Gavin Hackeling

Mastering Machine Learning with scikit-learn

Second Edition

Learn to implement and evaluate machine learning solutions with scikit-learn

Packt>

Mastering Machine Learning with scikit-learn - Second Edition

Machine learning is the buzzword that brings computer science and statistics together to build smart and efficient models. Using the powerful algorithms and techniques offered by machine learning you can automate any analytical model.

This book examines a variety of machine learning models, including popular machine learning algorithms such as k-Nearest Neighbors, logistic regression, Naive Bayes, K-means, decision trees, and artificial neural networks. It discusses data preprocessing, hyperparameter optimization, and ensemble methods. You will build systems that classify documents, recognize images, detect ads, and more. You will learn how to use scikit-learn's API to extract features from categorical variables, text, and images; evaluate model performance; and develop an intuition for how to improve your model's performance.

By the end of this book, you will master all the concepts of scikit-learn that are required to build efficient models at work to carry out advanced tasks with a practical approach.

Things you will learn:

- Review fundamental concepts such as bias and variance
- Extract features from categorical variables, text, and images
- Predict the values of continuous variables using linear regression and k-Nearest Neighbors
- Classify documents and images using logistic regression and support vector machines
- Create ensembles of estimators using bagging and boosting techniques
- Discover hidden structures in data using K-means clustering
- Evaluate the performance of machine learning systems in common tasks

Packt>
www.packtpub.com

$ **44.99** US
£ **37.99** UK

Prices do not include local sales
Tax or VAT where applicable

ISBN 978-1-78829-987-9
54499
9 781788 299879

Mastering Machine Learning with scikit-learn - Second Edition

Gavin Hackeling

Mastering Machine Learning with scikit-learn

Second Edition

Learn to implement and evaluate machine learning solutions with scikit-learn

Gavin Hackeling

Packt>

BIRMINGHAM - MUMBAI

Mastering Machine Learning with scikit-learn

Second Edition

First published: October 2014

second published: July 2017

Production reference: 1200717

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham
B3 2PB, UK.
ISBN 978-1-78829-987-9

www.packtpub.com

Credits

Author
Gavin Hackeling

Reviewer
Oleg Okun

Commissioning Editor
Amey Varangaonkar

Acquisition Editor
Aman Singh

Content Development Editor
Aishwarya Pandere

Technical Editor
Suwarna Patil

Copy Editors
Safis Editing
Vikrant Phadkay

Project Coordinator
Nidhi Joshi

Proofreader
Safis Editing

Indexer
Tejal Daruwale Soni

Graphics
Tania Dutta

Production Coordinator
Arvindkumar Gupta

About the Author

Gavin Hackeling is a data scientist and author. He was worked on a variety of machine learning problems, including automatic speech recognition, document classification, object recognition, and semantic segmentation. An alumnus of the University of North Carolina and New York University, he lives in Brooklyn with his wife and cat.

I would like to thank my wife, Hallie, and the scikti-learn community.

About the Reviewer

Oleg Okun is a machine learning expert and an author/editor of four books, numerous journal articles, and conference papers. His career spans more than a quarter of a century. He was employed in both academia and industry in his motherland, Belarus, and abroad (Finland, Sweden, and Germany). His work experience includes document image analysis, fingerprint biometrics, bioinformatics, online/offline marketing analytics, credit scoring analytics, and text analytics.

He is interested in all aspects of distributed machine learning and the Internet of Things. Oleg currently lives and works in Hamburg, Germany.

I would like to express my deepest gratitude to my parents for everything that they have done for me.

www.PacktPub.com

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://www.packtpub.com/mapt`

Get the most in-demand software skills with Mapt. Mapt gives you full access to all Packt books and video courses, as well as industry-leading tools to help you plan your personal development and advance your career.

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

Customer Feedback

Thanks for purchasing this Packt book. At Packt, quality is at the heart of our editorial process. To help us improve, please leave us an honest review on this book's Amazon page at `https://www.amazon.com/dp/1788299876`.

If you'd like to join our team of regular reviewers, you can e-mail us at `customerreviews@packtpub.com`. We award our regular reviewers with free eBooks and videos in exchange for their valuable feedback. Help us be relentless in improving our products!

Table of Contents

Preface

In recent years, popular imagination has become fascinated by machine learning. The discipline has found a variety of applications. Some of these applications, such as spam filtering, are ubiquitous and have been rendered mundane by their successes. 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 we will 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 state-of-the-art implementations of many machine learning algorithms under an intuitive and versatile API.

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, we will examine a machine learning model, apply it to a task, and measure its performance.

`Chapter 2`, *Simple Linear Regression*, discusses a model that relates a single feature to a continuous response variable. We will learn about cost functions and use the normal equation to optimize the model.

`Chapter 3`, *Classification and Regression with K-Nearest Neighbors*, introduces a simple, nonlinear model for classification and regression tasks.

`Chapter 4`, *Feature Extraction*, describes methods for representing text, images, and categorical variables as features that can be used in machine learning models.

`Chapter 5`, *From Simple Linear Regression to Multiple Linear Regression*, discusses a generalization of simple linear regression that regresses a continuous response variable onto multiple features.

`Chapter 6`, *From Linear Regression to Logistic Regression*, further generalizes multiple linear regression and introduces a model for binary classification tasks.

`Chapter 7`, *Naive Bayes*, discusses Bayes' theorem and the Naive Bayes family of classifiers, and compares generative and discriminative models.

`Chapter 8`, *Nonlinear Classification and Regression with Decision Trees*, introduces the decision tree, a simple, nonlinear model for classification and regression tasks.

`Chapter 9`, *From Decision Trees to Random Forests and other Ensemble Methods*, discusses three methods for combining models called bagging, boosting, and stacking.

`Chapter 10`, *The Perceptron*, introduces a simple online model for binary classification.

`Chapter 11`, *From the Perceptron to Support Vector Machines*, discusses a powerful, discriminative model for classification and regression called the support vector machine, and a technique for efficiently projecting features to higher dimensional spaces.

`Chapter 12`, *From the Perceptron to Artificial Neural Networks*, introduces powerful nonlinear models for classification and regression built from graphs of artificial neurons.

`Chapter 13`, *K-means*, discusses an algorithm that can be used to find structures in unlabeled data.

`Chapter 14`, *Dimensionality Reduction with Principal Component Analysis*, describes a method for reducing the dimensions of data that can mitigate the curse of dimensionality.

What you need for this book

The examples in this book require Python >= 2.7 or >= 3.3 and pip, the PyPA recommended tool for installing Python packages. The examples are intended to be executed in a Jupyter notebook or an IPython interpreter. `Chapter 1`, *The Fundamentals of Machine Learning* shows how to install scikit-learn 0.18.1, its dependencies, and other libraries on Ubuntu, Mac OS, and Windows.

Who this book is for

This book is intended for software engineers who want to understand how common machine learning algorithms work and develop an intuition for how to use them. It is also for data scientists who want to learn about the scikit-learn API. Familiarity with machine learning fundamentals and Python is helpful but not required.