

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

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable for prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 1999, 2003 and 2007

The use of general descriptive names, 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 protective laws and regulations and therefore free for general use.

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
London, UK

Michael R. Berthold
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
London, United Kingdom

Michael Berthold
David J. Hand

Table of Contents

1. Introduction	1
1.1 Why “Intelligent Data Analysis”?	1
1.2 How the Computer Is Changing Things	4
1.3 The Nature of Data	8
1.4 Modern Data Analytic Tools	12
1.5 Conclusion	14
2. Statistical Concepts	17
2.1 Introduction	17
2.2 Probability	18
2.3 Sampling and Sampling Distributions	29
2.4 Statistical Inference	33
2.5 Prediction and Prediction Error	46
2.6 Resampling	57
2.7 Conclusion	68
3. Statistical Methods	69
3.1 Introduction	69
3.2 Generalized Linear Models	70
3.3 Special Topics in Regression Modelling	93
3.4 Classical Multivariate Analysis	100
3.5 Conclusion	129
4. Bayesian Methods	131
4.1 Introduction	131
4.2 The Bayesian Paradigm	132
4.3 Bayesian Inference	135
4.4 Bayesian Modeling	143
4.5 Bayesian Networks	153
4.6 Conclusion	167
5. Support Vector and Kernel Methods	169
5.1 Example: Kernel Perceptron	170
5.2 Overfitting and Generalization Bounds	176
5.3 Support Vector Machines	181

5.4 Kernel PCA and CCA	194
5.5 Conclusion	196
6. Analysis of Time Series	199
6.1 Introduction	199
6.2 Linear Systems Analysis	202
6.3 Nonlinear Dynamics Basics	207
6.4 Delay-Coordinate Embedding	213
6.5 Examples	218
6.6 Conclusion	226
7. Rule Induction	229
7.1 Introduction	229
7.2 Propositional rule learning	232
7.3 Rule learning as search	236
7.4 Evaluating the quality of rules	242
7.5 Propositional rule induction at work	246
7.6 Learning first-order rules	250
7.7 Some ILP systems at work	262
7.8 Conclusion	267
8. Neural Networks	269
8.1 Introduction	269
8.2 Fundamentals	270
8.3 Multilayer Feedforward Neural Networks	278
8.4 Learning and Generalization	283
8.5 Radial Basis Function Networks	292
8.6 Competitive Learning	300
8.7 Principal Components Analysis and Neural Networks	307
8.8 Time Series Analysis	312
8.9 Conclusion	319
9. Fuzzy Logic	321
9.1 Introduction	321
9.2 Basics of Fuzzy Sets and Fuzzy Logic	322
9.3 Extracting Fuzzy Models from Data	336
9.4 Fuzzy Decision Trees	346
9.5 Conclusion	350
10. Stochastic Search Methods	351
10.1 Introduction	351
10.2 Stochastic Search by Simulated Annealing	354
10.3 Stochastic, Adaptive Search by Evolution	360
10.4 Evolution Strategies	362
10.5 Genetic Algorithms	374

10.6 Genetic Programming	390
10.7 Conclusion	400
11. Visualization	403
11.1 Introduction	403
11.2 Classification of Visual Data Analysis Techniques	405
11.3 Data Type to be Visualized	406
11.4 Visualization Techniques	411
11.5 Interaction Techniques	414
11.6 Specific Visual Data Analysis Techniques	418
11.7 Conclusion	426
12. Systems and Applications	429
12.1 Introduction	429
12.2 Diversity of IDA Applications	430
12.3 Several Development Issues	436
12.4 Conclusion	442
Appendix A: Tools	445
A.1 Tools for statistical analysis	445
A.2 Tools for exploration/modeling	447
A.3 Tools for Text and Web Mining	454
A.4 Data Analysis Suites	456
A.5 Conclusion	464
Appendix B: Information-Theoretic Tree and Rule Induction	465
B.1 Information and Uncertainty	465
B.2 Decision Tree Induction	468
B.3 Rule Induction	470
References	475
Index	501
Author Addresses	513