Community Experience Distilled

Mastering Machine Learning with scikit-learn

Apply effective learning algorithms to real-world problems using scikit-learn

Gavin Hackeling

[PACKT] open source
PUBLISHING

Table of Contents

Mastering Machine Learning with scikit-learn

Mastering Machine Learning with scikit-learn

First published: October 2014

Production reference: 1221014

Published by Packt Publishing Ltd.

Livery Place

35 Livery Street

Birmingham B3 2PB, UK.

ISBN 978-1-78398-836-5

www.packtpub.com

Cover image by Amy-Lee Winfield (<abjure@outlook.com>)

Credits

Author

Gavin Hackeling

Reviewers

Fahad Arshad

Sarah Guido

Mikhail Korobov

Aman Madaan

Acquisition Editor

Meeta Rajani

Content Development Editor

Neeshma Ramakrishnan

Technical Editor

Faisal Siddiqui

Copy Editors

Roshni Banerjee

Adithi Shetty

Project Coordinator

Danuta Jones

Proofreaders

Simran Bhogal

Tarsonia Sanghera

Lindsey Thomas

Indexer

Monica Ajmera Mehta

Graphics

Sheetal Aute

Ronak Dhruv

Disha Haria

Production Coordinator

Kyle Albuquerque

Cover Work

Kyle Albuquerque

About the Author

Gavin Hackeling develops machine learning services for large-scale documents and image classification at an advertising network in New York. He received his Master's degree from New York University's Interactive Telecommunications Program, and his Bachelor's degree from the University of North Carolina.

To Hallie, for her support, and Zipper, without whose contributions this book would have been completed in half the time.

About the Reviewers

Fahad Arshad completed his PhD at Purdue University in the Department of Electrical and Computer Engineering. His research interests focus on developing algorithms for software testing, error detection, and failure diagnosis in distributed systems. He is particularly interested in data-driven analysis of computer systems. His work has appeared at top dependability conferences—DSN, ISSRE, ICAC, Middleware, and SRDS—and he has been awarded grants to attend DSN, ICAC, and ICNP. Fahad has also been an active contributor to security research while working as a cybersecurity engineer at NEEScomm IT. He has recently taken on a position as a systems engineer in the industry.

Sarah Guido is a data scientist at Reonomy, where she's helping build disruptive technology in the commercial real estate industry. She loves Python, machine learning, and the startup world. She is an accomplished conference speaker and an O'Reilly Media author, and is very involved in the Python community. Prior to joining Reonomy, Sarah earned a Master's degree from the University of Michigan School of Information.

Mikhail Korobov is a software developer at ScrapingHub Inc., where he works on web scraping, information extraction, natural language processing, machine learning, and web development tasks. He is an NLTK team member, Scrapy team member, and an author or contributor to many other open source projects.

I'd like to thank my wife, Aleksandra, for her support and patience and for the cookies.

Aman Madaan is currently pursuing his Master's in Computer Science and Engineering. His interests span across machine learning, information extraction, natural language processing, and distributed computing. More details about his skills, interests, and experience can be found at http://www.amanmadaan.in.

www.PacktPub.com

Support files, eBooks, discount offers, and more

You might want to visit www.PacktPub.com for support files and downloads related to your book.

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.

http://PacktLib.PacktPub.com

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can access, read, and search across 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 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 nine entirely free books. Simply use your login credentials for immediate access.

Preface

Recent years have seen the rise of machine learning, the study of software that learns from experience. While machine learning is a new discipline, it has found many applications. We rely on some of these applications daily; in some cases, their successes have already rendered them mundane. Many other applications have only recently been conceived, and hint at machine learning's potential.

In this book, we will examine several machine learning models and learning algorithms. We will discuss tasks that machine learning is commonly applied to, and learn to measure the performance of machine learning systems. We will work with a popular library for the Python programming language called scikit-learn, which has assembled excellent implementations of many machine learning models and algorithms under a simple yet versatile API.

This book is motivated by two goals:

- Its content should be accessible. The book only assumes familiarity with basic programming and math.
- Its content should be practical. This book offers hands-on examples that readers can adapt to problems in the real world.

What this book covers

Chapter 1, *The Fundamentals of Machine Learning*, defines machine learning as the study and design of programs that improve their performance of a task by learning from experience. This definition guides the other chapters; in each chapter, we will examine a machine learning model, apply it to a task, and measure its performance.

Chapter 2, *Linear Regression*, discusses linear regression, a model that relates explanatory variables and model parameters to a continuous response variable. You will learn about cost functions, and use the normal equation to find the parameter values that produce the optimal model.

Chapter 3, *Feature Extraction and Preprocessing*, describes methods to represent text, images, and categorical variables as features that can be used in machine learning models.

Chapter 4, *From Linear Regression to Logistic Regression*, discusses generalizing linear regression to support classification tasks. We combine a model called logistic regression with some of the feature engineering techniques from the previous chapter to create a spam filter.

Chapter 5, *Nonlinear Classification and Regression with Decision Trees*, departs from linear models to discuss classification and regression with models called decision trees. We use an ensemble of decision trees to construct a banner advertisement blocker.

Chapter 6, *Clustering with K-Means*, introduces unsupervised learning. We examine the k-means algorithm, and combine it with logistic regression to create a semi-supervised photo classifier.

Chapter 7, *Dimensionality Reduction with PCA*, discusses another unsupervised learning task called dimensionality reduction. We use principal component analysis to visualize high-dimensional data and build a face recognizer.

Chapter 8, *The Perceptron*, describes an online, binary classifier called the perceptron. The limitations of the perceptron motivate the models described in the final chapters.

Chapter 9, *From the Perceptron to Support Vector Machines*, discusses efficient nonlinear classification and regression with support vector machines. We use support vector machines to recognize the characters in photographs of street signs.

Chapter 10, *From the Perceptron to Artificial Neural Networks*, introduces powerful

nonlinear models for classification and regression called artificial neural networks. We build a network that can recognize handwritten digits.

What you need for this book

The examples in this book assume that you have an installation of Python 2.7. The first chapter will describe methods to install scikit-learn 0.15.2, its dependencies, and other libraries on Linux, OS X, and Windows.

Who this book is for

This book is intended for software developers who have some experience with machine learning. scikit-learn's API is well-documented, but assumes that the reader understands how machine learning algorithms work and when it is appropriate to use them. This book does not attempt to reproduce the API's documentation. Instead, it describes how machine learning models work, how their parameters are learned, and how they can be evaluated. When practical, we will work through toy examples of the algorithms in detail to build the understanding required to apply them effectively.

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.

In-line code is formatted as follows: "The `TfidfVectorizer` combines the `CountVectorizer` and the `TfidfTransformer`."

A block of code is indicated as follows:

```
>>> import pandas as pd
>>> from sklearn.feature_extraction.text import
TfidfVectorizer
>>> from sklearn.linear_model.logistic import
LogisticRegression
>>> from sklearn.cross_validation import train_test_split
>>> df = pd.read_csv('sms/sms.csv')
>>> X_train_raw, X_test_raw, y_train, y_test =
train_test_split(df['message'], df['label'])
>>> vectorizer = TfidfVectorizer()
>>> X_train = vectorizer.fit_transform(X_train_raw)
>>> X_test = vectorizer.transform(X_test_raw)
>>> classifier = LogisticRegression()
>>> classifier.fit(X_train, y_train)
```

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.