Texts in Computer Science

Titles in this series now included in the Thomson Reuters Book Citation Index!
'Texts in Computer Science' (TCS) delivers high-quality instructional content for undergraduates and graduates in all areas of computing and information science, with a strong emphasis on core foundational and theoretical material but inclusive of some prominent applications-related content. TCS books should be reasonably self-contained and aim to provide students with modern and clear accounts of topics ranging across the computing curriculum. As a result, the books are ideal for semester courses or for individual self-study in cases where people need to expand their knowledge. All texts are authored by established experts in their fields, reviewed internally and by the series editors, and provide numerous examples, problems, and other pedagogical tools; many contain fully worked solutions.
The TCS series is comprised of high-quality, self-contained books that have broad and comprehensive coverage and are generally in hardback format and sometimes contain color. For undergraduate textbooks that are likely to be more brief and modular in their approach, require only black and white, and are under 275 pages, Springer offers the flexibly designed Undergraduate Topics in Computer Science series, to which we refer potential authors.

More information about this series at
www.springer.com/series/3191

Michael R. Berthold · Christian Borgelt ·
Frank Höppner · Frank Klawonn ·
Rosaria Silipo

Guide to Intelligent Data Science

How to Intelligently Make Use of Real Data

Second Edition

Springer

Michael R. Berthold
Department of Computer and Information Science
University of Konstanz
Konstanz, Germany

Christian Borgelt
Department of Computer Sciences
University of Salzburg
Salzburg, Austria

Series Editors
David Gries
Department of Computer Science
Cornell University
Ithaca, NY, USA

Frank Höppner
Department of Computer Science
Ostfalia University of Applied Sciences
Wolfenbüttel, Germany

Frank Klawonn
Department of Computer Science
Ostfalia University of Applied Sciences
Wolfenbüttel, Germany

Rosaria Silipo
KNIME AG
Zurich, Switzerland

Orit Hazzan
Faculty of Education in Technology and Science
Technion – Israel Institute of Technology
Haifa, Israel

ISSN 1868-0941
Texts in Computer Science
ISBN 978-3-030-45573-6
https://doi.org/10.1007/978-3-030-45574-3

ISSN 1868-095X (electronic)

ISBN 978-3-030-45574-3 (eBook)

Preface

The interest in making sense of data has become central to pretty much every business and project in the past decade. When we wrote the first edition of this volume, it was aimed at the people whose task it was to analyze real world data—often in a side department, prepared to solve problems given to them. In the meantime, data science use has spread from those isolated groups to the entire organization. This fits our goal from the first edition even more: providing different levels of detail for each section of the book. Readers more interested in the concepts and how to apply methods, focus more on the first half of each chapter. Readers, also interested in the underpinnings of the presented methods and algorithms, find the theory in the middle. And for both audiences, the end of most chapters shows practical examples of how to apply these to real data.

Still, given the advance in research we decided it worth revising, updating, and expanding the content to also cover select new methods and tools that have been invented in the meantime. We also adjusted to the new way of thinking about data analysis and how many organizations now see this in the bigger context of Data Science, as it is visible by the many universities adding Data Science curriculum or even entire Data Science departments. As the original volume covered the entire process from data ingestion to deployment and management (which we expanded substantially in this edition), we decided to also follow this trend and use Data Science as the overarching umbrella term. This also made it obvious that having someone else with more real world data science experience would be beneficial so we were glad when Rosaria agreed to share the (re)writing duty and add decades of industrial data science experience to the team.

And finally we decided to put the focus of the practical exercises more on KNIME Analytics Platform—not because there is only one tool out there but because visual workflows are easier to explain and therefore lend themselves to be included in printed material to illustrate how the methods are being used in practice. On the book's website: www.datascienceguide.org you can still find the R examples from the first edition and we have also added examples in Python as well as all of the KNIME workflows described in the book. We are also providing teaching material and will continuously update the site in the coming years.

There are many people to be thanked, and we will not attempt to list them all. However, Iris Adä and Martin Horn deserve mentioning for all their help with the first edition. For the second round, we owe thanks to Satoru Hayasaka, Kathrin Melcher, and Emilio Silvestri who have spent many hours proof reading and updating/creating workflows.

Konstanz, Germany — Michael R. Berthold
Salzburg, Austria — Christian Borgelt
Braunschweig, Germany — Frank Höppner and Frank Klawonn
Zurich, Switzerland — Rosaria Silipo

Contents

Symbols

A, A_i	attribute, variable [e.g., $A_1 = color$, $A_2 = price$, $A_3 = category$]
ω	a possible value of an attribute [e.g., $\omega = red$]
Ω, dom$(\cdot)$	set of possible values of an attribute [e.g., $\Omega_1 = \Omega_{\text{color}} = \text{dom}(A_i) = \{red, blue, green\}$]
$\mathcal{A}$	set of all attributes [e.g., $\mathcal{A} = \{color, price, category\}$]
m	number of considered attributes [e.g., 3]
x	a specific value of an attribute [e.g., $x_2 = x_{\text{price}} = 4000$]
$\mathcal{X}$	space of possible data records [e.g., $\mathcal{X} = \Omega_{A_1} \times \cdots \times \Omega_{A_m}$]
$\mathcal{D}$	set of all records, data set, $\mathcal{D} \subseteq \mathcal{X}$ [e.g., $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$]
n	number of records in data set
$\mathbf{x}$	record in database [e.g., $\mathbf{x} = (x_1, x_2, x_3) = (red, 4000, luxury)$]
$\mathbf{x}_A$	attribute A of record $\mathbf{x}$ [e.g., $\mathbf{x}_{\text{price}} = 4000$]
$\mathbf{x}_{2,A}$	attribute A of record $\mathbf{x}_2$
$\mathcal{D}_{A=v}$	set of all records $\mathbf{x} \in \mathcal{D}$ with $\mathbf{x}_A = v$
C	a selected categorical target attribute [e.g., $C = A_3 = category$]
Ω_C	set of all possible classes [e.g., $\Omega_C = \{\text{quits, stays, unknown}\}$]
Y	a selected continuous target attribute [e.g., $Y = A_2 = price$]
$\mathcal{C}$	cluster (set of associated data objects) [e.g., $\mathcal{C} \subseteq \mathcal{D}$]
c	number of clusters
$\mathcal{P}$	partition, set of clusters $\{\mathcal{C}_1, \ldots, \mathcal{C}_c\}$
$p_{i\|j}$	membership degree of data #j to cluster #i
$[p_{i\|j}]$	membership matrix
$d_\cdot$	distance function, metric (d_E: Euclidean)
$[d_{i,j}]$	distance matrix