

Undergraduate Topics in Computer Science

Boris Mirkin

Core Concepts in Data Analysis: Summarization, Correlation and Visualization



 Springer

The Springer logo consists of a stylized chess knight piece facing left, positioned above the word "Springer" in a serif font.

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

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

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

© Springer-Verlag London Limited 2011

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licenses issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

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

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

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 proactive 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

1 Introduction: What Is Core	1
1.1 Summarization and Correlation: Two Main Goals of Data Analysis	1
1.2 Case Study Problems	9
1.3 An Account of Data Visualization	21
1.3.1 General	21
1.3.2 Highlighting	22
1.3.3 Integrating Different Aspects	25
1.3.4 Narrating a Story	28
1.4 Summary	28
References	29
2 1D Analysis: Summarization and Visualization of a Single Feature	31
2.1 Quantitative Feature: Distribution and Histogram	31
P2.1.1 Presentation	31
F2.1.2 Formulation	33
C2.1.3 Computation	35
2.2 Further Summarization: Centers and Spreads	36
P2.2.1 Centers and Spreads: Presentation	36
F2.2.2 Centers and Spreads: Formulation	39
C2.2.3 Centers and Spreads: Computation	43
2.3 Binary and Categorical Features	43
P2.3.1 Presentation	43
F2.3.2 Formulation	46
C2.3.3 Computation	49
2.4 Modeling Uncertainty: Intervals and Fuzzy Sets	49
2.4.1 Individual Membership Functions	49
2.4.2 Central Fuzzy Set	52
2.5 Summary	64
References	65

3	2D Analysis: Correlation and Visualization of Two Features	67
3.1	General	67
3.2	Two Quantitative Features Case	68
	P3.2.1 Scatter-Plot, Linear Regression and Correlation Coefficients	68
	P3.2.2 Validity of the Regression	70
	F3.2.3 Linear Regression: Formulation	74
	C3.2.4 Linear Regression: Computation	78
3.3	Mixed Scale Case: Nominal Feature Versus a Quantitative One	89
	P3.3.1 Box-Plot, Tabular Regression and Correlation Ratio	89
	F3.3.2 Tabular Regression: Formulation	93
	3.3.3 Nominal Target	95
3.4	Two Nominal Features Case	100
	P3.4.1 Analysis of Contingency Tables: Presentation	100
	F3.4.2 Analysis of Contingency Tables: Formulation	107
3.5	Summary	111
	References	112
4	Learning Multivariate Correlations in Data	113
4.1	General: Decision Rules, Fitting Criteria, and Learning Protocols	113
4.2	Naïve Bayes Approach	118
	4.2.1 Bayes Decision Rule	118
	4.2.2 Naïve Bayes Classifier	120
	4.2.3 Metrics of Accuracy	123
4.3	Linear Regression	128
	P4.3.1 Linear Regression: Presentation	128
	F4.3.2 Linear Regression: Formulation	131
4.4	Linear Discrimination and SVM	133
	P4.4.1 Linear Discrimination and SVM: Presentation	133
	F4.4.2 Linear Discrimination and SVM: Formulation	137
4.5	Decision Trees	141
	P4.5.1 General: Presentation	141
	F4.5.2 General: Formulation	142
	4.5.3 Measuring Correlation for Classification Trees	145
	4.5.4 Building Classification Trees	152
	C4.5.5 Building Classification Trees: Computation	157
4.6	Learning Correlation with Neural Networks	159
	4.6.1 General	159
	4.6.2 Learning a Multi-layer Network	163
4.7	Summary	171
	References	171
5	Principal Component Analysis and SVD	173
5.1	Decoder Based Data Summarization	173
	5.1.1 Structure of a Summarization Problem with Decoder	173