

Springer Series in Statistics

Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

统计学习基础 第2版

Springer

世界图书出版公司
www.wpcbj.com.cn

Trevor Hastie

Robert Tibshirani

Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

 Springer

图书在版编目 (C I P) 数据

统计学习基础 = The elements of statistical learning : 第2版 : 英文 / (德) 黑斯蒂 (Hastie, T.) 著. — 影印本. — 北京 : 世界图书出版公司北京公司, 2014. 8

ISBN 978-7-5100-8450-8

I. ①统… II. ①黑… III. ①统计学—英文 IV. CS

中国版本图书馆CIP数据核字 (2014) 第185718号

书 名:	The Elements of Statistical Learning: Data Mining, Inference, and Prediction 2nd ed.
作 者:	Trevor Hastie, Robert Tibshirani, Jerome Friedman
中 译 名:	统计学习基础 第2版
责任编辑:	高蓉 刘慧
出 版 者:	世界图书出版公司北京公司
印 刷 者:	北京博图彩色印刷有限公司
发 行 者:	世界图书出版公司北京公司 (北京朝内大街137号 100010)
联系电话:	010-64021602, 010-64015659
电子信箱:	kjb@wpcbj.com.cn

开 本:	24开
印 张:	32
版 次:	2015年1月
版权登记:	图字: 01-2013-5101
书 号:	978-7-5100-8450-8
定 价:	119.00元

Springer Series in Statistics

Advisors:

P. Bickel, P. Diggle, S. Fienberg, U. Gather,
I. Olkin, S. Zeger

1. Title
2. Author
3. Editor
4. Publisher

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

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

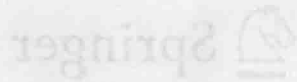
Reprint from English language edition:

The Elements of Statistical Learning: Data Mining, Inference, and Prediction Second Edition

by Trevor Hastie, Robert Tibshirani, Jerome Friedman

Copyright © 2009, Springer Science+Business Media

This reprint has been authorized by Springer Science & Business Media for distribution in China Mainland only and not for export therefrom.



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:

¹On 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 11) 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 Hoeffling, 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

Preface to the Second Edition	vii
Preface to the First Edition	xi
1 Introduction	1
2 Overview of Supervised Learning	9
2.1 Introduction	9
2.2 Variable Types and Terminology	9
2.3 Two Simple Approaches to Prediction:	
Least Squares and Nearest Neighbors	11
2.3.1 Linear Models and Least Squares	11
2.3.2 Nearest-Neighbor Methods	14
2.3.3 From Least Squares to Nearest Neighbors	16
2.4 Statistical Decision Theory	18
2.5 Local Methods in High Dimensions	22
2.6 Statistical Models, Supervised Learning	
and Function Approximation	28
2.6.1 A Statistical Model	
for the Joint Distribution $\Pr(X, Y)$	28
2.6.2 Supervised Learning	29
2.6.3 Function Approximation	29
2.7 Structured Regression Models	32
2.7.1 Difficulty of the Problem	32

2.8	Classes of Restricted Estimators	33
2.8.1	Roughness Penalty and Bayesian Methods	34
2.8.2	Kernel Methods and Local Regression	34
2.8.3	Basis Functions and Dictionary Methods	35
2.9	Model Selection and the Bias–Variance Tradeoff	37
	Bibliographic Notes	39
	Exercises	39
3	Linear Methods for Regression	43
3.1	Introduction	43
3.2	Linear Regression Models and Least Squares	44
3.2.1	Example: Prostate Cancer	49
3.2.2	The Gauss–Markov Theorem	51
3.2.3	Multiple Regression from Simple Univariate Regression	52
3.2.4	Multiple Outputs	56
3.3	Subset Selection	57
3.3.1	Best-Subset Selection	57
3.3.2	Forward- and Backward-Stepwise Selection	58
3.3.3	Forward-Stagewise Regression	60
3.3.4	Prostate Cancer Data Example (Continued)	61
3.4	Shrinkage Methods	61
3.4.1	Ridge Regression	61
3.4.2	The Lasso	68
3.4.3	Discussion: Subset Selection, Ridge Regression and the Lasso	69
3.4.4	Least Angle Regression	73
3.5	Methods Using Derived Input Directions	79
3.5.1	Principal Components Regression	79
3.5.2	Partial Least Squares	80
3.6	Discussion: A Comparison of the Selection and Shrinkage Methods	82
3.7	Multiple Outcome Shrinkage and Selection	84
3.8	More on the Lasso and Related Path Algorithms	86
3.8.1	Incremental Forward Stagewise Regression	86
3.8.2	Piecewise-Linear Path Algorithms	89
3.8.3	The Dantzig Selector	89
3.8.4	The Grouped Lasso	90
3.8.5	Further Properties of the Lasso	91
3.8.6	Pathwise Coordinate Optimization	92
3.9	Computational Considerations	93
	Bibliographic Notes	94
	Exercises	94

