coffee, write his personal journal, or philosophize on various aspects of life. His website and LinkedIn profile are `http://maurice.vodien.com` and `http://www.linkedin.com/in/mauriceling`, respectively.

Radim Řehůřek is a tech geek and developer at heart. He founded and led the research department at Seznam.cz, a major search engine company in central Europe. After finishing his PhD, he decided to move on and spread the machine learning love, starting his own privately owned R&D company, RaRe Consulting Ltd. RaRe specializes in made-to-measure data mining solutions, delivering cutting-edge systems for clients ranging from large multinationals to nascent start-ups.

Radim is also the author of a number of popular open source projects, including gensim and smart_open.

A big fan of experiencing different cultures, Radim has lived around the globe with his wife for the past decade, with his next steps leading to South Korea. No matter where he stays, Radim and his team always try to evangelize data-driven solutions and help companies worldwide make the most of their machine learning opportunities.

www.PacktPub.com

Support files, eBooks, discount offers, and more

For support files and downloads related to your book, please visit `www.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.

`https://www2.packtpub.com/books/subscription/packtlib`

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can search, access, and read Packt's entire library of books.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Free access for Packt account holders

If you have an account with Packt at `www.PacktPub.com`, you can use this to access PacktLib today and view 9 entirely free books. Simply use your login credentials for immediate access.

Preface

One could argue that it is a fortunate coincidence that you are holding this book in your hands (or have it on your eBook reader). After all, there are millions of books printed every year, which are read by millions of readers. And then there is this book read by you. One could also argue that a couple of machine learning algorithms played their role in leading you to this book—or this book to you. And we, the authors, are happy that you want to understand more about the hows and whys.

Most of the book will cover the *how*. How has data to be processed so that machine learning algorithms can make the most out of it? How should one choose the right algorithm for a problem at hand?

Occasionally, we will also cover the *why*. Why is it important to measure correctly? Why does one algorithm outperform another one in a given scenario?

We know that there is much more to learn to be an expert in the field. After all, we only covered some *hows* and just a tiny fraction of the *whys*. But in the end, we hope that this mixture will help you to get up and running as quickly as possible.

What this book covers

Chapter 1, Getting Started with Python Machine Learning, introduces the basic idea of machine learning with a very simple example. Despite its simplicity, it will challenge us with the risk of overfitting.

Chapter 2, Classifying with Real-world Examples, uses real data to learn about classification, whereby we train a computer to be able to distinguish different classes of flowers.

Chapter 3, Clustering - Finding Related Posts, teaches how powerful the bag of words approach is, when we apply it to finding similar posts without really "understanding" them.

Chapter 4, Topic Modeling, moves beyond assigning each post to a single cluster and assigns them to several topics as a real text can deal with multiple topics.

Chapter 5, Classification – Detecting Poor Answers, teaches how to use the bias-variance trade-off to debug machine learning models though this chapter is mainly on using a logistic regression to find whether a user's answer to a question is good or bad.

Chapter 6, Classification II – Sentiment Analysis, explains how Naïve Bayes works, and how to use it to classify tweets to see whether they are positive or negative.

Chapter 7, Regression, explains how to use the classical topic, regression, in handling data, which is still relevant today. You will also learn about advanced regression techniques such as the Lasso and ElasticNets.

Chapter 8, Recommendations, builds recommendation systems based on costumer product ratings. We will also see how to build recommendations just from shopping data without the need for ratings data (which users do not always provide).

Chapter 9, Classification – Music Genre Classification, makes us pretend that someone has scrambled our huge music collection, and our only hope to create order is to let a machine learner classify our songs. It will turn out that it is sometimes better to trust someone else's expertise than creating features ourselves.

Chapter 10, Computer Vision, teaches how to apply classification in the specific context of handling images by extracting features from data. We will also see how these methods can be adapted to find similar images in a collection.

Chapter 11, Dimensionality Reduction, teaches us what other methods exist that can help us in downsizing data so that it is chewable by our machine learning algorithms.

