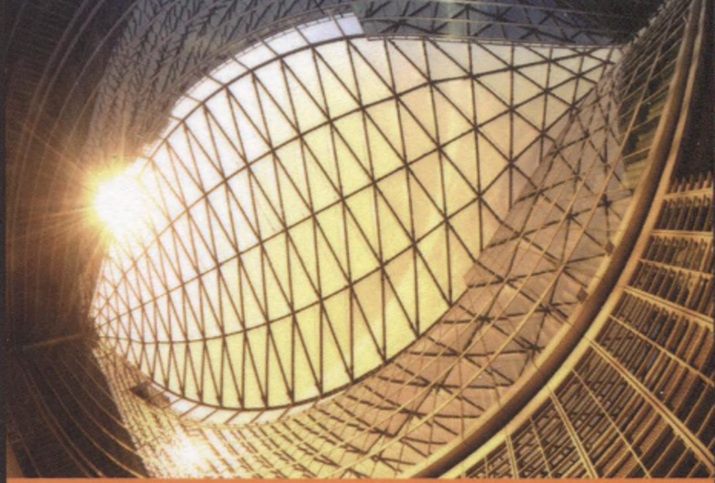

Learning Data Mining with Python

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 this insight, and Python is one of the most popular languages for data mining, providing both power and flexibility in analysis.

This book teaches you to design and develop data mining applications using a variety of datasets, starting with basic classification and affinity analysis. Next, we move on to more complex data types including text, images, and graphs. In every chapter, we create models that solve real-world problems.

There is a rich and varied set of libraries available in Python for data mining. This book covers a large number, including the IPython Notebook, pandas, scikit-learn and NLTK.

Each chapter of this book introduces you to new algorithms and techniques. By the end of the book, you will gain a large insight into using Python for data mining, with a good knowledge and understanding of the algorithms and implementations.

Who this book is written for

If you are a programmer who wants to get started with data mining, then this book is for you.

What you will learn from this book

- 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
- Set up reproducible experiments and generate robust results
- Recommend movies, online celebrities, and news articles based on personal preferences
- Compute on big data, including real-time data from the Internet

$ 44.99 US
£ 28.99 UK

Prices do not include
local sales tax or VAT
where applicable

ISBN 978-1-78439-605-3
54499
9 781784 396053

[PACKT] PUBLISHING open source*
community experience distilled

Visit **www.PacktPub.com** for books, eBooks, code, downloads, and PacktLib.

Preface ix
Chapter 1: Getting Started with Data Mining 1
Introducing data mining 2
Using Python and the IPython Notebook 3
Installing Python 3
Installing IPython 5
Installing scikit-learn 6
A simple affinity analysis example 7
What is affinity analysis? 7
Product recommendations 8
Loading the dataset with NumPy 8
Implementing a simple ranking of rules 10
Ranking to find the best rules 13
A simple classification example 16
What is classification? 16
Loading and preparing the dataset 16
Implementing the OneR algorithm 18
Testing the algorithm 20
Summary 23
Chapter 2: Classifying with scikit-learn Estimators 25
scikit-learn estimators 25
Nearest neighbors 26
Distance metrics 27
Loading the dataset 29
Moving towards a standard workflow 31
Running the algorithm 32
Setting parameters 33

Preprocessing using pipelines **35**
An example 36
Standard preprocessing 37
Putting it all together 38
Pipelines **38**
Summary **40**
Chapter 3: Predicting Sports Winners with Decision Trees **41**
Loading the dataset **41**
Collecting the data 42
Using pandas to load the dataset 43
Cleaning up the dataset 44
Extracting new features 45
Decision trees **47**
Parameters in decision trees 48
Using decision trees 49
Sports outcome prediction **49**
Putting it all together 50
Random forests **54**
How do ensembles work? 55
Parameters in Random forests 56
Applying Random forests 56
Engineering new features 58
Summary **59**
Chapter 4: Recommending Movies Using Affinity Analysis **61**
Affinity analysis **61**
Algorithms for affinity analysis 62
Choosing parameters 63
The movie recommendation problem **64**
Obtaining the dataset 64
Loading with pandas 64
Sparse data formats 65
The Apriori implementation **66**
The Apriori algorithm 68
Implementation 69
Extracting association rules **72**
Evaluation 76
Summary **79**