4	Linear Methods for Classification	101
4.1	Introduction	101
4.2	Linear Regression of an Indicator Matrix	103
4.3	Linear Discriminant Analysis	106
	4.3.1 Regularized Discriminant Analysis	112
	4.3.2 Computations for LDA	113
	4.3.3 Reduced-Rank Linear Discriminant Analysis	113
4.4	Logistic Regression	119
	4.4.1 Fitting Logistic Regression Models	120
	4.4.2 Example: South African Heart Disease	122
	4.4.3 Quadratic Approximations and Inference	124
	4.4.4 L_1 Regularized Logistic Regression	125
	4.4.5 Logistic Regression or LDA?	127
4.5	Separating Hyperplanes	129
	4.5.1 Rosenblatt's Perceptron Learning Algorithm	130
	4.5.2 Optimal Separating Hyperplanes	132
	Bibliographic Notes	135
	Exercises	135
5	Basis Expansions and Regularization	139
5.1	Introduction	139
5.2	Piecewise Polynomials and Splines	141
	5.2.1 Natural Cubic Splines	144
	5.2.2 Example: South African Heart Disease (Continued)	146
	5.2.3 Example: Phoneme Recognition	148
5.3	Filtering and Feature Extraction	150
5.4	Smoothing Splines	151
	5.4.1 Degrees of Freedom and Smoother Matrices	153
5.5	Automatic Selection of the Smoothing Parameters	156
	5.5.1 Fixing the Degrees of Freedom	158
	5.5.2 The Bias-Variance Tradeoff	158
5.6	Nonparametric Logistic Regression	161
5.7	Multidimensional Splines	162
5.8	Regularization and Reproducing Kernel Hilbert Spaces	167
	5.8.1 Spaces of Functions Generated by Kernels	168
	5.8.2 Examples of RKHS	170
5.9	Wavelet Smoothing	174
	5.9.1 Wavelet Bases and the Wavelet Transform	176
	5.9.2 Adaptive Wavelet Filtering	179
	Bibliographic Notes	181
	Exercises	181
	Appendix: Computational Considerations for Splines	186
	Appendix: B -splines	186
	Appendix: Computations for Smoothing Splines	189

6	Kernel Smoothing Methods	191
6.1	One-Dimensional Kernel Smoothers	192
6.1.1	Local Linear Regression	194
6.1.2	Local Polynomial Regression	197
6.2	Selecting the Width of the Kernel	198
6.3	Local Regression in \mathbb{R}^p	200
6.4	Structured Local Regression Models in \mathbb{R}^p	201
6.4.1	Structured Kernels	203
6.4.2	Structured Regression Functions	203
6.5	Local Likelihood and Other Models	205
6.6	Kernel Density Estimation and Classification	208
6.6.1	Kernel Density Estimation	208
6.6.2	Kernel Density Classification	210
6.6.3	The Naive Bayes Classifier	210
6.7	Radial Basis Functions and Kernels	212
6.8	Mixture Models for Density Estimation and Classification	214
6.9	Computational Considerations	216
	Bibliographic Notes	216
	Exercises	216
7	Model Assessment and Selection	219
7.1	Introduction	219
7.2	Bias, Variance and Model Complexity	219
7.3	The Bias–Variance Decomposition	223
7.3.1	Example: Bias–Variance Tradeoff	226
7.4	Optimism of the Training Error Rate	228
7.5	Estimates of In-Sample Prediction Error	230
7.6	The Effective Number of Parameters	232
7.7	The Bayesian Approach and BIC	233
7.8	Minimum Description Length	235
7.9	Vapnik–Chervonenkis Dimension	237
7.9.1	Example (Continued)	239
7.10	Cross-Validation	241
7.10.1	K -Fold Cross-Validation	241
7.10.2	The Wrong and Right Way to Do Cross-validation	245
7.10.3	Does Cross-Validation Really Work?	247
7.11	Bootstrap Methods	249
7.11.1	Example (Continued)	252
7.12	Conditional or Expected Test Error?	254
	Bibliographic Notes	257
	Exercises	257
8	Model Inference and Averaging	261
8.1	Introduction	261

8.2	The Bootstrap and Maximum Likelihood Methods	261
8.2.1	A Smoothing Example	261
8.2.2	Maximum Likelihood Inference	265
8.2.3	Bootstrap versus Maximum Likelihood	267
8.3	Bayesian Methods	267
8.4	Relationship Between the Bootstrap and Bayesian Inference	271
8.5	The EM Algorithm	272
8.5.1	Two-Component Mixture Model	272
8.5.2	The EM Algorithm in General	276
8.5.3	EM as a Maximization–Maximization Procedure	277
8.6	MCMC for Sampling from the Posterior	279
8.7	Bagging	282
8.7.1	Example: Trees with Simulated Data	283
8.8	Model Averaging and Stacking	288
8.9	Stochastic Search: Bumping	290
	Bibliographic Notes	292
	Exercises	293
9	Additive Models, Trees, and Related Methods	295
9.1	Generalized Additive Models	295
9.1.1	Fitting Additive Models	297
9.1.2	Example: Additive Logistic Regression	299
9.1.3	Summary	304
9.2	Tree-Based Methods	305
9.2.1	Background	305
9.2.2	Regression Trees	307
9.2.3	Classification Trees	308
9.2.4	Other Issues	310
9.2.5	Spam Example (Continued)	313
9.3	PRIM: Bump Hunting	317
9.3.1	Spam Example (Continued)	320
9.4	MARS: Multivariate Adaptive Regression Splines	321
9.4.1	Spam Example (Continued)	326
9.4.2	Example (Simulated Data)	327
9.4.3	Other Issues	328
9.5	Hierarchical Mixtures of Experts	329
9.6	Missing Data	332
9.7	Computational Considerations	334
	Bibliographic Notes	334
	Exercises	335
10	Boosting and Additive Trees	337
10.1	Boosting Methods	337
10.1.1	Outline of This Chapter	340