Chapter 12, Bigger Data, explores some approaches to deal with larger data by taking advantage of multiple cores or computing clusters. We also have an introduction to using cloud computing (using Amazon Web Services as our cloud provider).

Appendix, Where to Learn More Machine Learning, lists many wonderful resources available to learn more about machine learning.

What you need for this book

This book assumes you know Python and how to install a library using easy_install or pip. We do not rely on any advanced mathematics such as calculus or matrix algebra.

We are using the following versions throughout the book, but you should be fine with any more recent ones:

- Python 2.7 (all the code is compatible with version 3.3 and 3.4 as well)
- NumPy 1.8.1
- SciPy 0.13
- scikit-learn 0.14.0

Who this book is for

This book is for Python programmers who want to learn how to perform machine learning using open source libraries. We will walk through the basic modes of machine learning based on realistic examples.

This book is also for machine learners who want to start using Python to build their systems. Python is a flexible language for rapid prototyping, while the underlying algorithms are all written in optimized C or C++. Thus the resulting code is fast and robust enough to be used in production as well.

Conventions

In this book, you will find a number of styles of text that distinguish between different kinds of information. Here are some examples of these styles, and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows: "We then use `polyld()` to create a model function from the model parameters."

A block of code is set as follows:

```
[aws info]
AWS_ACCESS_KEY_ID =  AAKIIT7HHF6IUSN3OCAA
AWS_SECRET_ACCESS_KEY = <your secret key>
```

Any command-line input or output is written as follows:

```
>>> import numpy
>>> numpy.version.full_version
1.8.1
```

New terms and **important words** are shown in bold. Words that you see on the screen, in menus or dialog boxes for example, appear in the text like this: "Once the machine is stopped, the **Change instance type** option becomes available."

Warnings or important notes appear in a box like this.

Tips and tricks appear like this.

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book—what you liked or may have disliked. Reader feedback is important for us to develop titles that you really get the most out of.

To send us general feedback, simply send an e-mail to `feedback@packtpub.com`, and mention the book title via the subject of your message. If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide on `www.packtpub.com/authors`.

Customer support

Now that you are the proud owner of a Packt book, we have a number of things to help you to get the most from your purchase.

Downloading the example code

You can download the example code files from your account at `http://www.packtpub.com` for all the Packt Publishing books you have purchased. If you purchased this book elsewhere, you can visit `http://www.packtpub.com/support` and register to have the files e-mailed directly to you.

The code for this book is also available on GitHub at `https://github.com/luispedro/BuildingMachineLearningSystemsWithPython`. This repository is kept up-to-date so that it will incorporate both errata and any necessary updates for newer versions of Python or of the packages we use in the book.

Errata

Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you find a mistake in one of our books—maybe a mistake in the text or the code—we would be grateful if you could report this to us. By doing so, you can save other readers from frustration and help us improve subsequent versions of this book. If you find any errata, please report them by visiting `http://www.packtpub.com/submit-errata`, selecting your book, clicking on the **Errata Submission Form** link, and entering the details of your errata. Once your errata are verified, your submission will be accepted and the errata will be uploaded to our website or added to any list of existing errata under the Errata section of that title.

To view the previously submitted errata, go to `https://www.packtpub.com/books/content/support` and enter the name of the book in the search field. The required information will appear under the **Errata** section.

Another excellent way would be to visit `www.TwoToReal.com` where the authors try to provide support and answer all your questions.

Piracy

Piracy of copyright material on the Internet is an ongoing problem across all media. At Packt, we take the protection of our copyright and licenses very seriously. If you come across any illegal copies of our works, in any form, on the Internet, please provide us with the location address or website name immediately so that we can pursue a remedy.

Please contact us at `copyright@packtpub.com` with a link to the suspected pirated material.

We appreciate your help in protecting our authors, and our ability to bring you valuable content.

Questions

You can contact us at `questions@packtpub.com` if you are having a problem with any aspect of the book, and we will do our best to address it.

Table of Contents