Springer Series in Statistics

Springer Series in Statistics

For other titles published in this series, go to
http://www.springer.com/series/692

Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

Springer

Trevor Hastie
Stanford University
Dept. of Statistics
Stanford CA 94305
USA
hastie@stanford.edu

Robert Tibshirani
Stanford University
Dept. of Statistics
Stanford CA 94305
USA
tibs@stanford.edu

Jerome Friedman
Stanford University
Dept. of Statistics
Stanford CA 94305
USA
jhf@stanford.edu

ISSN: 0172-7397
ISBN: 978-0-387-84857-0 e-ISBN: 978-0-387-84858-7
DOI: 10.1007/b94608

Library of Congress Control Number: 2008941148

Printed on acid-free paper

springer.com

To our parents:

Valerie and Patrick Hastie

Vera and Sami Tibshirani

Florence and Harry Friedman

and to our families:

Samantha, Timothy, and Lynda

Charlie, Ryan, Julie, and Cheryl

Melanie, Dora, Monika, and Ildiko

Preface to the Second Edition

In God we trust, all others bring data.

–William Edwards Deming (1900-1993)[1]

We have been gratified by the popularity of the first edition of *The Elements of Statistical Learning.* This, along with the fast pace of research in the statistical learning field, motivated us to update our book with a second edition.

We have added four new chapters and updated some of the existing chapters. Because many readers are familiar with the layout of the first edition, we have tried to change it as little as possible. Here is a summary of the main changes:

[1]On the Web, this quote has been widely attributed to both Deming and Robert W. Hayden; however Professor Hayden told us that he can claim no credit for this quote, and ironically we could find no "data" confirming that Deming actually said this.

Chapter	What's new
1. Introduction	
2. Overview of Supervised Learning	
3. Linear Methods for Regression	LAR algorithm and generalizations of the lasso
4. Linear Methods for Classification	Lasso path for logistic regression
5. Basis Expansions and Regularization	Additional illustrations of RKHS
6. Kernel Smoothing Methods	
7. Model Assessment and Selection	Strengths and pitfalls of cross-validation
8. Model Inference and Averaging	
9. Additive Models, Trees, and Related Methods	
10. Boosting and Additive Trees	New example from ecology; some material split off to Chapter 16.
11. Neural Networks	Bayesian neural nets and the NIPS 2003 challenge
12. Support Vector Machines and Flexible Discriminants	Path algorithm for SVM classifier
13. Prototype Methods and Nearest-Neighbors	
14. Unsupervised Learning	Spectral clustering, kernel PCA, sparse PCA, non-negative matrix factorization archetypal analysis, nonlinear dimension reduction, Google page rank algorithm, a direct approach to ICA
15. Random Forests	New
16. Ensemble Learning	New
17. Undirected Graphical Models	New
18. High-Dimensional Problems	New

Some further notes:

- Our first edition was unfriendly to colorblind readers; in particular, we tended to favor red/green contrasts which are particularly troublesome. We have changed the color palette in this edition to a large extent, replacing the above with an orange/blue contrast.

- We have changed the name of Chapter 6 from "Kernel Methods" to "Kernel Smoothing Methods", to avoid confusion with the machine-learning kernel method that is discussed in the context of support vector machines (Chapter 12) and more generally in Chapters 5 and 14.

- In the first edition, the discussion of error-rate estimation in Chapter 7 was sloppy, as we did not clearly differentiate the notions of conditional error rates (conditional on the training set) and unconditional rates. We have fixed this in the new edition.

- Chapters 15 and 16 follow naturally from Chapter 10, and the chapters are probably best read in that order.

- In Chapter 17, we have not attempted a comprehensive treatment of graphical models, and discuss only undirected models and some new methods for their estimation. Due to a lack of space, we have specifically omitted coverage of directed graphical models.

- Chapter 18 explores the "$p \gg N$" problem, which is learning in high-dimensional feature spaces. These problems arise in many areas, including genomic and proteomic studies, and document classification.

