

Robert Layton

Learning Data Mining with Python

Second Edition

Use Python to manipulate data and build
predictive models



Packt>

Learning Data Mining with Python

- Second Edition

The next step in the information age is to gain insights from the deluge of data coming our way. Data mining provides a way of finding these insights, and Python is one of the most popular languages for data mining, providing both power and flexibility in analysis.

This book teaches you how to design and develop data mining applications using a variety of datasets, starting with basic classification and affinity analysis. This book covers a large number of libraries available in Python, including the Jupyter Notebook, pandas, scikit-learn, spaCy and Tensorflow.

You will gain hands-on experience with complex data types including text, images, and graphs. You will also discover object detection using Deep Neural Networks, which is one of the big and difficult areas of machine learning right now.

With restructured examples and code samples updated for the latest edition of Python, each chapter of this book introduces you to new algorithms and techniques. By the end of the book, you will have a great insight into using Python for data mining and understanding of the algorithms as well as implementations.

Things you will learn:

- Apply data mining concepts to real-world problems
- Predict the outcome of sports matches based on past results
- Determine the author of a document based on their writing style
- Use APIs to download datasets from social media and other online services
- Find and extract good features from difficult datasets
- Create models that solve real-world problems
- Design and develop data mining applications using a variety of datasets
- Perform object detection in images using Deep Neural Networks
- Get meaningful insights from your data through intuitive visualizations
- Compute on big data, including real-time data from the Internet

Packt
www.packtpub.com

\$ 44.99 US
£ 37.99 UK

Prices do not include local sales
Tax or VAT where applicable



Learning Data Mining with Python - Second Edition

Robert Layton



Learning Data Mining with Python

Second Edition

Use Python to manipulate data and build predictive models

Robert Layton

Packt>

BIRMINGHAM - MUMBAI

Learning Data Mining with Python

Second Edition

Copyright © 2017 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor Packt Publishing, and its dealers and distributors will be held liable for any damages caused or alleged to be caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

First published: July 2015

Second edition: April 2017

Production reference: 1250417

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham
B3 2PB, UK.

ISBN 978-1-78712-678-7

www.packtpub.com

Credits

Author

Robert Layton

Reviewer

Asad Ahamad

Commissioning Editor

Veena Pagare

Acquisition Editor

Divya Poojari

Content Development Editor

Tejas Limkar

Technical Editor

Danish Shaikh

Copy Editor

Vikrant Phadkay

Project Coordinator

Nidhi Joshi

Proofreader

Safis Editing

Indexer

Mariammal Chettiyar

Graphics

Tania Dutta

Production Coordinator

Aparna Bhagat

About the Author

Robert Layton is a data scientist investigating data-driven applications to businesses across a number of sectors. He received a PhD investigating cybercrime analytics from the Internet Commerce Security Laboratory at Federation University Australia, before moving into industry, starting his own data analytics company dataPipeline (www.datapipeline.com.au). Next, he created Eureactive (www.eureactive.com.au), which works with tech-based startups on developing their proof-of-concepts and early-stage prototypes. Robert also runs www.learningtensorflow.com, which is one of the world's premier tutorial websites for Google's TensorFlow library.

Robert is an active member of the Python community, having used Python for more than 8 years. He has presented at PyConAU for the last four years and works with Python Charmers to provide Python-based training for businesses and professionals from a wide range of organisations.

Robert can be best reached via Twitter @robertlayton

Thank you to my family for supporting me on this journey, thanks to all the readers of revision 1 for making it a success, and thanks to Matty for his assistance behind-the-scenes with the book.

About the Reviewer

Asad Ahamad is a data enthusiast and loves to work on data to solve challenging problems.

He did his masters in Industrial Mathematics with Computer Application from Jamia Millia Islamia, New Delhi. He admires Mathematics a lot and always tries to use it to gain maximum profit for business.

He has good experience working on data mining, machine learning and data science and worked for various multinationals in India. He mainly uses R and Python to perform data wrangling and modeling. He is fond of using open source tools for data analysis.

He is active social media user. Feel free to connect him on twitter @asadtaj88

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/1787126781>.

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	1
Chapter 1: Getting Started with Data Mining	7
Introducing data mining	7
Using Python and the Jupyter Notebook	9
Installing Python	10
Installing Jupyter Notebook	11
Installing scikit-learn	13
A simple affinity analysis example	14
What is affinity analysis?	14
Product recommendations	14
Loading the dataset with NumPy	15
Downloading the example code	17
Implementing a simple ranking of rules	18
Ranking to find the best rules	21
A simple classification example	23
What is classification?	24
Loading and preparing the dataset	24
Implementing the OneR algorithm	26
Testing the algorithm	28
Summary	31
Chapter 2: Classifying with scikit-learn Estimators	33
scikit-learn estimators	33
Nearest neighbors	34
Distance metrics	35
Loading the dataset	38
Moving towards a standard workflow	40
Running the algorithm	41
Setting parameters	42
Preprocessing	44
Standard pre-processing	46
Putting it all together	47
Pipelines	47
Summary	49

