Undergraduate Topics in Computer Science

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

Springer

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

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

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.