Intelligent Data Analysis

Michael Berthold · David J. Hand (Eds.)

# Intelligent Data Analysis

## An Introduction

2nd revised and extended Edition

With 140 Figures, 22 in color and 50 Tables

Springer

*Editors*

Michael Berthold
Universität Konstanz
FB Informatik und Informationswissenschaft
78457 Konstanz
Germany
Michael.Berthold@uni-konstanz.de

David J. Hand
Department of Mathematics
Imperial College
Huxley Building
180 Queen's Gate
London, SW7 2BZ
UK
d.j.hand@ic.ac.uk

Library of Congress Control Number: 2003041211

ACM Computing Classification (1998): I.2, H.3, G.3, I.5.1, I.4, J.2, J.1, J.3, F.4.1, F.1

corrected second printing 2007

ISBN-10 3-540-43060-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-43060-5 Springer Berlin Heidelberg New York
ISBN-10 3-540-65808-4 1. Edition Springer Berlin Heidelberg New York



Springer is a part of Springer Science+Business Media

springer.com



Typesetting: by the Editors
Production: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig
Cover: KünkelLopka, Heidelberg

Printed on acid-free paper 45/3100YL - 5 4 3 2 1 0

# Preface to the Second Edition

We were pleasantly surprised by the success of the first edition of this book. Many of our colleagues have started to use it for teaching purposes, and feedback from industrial researchers has also shown that it is useful for practitioners. So, when Springer-Verlag approached us and asked us to revise the material for a second edition, we gladly took the opportunity to rearrange some of the existing material and to invite new authors to write two new chapters. These additional chapters cover material that has attracted considerable attention since the first edition of the book appeared. They deal with kernel methods and support vector machines on the one hand, and visualization on the other. Kernel methods represent a relatively new technology, but one which is showing great promise. Visualization methods have been around in some form or other ever since data analysis began, but are currently experiencing a renaissance in response to the increase, in numbers and size, of large data sets. In addition the chapter on rule induction has been replaced with a new version, covering this topic in much more detail.

As research continues, and new tools and methods for data analysis continue to be developed, so it becomes ever more difficult to cover all of the important techniques. Indeed, we are probably further from this goal than we were with the original edition – too many new fields have emerged over the past three years. However, we believe that this revision still provides a solid basis for anyone interested in the analysis of real data.

We are very grateful to the authors of the new chapters for working with us to an extremely tight schedule. We also would like to thank the authors of the existing chapters for spending so much time carefully revising and updating their chapters. And, again, all this would not have been possible without the help of many people, including Olfa Nasraoui, Ashley Morris, and Jim Farrand.

Once again, we owe especial thanks to Alfred Hofmann and Ingeborg Mayer of Springer-Verlag, for their continued support for this book and their patience with various delays during the preparation of this second edition.

November 2002

*South San Francisco, CA, USA* Michael R. Berthold
*London, UK* David J. Hand

# Preface to the First Edition

The obvious question, when confronted with a book with the title of this one, is why "intelligent" data analysis? The answer is that modern data analysis uses tools developed by a wide variety of intellectual communities and that "intelligent data analysis", or IDA, has been adopted as an overall term. It should be taken to imply the intelligent application of data analytic tools, and also the application of "intelligent" data analytic tools, computer programs which probe more deeply into structure than first generation methods. These aspects reflect the distinct influences of statistics and machine learning on the subject matter.

The importance of intelligent data analysis arises from the fact that the modern world is a data-driven world. We are surrounded by data, numerical and otherwise, which must be analysed and processed to convert it into *information* which informs, instructs, answers, or otherwise aids understanding and decision making. The quantity of such data is huge and growing, the number of sources is effectively unlimited, and the range of areas covered is vast: industrial, commercial, financial, and scientific activities are all generating such data.

The origin of this book was a wish to have a single introductory source to which we could direct students, rather than having to direct them to multiple sources. However, it soon became apparent that wider interest existed, and that potential readers other than our students would appreciate a compilation of some of the most important tools of intelligent data analysis. Such readers include people from a wide variety of backgrounds and positions who find themselves confronted by the need to make sense of data.

Given the wide range of topics we hoped to cover, we rapidly abandoned the idea of writing the entire volume ourselves, and instead decided to invite appropriate experts to contribute separate chapters. We did, however, make considerable efforts to ensure that these chapters complemented and built on each other, so that a rounded picture resulted. We are especially grateful to the authors for their patience in putting up with repeated requests for revision so as to make the chapters meld better.

In a volume such as this there are many people whose names do not explicitly appear as contributors, but without whom the work would be of substantially reduced quality. These people include Jay Diamond, Matt Easley, Sibylle Frank, Steven Greenberg, Thomas Hofmann, Joy Hollenback, Joe Iwanski, Carlo Marchesi, Roger Mitton, Vanessa Robins, Nancy Shaw, and Camille Sinanan for their painstaking proofreading and other help, as well as Stefan Wrobel, Chris Road-

knight and Dominic Palmer-Brown for stimulating discussions and contributions which, though not appearing in print, have led to critical reassessment of how we thought some of the material should be presented.

Finally, we owe especial thanks to Alfred Hofmann from Springer-Verlag, for his enthusiasm and support for this book right from the start.

February 1999

*Berkeley, California* Michael Berthold
*London, United Kingdom* David J. Hand

# Table of Contents