
Undergraduate Topics in Computer Science

Series editor

Ian Mackie

Advisory Board

Samson Abramsky, University of Oxford, Oxford, UK

Karin Breitman, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil

Chris Hankin, Imperial College London, London, UK

Dexter Kozen, Cornell University, Ithaca, USA

Andrew Pitts, University of Cambridge, Cambridge, UK

Hanne Riis Nielson, Technical University of Denmark, Kongens Lyngby, Denmark

Steven Skiena, Stony Brook University, Stony Brook, USA

Iain Stewart, University of Durham, Durham, UK

Undergraduate Topics in Computer Science (UTiCS) delivers high-quality instructional content for undergraduates studying in all areas of computing and information science. From core foundational and theoretical material to final-year topics and applications, UTiCS books take a fresh, concise, and modern approach and are ideal for self-study or for a one- or two-semester course. The texts are all authored by established experts in their fields, reviewed by an international advisory board, and contain numerous examples and problems. Many include fully worked solutions.

More information about this series at <http://www.springer.com/series/7592>

Laura Igual · Santi Seguí

Introduction to Data Science

A Python Approach to Concepts,
Techniques and Applications

With contributions from Jordi Vitrià, Eloi Puertas
Petia Radeva, Oriol Pujol, Sergio Escalera, Francesc Dantí
and Lluís Garrido

Laura Igual
Departament de Matemàtiques i Informàtica
Universitat de Barcelona
Barcelona
Spain

Santi Seguí
Departament de Matemàtiques i Informàtica
Universitat de Barcelona
Barcelona
Spain

With contributions from Jordi Vitrià, Eloi Puertas, Petia Radeva, Oriol Pujol, Sergio Escalera, Francesc Dantí and Lluís Garrido

ISSN 1863-7310

ISSN 2197-1781 (electronic)

Undergraduate Topics in Computer Science

ISBN 978-3-319-50016-4

ISBN 978-3-319-50017-1 (eBook)

DOI 10.1007/978-3-319-50017-1

Library of Congress Control Number: 2016962046

© Springer International Publishing Switzerland 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature

The registered company is Springer International Publishing AG

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Subject Area of the Book

In this era, where a huge amount of information from different fields is gathered and stored, its analysis and the extraction of value have become one of the most attractive tasks for companies and society in general. The design of solutions for the new questions emerged from data has required multidisciplinary teams. Computer scientists, statisticians, mathematicians, biologists, journalists and sociologists, as well as many others are now working together in order to provide knowledge from data. This new interdisciplinary field is called *data science*.

The pipeline of any data science goes through asking the right questions, gathering data, cleaning data, generating hypothesis, making inferences, visualizing data, assessing solutions, etc.

Organization and Feature of the Book

This book is an introduction to concepts, techniques, and applications in data science. This book focuses on the analysis of data, covering concepts from statistics to machine learning, techniques for graph analysis and parallel programming, and applications such as recommender systems or sentiment analysis.

All chapters introduce new concepts that are illustrated by practical cases using real data. Public databases such as Eurostat, different social networks, and MovieLens are used. Specific questions about the data are posed in each chapter. The solutions to these questions are implemented using Python programming language and presented in code boxes properly commented. This allows the reader to learn data science by solving problems which can generalize to other problems.

This book is not intended to cover the whole set of data science methods neither to provide a complete collection of references. Currently, data science is an increasing and emerging field, so readers are encouraged to look for specific methods and references using keywords in the net.

Target Audiences

This book is addressed to upper-tier undergraduate and beginning graduate students from technical disciplines. Moreover, this book is also addressed to professional audiences following continuous education short courses and to researchers from diverse areas following self-study courses.

Basic skills in computer science, mathematics, and statistics are required. Code programming in Python is of benefit. However, even if the reader is new to Python, this should not be a problem, since acquiring the Python basics is manageable in a short period of time.

Previous Uses of the Materials

Parts of the presented materials have been used in the postgraduate course of *Data Science and Big Data* from Universitat de Barcelona. All contributing authors are involved in this course.

Suggested Uses of the Book

This book can be used in any introductory data science course. The problem-based approach adopted to introduce new concepts can be useful for the beginners. The implemented code solutions for different problems are a good set of exercises for the students. Moreover, these codes can serve as a baseline when students face bigger projects.

Supplemental Resources

This book is accompanied by a set of IPython Notebooks containing all the codes necessary to solve the practical cases of the book. The Notebooks can be found on the following GitHub repository: <https://github.com/DataScienceUB/introduction-datascience-python-book>.

Acknowledgements

We acknowledge all the contributing authors: J. Vitrià, E. Puertas, P. Radeva, O. Pujol, S. Escalera, L. Garrido, and F. Dantí.