Chapter 5: Extracting Features with Transformers 81

Feature extraction 82

Representing reality in models 82

Common feature patterns 84

Creating good features 87

Feature selection 88

Selecting the best individual features 90

Feature creation 93

Principal Component Analysis 96

Creating your own transformer 98

The transformer API 99

Implementation details 99

Unit testing 101

Putting it all together 103

Summary 103

Chapter 6: Social Media Insight Using Naive Bayes 105

Disambiguation 106

Downloading data from a social network 107

Loading and classifying the dataset 109

Creating a replicable dataset from Twitter 114

Text transformers 118

Bag-of-words 118

N-grams 120

Other features 121

Naive Bayes 121

Bayes' theorem 122

Naive Bayes algorithm 123

How it works 124

Application 126

Extracting word counts 126

Converting dictionaries to a matrix 127

Training the Naive Bayes classifier 127

Putting it all together 128

Evaluation using the F1-score 129

Getting useful features from models 130

Summary 133

Chapter 7: Discovering Accounts to Follow Using Graph Mining 135
Loading the dataset 136
Classifying with an existing model 137
Getting follower information from Twitter 140
Building the network 142
Creating a graph 145
Creating a similarity graph 147
Finding subgraphs 151
Connected components 151
Optimizing criteria 155
Summary 159
Chapter 8: Beating CAPTCHAs with Neural Networks 161
Artificial neural networks 162
An introduction to neural networks 163
Creating the dataset 165
Drawing basic CAPTCHAs 166
Splitting the image into individual letters 167
Creating a training dataset 169
Adjusting our training dataset to our methodology 171
Training and classifying 172
Back propagation 173
Predicting words 175
Improving accuracy using a dictionary 180
Ranking mechanisms for words 180
Putting it all together 181
Summary 183
Chapter 9: Authorship Attribution 185
Attributing documents to authors 186
Applications and use cases 186
Attributing authorship 187
Getting the data 189
Function words 192
Counting function words 193
Classifying with function words 195
Support vector machines 196
Classifying with SVMs 197
Kernels 198
Character n-grams 198
Extracting character n-grams 199

Using the Enron dataset **200**
Accessing the Enron dataset 201
Creating a dataset loader 201
Putting it all together 206
Evaluation 207
Summary **209**

Chapter 10: Clustering News Articles **211**

Obtaining news articles **211**
Using a Web API to get data 212
Reddit as a data source 215
Getting the data 216
Extracting text from arbitrary websites **218**
Finding the stories in arbitrary websites 218
Putting it all together 220
Grouping news articles **222**
The k-means algorithm 223
Evaluating the results 226
Extracting topic information from clusters 228
Using clustering algorithms as transformers 229
Clustering ensembles **230**
Evidence accumulation 230
How it works 234
Implementation 235
Online learning **236**
An introduction to online learning 236
Implementation 237
Summary **240**

Chapter 11: Classifying Objects in Images Using Deep Learning **241**

Object classification **242**
Application scenario and goals **242**
Use cases 245
Deep neural networks **246**
Intuition 247
Implementation 247
An introduction to Theano 248
An introduction to Lasagne 249
Implementing neural networks with nolearn 254

GPU optimization **258**
When to use GPUs for computation 258
Running our code on a GPU 259
Setting up the environment **261**
Application **263**
Getting the data 263
Creating the neural network 264
Putting it all together 267
Summary **269**
Chapter 12: Working with Big Data **271**
Big data **272**
Application scenario and goals **273**
MapReduce **274**
Intuition 275
A word count example 276
Hadoop MapReduce 278
Application **279**
Getting the data 280
Naive Bayes prediction 282
The mrjob package 282
Extracting the blog posts 283
Training Naive Bayes 285
Putting it all together 288
Training on Amazon's EMR infrastructure 293
Summary **296**
Appendix: Next Steps… **297**
Chapter 1 – Getting Started with Data Mining **297**
Scikit-learn tutorials 297
Extending the IPython Notebook 298
Chapter 2 – Classifying with scikit-learn Estimators **298**
Scalability with the nearest neighbor 298
More complex pipelines 299
Comparing classifiers 299
Chapter 3: Predicting Sports Winners with Decision Trees **300**
More on pandas 300
More complex features 300
Chapter 4 – Recommending Movies Using Affinity Analysis **301**
New datasets 301
The Eclat algorithm 301

Chapter 5 – Extracting Features with Transformers **301**
Adding noise 301
Vowpal Wabbit 302
Chapter 6 – Social Media Insight Using Naive Bayes **302**
Spam detection 302
Natural language processing and part-of-speech tagging 302
Chapter 7 – Discovering Accounts to Follow Using Graph Mining **303**
More complex algorithms 303
NetworkX 303
Chapter 8 – Beating CAPTCHAs with Neural Networks **303**
Better (worse?) CAPTCHAs 303
Deeper networks 304
Reinforcement learning 304
Chapter 9 – Authorship Attribution **304**
Increasing the sample size 304
Blogs dataset 304
Local n-grams 305
Chapter 10 – Clustering News Articles **305**
Evaluation 305
Temporal analysis 305
Real-time clusterings 306
Chapter 11: Classifying Objects in Images Using Deep Learning **306**
Keras and Pylearn2 306
Mahotas 306
Chapter 12 – Working with Big Data **307**
Courses on Hadoop 307
Pydoop 307
Recommendation engine 307
More resources **308**
Index **309**