Chapter 3: Predicting Sports Winners with Decision Trees	51
Loading the dataset	51
Collecting the data	52
Using pandas to load the dataset	53
Cleaning up the dataset	54
Extracting new features	56
Decision trees	58
Parameters in decision trees	59
Using decision trees	60
Sports outcome prediction	61
Putting it all together	61
Random forests	65
How do ensembles work?	66
Setting parameters in Random Forests	67
Applying random forests	67
Engineering new features	69
Summary	70
Chapter 4: Recommending Movies Using Affinity Analysis	71
Affinity analysis	72
Algorithms for affinity analysis	73
Overall methodology	74
Dealing with the movie recommendation problem	74
Obtaining the dataset	75
Loading with pandas	75
Sparse data formats	76
Understanding the Apriori algorithm and its implementation	77
Looking into the basics of the Apriori algorithm	79
Implementing the Apriori algorithm	80
Extracting association rules	83
Evaluating the association rules	86
Summary	89
Chapter 5: Features and scikit-learn Transformers	91
Feature extraction	92
Representing reality in models	92
Common feature patterns	95
Creating good features	99
Feature selection	100
Selecting the best individual features	102

Feature creation	105
Principal Component Analysis	108
Creating your own transformer	111
The transformer API	111
Implementing a Transformer	112
Unit testing	113
Putting it all together	114
Summary	115
Chapter 6: Social Media Insight using Naive Bayes	117
Disambiguation	118
Downloading data from a social network	119
Loading and classifying the dataset	121
Creating a replicable dataset from Twitter	125
Text transformers	129
Bag-of-words models	129
n-gram features	131
Other text features	132
Naive Bayes	133
Understanding Bayes' theorem	133
Naive Bayes algorithm	134
How it works	135
Applying of Naive Bayes	137
Extracting word counts	137
Converting dictionaries to a matrix	138
Putting it all together	139
Evaluation using the F1-score	140
Getting useful features from models	141
Summary	144
Chapter 7: Follow Recommendations Using Graph Mining	145
Loading the dataset	145
Classifying with an existing model	147
Getting follower information from Twitter	150
Building the network	152
Creating a graph	155
Creating a similarity graph	157
Finding subgraphs	161
Connected components	161
Optimizing criteria	165

Summary	168
Chapter 8: Beating CAPTCHAs with Neural Networks	171
Artificial neural networks	172
An introduction to neural networks	173
Creating the dataset	175
Drawing basic CAPTCHAs	176
Splitting the image into individual letters	179
Creating a training dataset	182
Training and classifying	184
Back-propagation	187
Predicting words	188
Improving accuracy using a dictionary	193
Ranking mechanisms for word similarity	193
Putting it all together	194
Summary	195
Chapter 9: Authorship Attribution	197
Attributing documents to authors	198
Applications and use cases	199
Authorship attribution	200
Getting the data	202
Using function words	205
Counting function words	206
Classifying with function words	209
Support Vector Machines	210
Classifying with SVMs	211
Kernels	212
Character n-grams	212
Extracting character n-grams	213
The Enron dataset	214
Accessing the Enron dataset	215
Creating a dataset loader	215
Putting it all together	218
Evaluation	219
Summary	221
Chapter 10: Clustering News Articles	223
Trending topic discovery	224
Using a web API to get data	224
Reddit as a data source	227

Getting the data	228
Extracting text from arbitrary websites	231
Finding the stories in arbitrary websites	231
Extracting the content	233
Grouping news articles	235
The k-means algorithm	236
Evaluating the results	239
Extracting topic information from clusters	242
Using clustering algorithms as transformers	243
Clustering ensembles	244
Evidence accumulation	244
How it works	248
Implementation	250
Online learning	251
Implementation	252
Summary	255
Chapter 11: Object Detection in Images using Deep Neural Networks	257
Object classification	258
Use cases	258
Application scenario	260
Deep neural networks	263
Intuition	263
Implementing deep neural networks	265
An Introduction to TensorFlow	266
Using Keras	270
Convolutional Neural Networks	275
GPU optimization	277
When to use GPUs for computation	278
Running our code on a GPU	279
Setting up the environment	280
Application	281
Getting the data	282
Creating the neural network	283
Putting it all together	285
Summary	286
Chapter 12: Working with Big Data	289
Big data	290
Applications of big data	291

MapReduce	293
The intuition behind MapReduce	295
A word count example	297
Hadoop MapReduce	299
Applying MapReduce	300
Getting the data	300
Naive Bayes prediction	302
The mrjob package	302
Extracting the blog posts	303
Training Naive Bayes	305
Putting it all together	309
Training on Amazon's EMR infrastructure	314
Summary	318
Appendix: Next Steps...	319
Getting Started with Data Mining	319
Scikit-learn tutorials	319
Extending the Jupyter Notebook	320
More datasets	320
Other Evaluation Metrics	320
More application ideas	320
Classifying with scikit-learn Estimators	321
Scalability with the nearest neighbor	321
More complex pipelines	321
Comparing classifiers	322
Automated Learning	322
Predicting Sports Winners with Decision Trees	323
More complex features	323
Dask	324
Research	324
Recommending Movies Using Affinity Analysis	324
New datasets	324
The Eclat algorithm	325
Collaborative Filtering	325
Extracting Features with Transformers	325
Adding noise	325
Vowpal Wabbit	326
word2vec	326
Social Media Insight Using Naive Bayes	326
Spam detection	326