Barcelona, Spain

Laura Igual
Santi Seguí

Contents

1	Introduction to Data Science	1
1.1	What is Data Science?	1
1.2	About This Book	3
2	Toolboxes for Data Scientists	5
2.1	Introduction	5
2.2	Why Python?	6
2.3	Fundamental Python Libraries for Data Scientists	6
2.3.1	Numeric and Scientific Computation: NumPy and SciPy	7
2.3.2	SCIKIT-Learn: Machine Learning in Python	7
2.3.3	PANDAS: Python Data Analysis Library	7
2.4	Data Science Ecosystem Installation	7
2.5	Integrated Development Environments (IDE)	8
2.5.1	Web Integrated Development Environment (WIDE): Jupyter	9
2.6	Get Started with Python for Data Scientists	10
2.6.1	Reading	14
2.6.2	Selecting Data	16
2.6.3	Filtering Data	17
2.6.4	Filtering Missing Values	17
2.6.5	Manipulating Data	18
2.6.6	Sorting	22
2.6.7	Grouping Data	23
2.6.8	Rearranging Data	24
2.6.9	Ranking Data	25
2.6.10	Plotting	26
2.7	Conclusions	28
3	Descriptive Statistics	29
3.1	Introduction	29
3.2	Data Preparation	30
3.2.1	The Adult Example	30

3.3	Exploratory Data Analysis	32
3.3.1	Summarizing the Data	32
3.3.2	Data Distributions	36
3.3.3	Outlier Treatment	38
3.3.4	Measuring Asymmetry: Skewness and Pearson's Median Skewness Coefficient	41
3.3.5	Continuous Distribution	42
3.3.6	Kernel Density	44
3.4	Estimation	46
3.4.1	Sample and Estimated Mean, Variance and Standard Scores	46
3.4.2	Covariance, and Pearson's and Spearman's Rank Correlation.	47
3.5	Conclusions	50
	References	50
4	Statistical Inference	51
4.1	Introduction	51
4.2	Statistical Inference: The Frequentist Approach	52
4.3	Measuring the Variability in Estimates.	52
4.3.1	Point Estimates	53
4.3.2	Confidence Intervals	56
4.4	Hypothesis Testing.	59
4.4.1	Testing Hypotheses Using Confidence Intervals	60
4.4.2	Testing Hypotheses Using p -Values	61
4.5	But Is the Effect E Real?	64
4.6	Conclusions	64
	References	65
5	Supervised Learning.	67
5.1	Introduction	67
5.2	The Problem	68
5.3	First Steps	69
5.4	What Is Learning?	78
5.5	Learning Curves.	79
5.6	Training, Validation and Test.	82
5.7	Two Learning Models	86
5.7.1	Generalities Concerning Learning Models	86
5.7.2	Support Vector Machines	87
5.7.3	Random Forest	90
5.8	Ending the Learning Process	91
5.9	A Toy Business Case.	92
5.10	Conclusion	95
	Reference	96

6	Regression Analysis	97
6.1	Introduction	97
6.2	Linear Regression	98
6.2.1	Simple Linear Regression	98
6.2.2	Multiple Linear Regression and Polynomial Regression	103
6.2.3	Sparse Model	104
6.3	Logistic Regression	110
6.4	Conclusions	113
	References	114
7	Unsupervised Learning	115
7.1	Introduction	115
7.2	Clustering.	116
7.2.1	Similarity and Distances	117
7.2.2	What Constitutes a Good Clustering? Defining Metrics to Measure Clustering Quality	117
7.2.3	Taxonomies of Clustering Techniques	120
7.3	Case Study.	132
7.4	Conclusions	138
	References	139
8	Network Analysis	141
8.1	Introduction	141
8.2	Basic Definitions in Graphs	142
8.3	Social Network Analysis	144
8.3.1	Basics in NetworkX	144
8.3.2	Practical Case: Facebook Dataset	145
8.4	Centrality	147
8.4.1	Drawing Centrality in Graphs	152
8.4.2	PageRank	154
8.5	Ego-Networks	157
8.6	Community Detection	162
8.7	Conclusions	163
	References	164
9	Recommender Systems	165
9.1	Introduction	165
9.2	How Do Recommender Systems Work?	166
9.2.1	Content-Based Filtering	166
9.2.2	Collaborative Filtering	167
9.2.3	Hybrid Recommenders	167
9.3	Modeling User Preferences	167
9.4	Evaluating Recommenders	168

9.5	Practical Case	169
9.5.1	MovieLens Dataset	169
9.5.2	User-Based Collaborative Filtering	171
9.6	Conclusions	179
	References	179
10	Statistical Natural Language Processing for Sentiment	
	Analysis	181
10.1	Introduction	181
10.2	Data Cleaning	182
10.3	Text Representation	185
10.3.1	Bi-Grams and n-Grams	190
10.4	Practical Cases	191
10.5	Conclusions	196
	References	196
11	Parallel Computing	199
11.1	Introduction	199
11.2	Architecture	200
11.2.1	Getting Started	201
11.2.2	Connecting to the Cluster (The Engines)	202
11.3	Multicore Programming	203
11.3.1	Direct View of Engines	203
11.3.2	Load-Balanced View of Engines	206
11.4	Distributed Computing	207
11.5	A Real Application: New York Taxi Trips	208
11.5.1	A Direct View Non-Blocking Proposal	209
11.5.2	Results	212
11.6	Conclusions	214
	References	215
	Index	217