We thank the many readers who have found the (too numerous) errors in the first edition. We apologize for those and have done our best to avoid errors in this new edition. We thank Mark Segal, Bala Rajaratnam, and Larry Wasserman for comments on some of the new chapters, and many Stanford graduate and post-doctoral students who offered comments, in particular Mohammed AlQuraishi, John Boik, Holger Hoefling, Arian Maleki, Donal McMahon, Saharon Rosset, Babak Shababa, Daniela Witten, Ji Zhu and Hui Zou. We thank John Kimmel for his patience in guiding us through this new edition. RT dedicates this edition to the memory of Anna McPhee.

Trevor Hastie
Robert Tibshirani
Jerome Friedman

Stanford, California
August 2008

Preface to the First Edition

We are drowning in information and starving for knowledge.

–Rutherford D. Roger

The field of Statistics is constantly challenged by the problems that science and industry brings to its door. In the early days, these problems often came from agricultural and industrial experiments and were relatively small in scope. With the advent of computers and the information age, statistical problems have exploded both in size and complexity. Challenges in the areas of data storage, organization and searching have led to the new field of "data mining"; statistical and computational problems in biology and medicine have created "bioinformatics." Vast amounts of data are being generated in many fields, and the statistician's job is to make sense of it all: to extract important patterns and trends, and understand "what the data says." We call this *learning from data*.

The challenges in learning from data have led to a revolution in the statistical sciences. Since computation plays such a key role, it is not surprising that much of this new development has been done by researchers in other fields such as computer science and engineering.

The learning problems that we consider can be roughly categorized as either *supervised* or *unsupervised*. In supervised learning, the goal is to predict the value of an outcome measure based on a number of input measures; in unsupervised learning, there is no outcome measure, and the goal is to describe the associations and patterns among a set of input measures.

This book is our attempt to bring together many of the important new ideas in learning, and explain them in a statistical framework. While some mathematical details are needed, we emphasize the methods and their conceptual underpinnings rather than their theoretical properties. As a result, we hope that this book will appeal not just to statisticians but also to researchers and practitioners in a wide variety of fields.

Just as we have learned a great deal from researchers outside of the field of statistics, our statistical viewpoint may help others to better understand different aspects of learning:

> *There is no true interpretation of anything; interpretation is a vehicle in the service of human comprehension. The value of interpretation is in enabling others to fruitfully think about an idea.*

–Andreas Buja

We would like to acknowledge the contribution of many people to the conception and completion of this book. David Andrews, Leo Breiman, Andreas Buja, John Chambers, Bradley Efron, Geoffrey Hinton, Werner Stuetzle, and John Tukey have greatly influenced our careers. Balasubramanian Narasimhan gave us advice and help on many computational problems, and maintained an excellent computing environment. Shin-Ho Bang helped in the production of a number of the figures. Lee Wilkinson gave valuable tips on color production. Ilana Belitskaya, Eva Cantoni, Maya Gupta, Michael Jordan, Shanti Gopatam, Radford Neal, Jorge Picazo, Bogdan Popescu, Olivier Renaud, Saharon Rosset, John Storey, Ji Zhu, Mu Zhu, two reviewers and many students read parts of the manuscript and offered helpful suggestions. John Kimmel was supportive, patient and helpful at every phase; MaryAnn Brickner and Frank Ganz headed a superb production team at Springer. Trevor Hastie would like to thank the statistics department at the University of Cape Town for their hospitality during the final stages of this book. We gratefully acknowledge NSF and NIH for their support of this work. Finally, we would like to thank our families and our parents for their love and support.

Trevor Hastie
Robert Tibshirani
Jerome Friedman

Stanford, California
May 2001

> *The quiet statisticians have changed our world; not by discovering new facts or technical developments, but by changing the ways that we reason, experiment and form our opinions ....*

–Ian Hacking

Contents