Undergraduate Topics in Computer Science

Boris Mirkin

Core Concepts in Data Analysis: Summarization, Correlation and Visualization

UTiCS

Springer

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.

For further volumes:
http://www.springer.com/series/7592

Boris Mirkin

Core Concepts in Data Analysis: Summarization, Correlation and Visualization

Springer

Boris Mirkin
Research University – Higher School
of Economics
School of Applied Mathematics
and Informatics
11 Pokrovsky Boulevard
Moscow RF
Russia
and
Department of Computer Science
Birkbeck University of London
Malet Street, London
UK

ISSN 1863-7310
ISBN 978-0-85729-286-5 e-ISBN 978-0-85729-287-2
DOI 10.1007/978-0-85729-287-2
Springer London Dordrecht Heidelberg New York

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2011922052

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

In this textbook, I take an unconventional approach to data analysis. Its contents are heavily influenced by the idea that data analysis should help in enhancing and augmenting knowledge of the domain as represented by the concepts and statements of relation between them. According to this view, two main pathways for data analysis are summarization, for developing and augmenting concepts, and correlation, for enhancing and establishing relations. Visualization, in this context, is a way of presenting results in a cognitively comfortable way. The term *summarization* is understood quite broadly here to embrace not only simple summaries like totals and means, but also more complex summaries such as the principal components of a set of features or cluster structures in a set of entities.

The material presented in this perspective makes a unique mix of subjects from the fields of statistical data analysis, data mining, and computational intelligence, which follow different systems of presentation.

Another feature of the text is that its main thrust is to give an in-depth understanding of a few basic techniques rather than to cover a broad spectrum of approaches developed so far. Most of the described methods fall under the same least-squares paradigm for mapping an "idealized" structure to the data. This allows me to bring forward a number of relations between methods that are usually overlooked. Just one example: a relation between the choice of a scoring function for classification trees and normalization options for dummies representing the target categories.

Although the in-depth study approach involves a great deal of technical details, these are encapsulated in specific fragments of the text termed "formulation" parts. The main, "presentation", part is written in a very different style. The presentation involves no mathematical formulas and explains a method by actually applying it to a small real-world dataset – this part can be read and studied with no concern for the formulation at all. There is one more part, "computation", targeted at a computer-oriented reader. This part describes the computational implementation of the methods, illustrated using the MatLab computing environment. I have arrived at this three-way narrative style as a result of my experiences in teaching data analysis and computational intelligence to students in Computer Science. Some students might be mainly interested in just one of the parts, whereas others might try to get to grips with two or even all three of them.

One more device to stimulate the reader's interest is a multi-layer system of pro-active learning materials for class- and self-study:

- *Worked examples* provided to show how specific methods apply to particular datasets;
- More complex problems solved, *case studies*, possibly involving a rule for data generation, rather than a pre-specified dataset, or an informal way of analyzing results;
- Even more complex problems, *projects*, possibly involving uncharted terrain and a small-scale investigation; these should be used as models for similar self-study projects on other data;
- A number of computational or theoretical problems, *questions*, formulated as self-study exercises; answers are provided for most of them.

The text is based on my courses for full-time and part-time students in the MS program in Computer Science at Birkbeck, University of London (2003–2010), in the BS and MS programs in Applied Mathematics and Informatics at Higher School of Economics, Moscow (2008–2010), and post-graduate School of Data Analysis at Yandex, a popular Russian search engine, Moscow (2009–2010). The material covers lectures and labs for about 35–40 lecture hours in advanced BS programs or MS programs in Computer Science or Engineering. It can also be used in application-oriented courses such as Bioinformatics or Methods in Marketing Research.

No prerequisite beyond a conventional school background for reading through the presentation part is required, yet some training in reading academic material is expected. The reader interested in studying the formulation part should have some background in: (a) basic calculus including the concepts of function, derivative and the first-order optimality conditions, (b) basic linear algebra including vectors, inner products, Euclidean distances and matrices (these are reviewed in the Appendix), and (c) basic set theory notation such as the symbols for inclusion and membership. The computation part is oriented towards those interested in coding for computer implementation, specifically focusing on working with MatLab as a user-friendly environment.

Acknowledgments

Too many people contributed to the material of the book to list all their names. First of all, my gratitude goes to Springer's editors who were instrumental in bringing forth the very idea of writing such a book and in channeling my efforts by providing good critical reviews. Then, of course, I thank the students at my classes in MS programs in Computer Science at Birkbeck and, more recently, in BS and MS programs in Applied Mathematics and Informatics at HSE. Here is a list of people who directly contributed to this book with advice, and sometimes with computation: I. Muchnik (Rutgers University), M. Levin (Higher School of Economics Moscow), T. Fenner (Birkbeck University of London), S. Nascimento (New University of Lisbon), T. Krauze (Hofstra University), I. Mandel (Telmar Inc), I. Mirkin (Yext), V. Sulimova (Tula Technical University), and V. Topinsky (Higher School of Economics Moscow). The HSE students J. Askarova, K. Chernyak, O. Chugunova, K. Kovaleva, and A Kramarenko helped in debugging the final version.

Contents