Authors and Contributors

About the Authors

Dr. Laura Igual is an associate professor from the Department of Mathematics and Computer Science at the Universitat de Barcelona. She received a degree in mathematics from Universitat de Valencia (Spain) in 2000 and a Ph.D. degree from the Universitat Pompeu Fabra (Spain) in 2006. Her particular areas of interest include computer vision, medical imaging, machine learning, and data science.

Dr. Laura Igual is coauthor of Chaps. 3, 6, and 8.

Dr. Santi Seguí is an assistant professor from the Department of Mathematics and Computer Science at the Universitat de Barcelona. He is a computer science engineer by the Universitat Autònoma de Barcelona (Spain) since 2007. He received his Ph.D. degree from the Universitat de Barcelona (Spain) in 2011. His particular areas of interest include computer vision, applied machine learning, and data science.

Dr. Santi Seguí is coauthor of Chaps. 8–10.

Contributors

Francesc Dantí is an adjunct professor and system administrator from the Department of Mathematics and Computer Science at the Universitat de Barcelona. He is a computer science engineer by the Universitat Oberta de Catalunya (Spain). His particular areas of interest are HPC and grid computing, parallel computing, and cybersecurity.

Francesc Dantí is coauthor of Chaps. 2 and 11.

Dr. Sergio Escalera is an associate professor from the Department of Mathematics and Computer Science at the Universitat de Barcelona. He is a computer science engineer by the Universitat Autònoma de Barcelona (Spain) since 2003. He received his Ph.D. degree from the Universitat Autònoma de Barcelona (Spain) in 2008. His research interests include, between others, statistical pattern recognition,

visual object recognition, with special interest in human pose recovery and behavior analysis from multimodal data.

Dr. Sergio Escalera is coauthor of Chaps. 4 and 10.

Dr. Lluís Garrido is an associate professor from the Department of Mathematics and Computer Science at the Universitat de Barcelona. He is a telecommunications engineer by the Universitat Politècnica de Catalunya (UPC) since 1996. He received his Ph.D. degree from the same university in 2002. His particular areas of interest include computer vision, image processing, numerical optimization, parallel computing, and data science.

Dr. Lluís Garrido is coauthor of Chap. 11.

Dr. Eloi Puertas is an assistant professor from the Department of Mathematics and Computer Science at the Universitat de Barcelona. He is a computer science engineer by the Universitat Autònoma de Barcelona (Spain) since 2002. He received his Ph.D. degree from the Universitat de Barcelona (Spain) in 2014. His particular areas of interest include artificial intelligence, software engineering, and data science.

Dr. Eloi Puertas is coauthor of Chaps. 2 and 9.

Dr. Oriol Pujol is a tenured associate professor from the Department of Mathematics and Computer Science at the Universitat de Barcelona. He received his Ph.D. degree from the Universitat Autònoma de Barcelona (Spain) in 2004 for his work in machine learning and computer vision. His particular areas of interest include machine learning, computer vision, and data science.

Dr. Oriol Pujol is coauthor of Chaps. 5 and 7.

Dr. Petia Radeva is a tenured associate professor and senior researcher from the Universitat de Barcelona. She graduated in applied mathematics and computer science in 1989 at the University of Sofia, Bulgaria, and received her Ph.D. degree in Computer Vision for Medical Imaging in 1998 from the Universitat Autònoma de Barcelona, Spain. She is Icrea Academia Researcher from 2015, head of the Consolidated Research Group “Computer Vision at the Universitat of Barcelona,” and head of MiLab of Computer Vision Center. Her present research interests are on the development of learning-based approaches for computer vision, deep learning, egocentric vision, lifelogging, and data science.

Dr. Petia Radeva is coauthor of Chaps. 3, 5, and 7.

Dr. Jordi Vitrià is a full professor from the Department of Mathematics and Computer Science at the Universitat de Barcelona. He received his Ph.D. degree from the Universitat Autònoma de Barcelona in 1990. Dr. Jordi Vitrià has published more than 100 papers in SCI-indexed journals and has more than 25 years of experience in working on computer vision and artificial intelligence and its applications to several fields. He is now leader of the “Data Science Group at UB,” a technology transfer unit that performs collaborative research projects between the Universitat de Barcelona and private companies.

Dr. Jordi Vitrià is coauthor of Chaps. 1, 4, and 6.