

Gareth James · Daniela Witten ·
Trevor Hastie · Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Second Edition

First Printing: August 4, 2021

To our parents:

Alison and Michael James

Chiara Nappi and Edward Witten

Valerie and Patrick Hastie

Vera and Sami Tibshirani

and to our families:

Michael, Daniel, and Catherine

Tessa, Theo, Otto, and Ari

Samantha, Timothy, and Lynda

Charlie, Ryan, Julie, and Cheryl

Preface

Statistical learning refers to a set of tools for *making sense of complex datasets*. In recent years, we have seen a staggering increase in the scale and scope of data collection across virtually all areas of science and industry. As a result, statistical learning has become a critical toolkit for anyone who wishes to understand data — and as more and more of today’s jobs involve data, this means that statistical learning is fast becoming a critical toolkit for *everyone*.

One of the first books on statistical learning — *The Elements of Statistical Learning* (ESL, by Hastie, Tibshirani, and Friedman) — was published in 2001, with a second edition in 2009. ESL has become a popular text not only in statistics but also in related fields. One of the reasons for ESL’s popularity is its relatively accessible style. But ESL is best-suited for individuals with advanced training in the mathematical sciences.

An Introduction to Statistical Learning (ISL) arose from the clear need for a broader and less technical treatment of the key topics in statistical learning. The intention behind ISL is to concentrate more on the applications of the methods and less on the mathematical details. Beginning with Chapter 2, each chapter in ISL contains a lab illustrating how to implement the statistical learning methods seen in that chapter using the popular statistical software package **R**. These labs provide the reader with valuable hands-on experience.

ISL is appropriate for advanced undergraduates or master’s students in Statistics or related quantitative fields, or for individuals in other disciplines who wish to use statistical learning tools to analyze their data. It can be used as a textbook for a course spanning two semesters.

The first edition of ISL covered a number of important topics, including sparse methods for classification and regression, decision trees, boosting, support vector machines, and clustering. Since it was published in 2013, it has become a mainstay of undergraduate and graduate classrooms across the United States and worldwide, as well as a key reference book for data scientists.

In this second edition of ISL, we have greatly expanded the set of topics covered. In particular, the second edition includes new chapters on deep learning (Chapter 10), survival analysis (Chapter 11), and multiple testing (Chapter 13). We have also substantially expanded some chapters that were part of the first edition: among other updates, we now include treatments of naive Bayes and generalized linear models in Chapter 4, Bayesian additive regression trees in Chapter 8, and matrix completion in Chapter 12. Furthermore, we have updated the **R** code throughout the labs to ensure that the results that they produce agree with recent **R** releases.

We are grateful to these readers for providing valuable comments on the first edition of this book: Pallavi Basu, Alexandra Chouldechova, Patrick Danaher, Will Fithian, Luella Fu, Sam Gross, Max Grazier G'Sell, Courtney Paulson, Xinghao Qiao, Elisa Sheng, Noah Simon, Kean Ming Tan, Xin Lu Tan. We thank these readers for helpful input on the second edition of this book: Alan Agresti, Iain Carmichael, Yiqun Chen, Erin Craig, Daisy Ding, Lucy Gao, Ismael Lemhadri, Bryan Martin, Anna Neufeld, Geoff Tims, Carsten Voelkmann, Steve Yadlowsky, and James Zou. We also thank Anna Neufeld for her assistance in reformatting the **R** code throughout this book. We are immensely grateful to Balasubramanian “Naras” Narasimhan for his assistance on both editions of this textbook.

It has been an honor and a privilege for us to see the considerable impact that the first edition of ISL has had on the way in which statistical learning is practiced, both in and out of the academic setting. We hope that this new edition will continue to give today's and tomorrow's applied statisticians and data scientists the tools they need for success in a data-driven world.

It's tough to make predictions, especially about the future.

-Yogi Berra

Contents

Preface	vii
1 Introduction	1
2 Statistical Learning	15
2.1 What Is Statistical Learning?	15
2.1.1 Why Estimate f ?	17
2.1.2 How Do We Estimate f ?	21
2.1.3 The Trade-Off Between Prediction Accuracy and Model Interpretability	24
2.1.4 Supervised Versus Unsupervised Learning	26
2.1.5 Regression Versus Classification Problems	28
2.2 Assessing Model Accuracy	29
2.2.1 Measuring the Quality of Fit	29
2.2.2 The Bias-Variance Trade-Off	33
2.2.3 The Classification Setting	37
2.3 Lab: Introduction to R	42
2.3.1 Basic Commands	43
2.3.2 Graphics	45
2.3.3 Indexing Data	47
2.3.4 Loading Data	48
2.3.5 Additional Graphical and Numerical Summaries . .	50
2.4 Exercises	52
3 Linear Regression	59
3.1 Simple Linear Regression	60
3.1.1 Estimating the Coefficients	61
3.1.2 Assessing the Accuracy of the Coefficient Estimates	63
3.1.3 Assessing the Accuracy of the Model	68
3.2 Multiple Linear Regression	71
3.2.1 Estimating the Regression Coefficients	72
3.2.2 Some Important Questions	75
3.3 Other Considerations in the Regression Model	83

3.3.1	Qualitative Predictors	83
3.3.2	Extensions of the Linear Model	87
3.3.3	Potential Problems	92
3.4	The Marketing Plan	103
3.5	Comparison of Linear Regression with K -Nearest Neighbors	105
3.6	Lab: Linear Regression	110
3.6.1	Libraries	110
3.6.2	Simple Linear Regression	111
3.6.3	Multiple Linear Regression	114
3.6.4	Interaction Terms	116
3.6.5	Non-linear Transformations of the Predictors	116
3.6.6	Qualitative Predictors	119
3.6.7	Writing Functions	120
3.7	Exercises	121
4	Classification	129
4.1	An Overview of Classification	130
4.2	Why Not Linear Regression?	131
4.3	Logistic Regression	133
4.3.1	The Logistic Model	133
4.3.2	Estimating the Regression Coefficients	135
4.3.3	Making Predictions	136
4.3.4	Multiple Logistic Regression	137
4.3.5	Multinomial Logistic Regression	140
4.4	Generative Models for Classification	141
4.4.1	Linear Discriminant Analysis for $p = 1$	142
4.4.2	Linear Discriminant Analysis for $p > 1$	145
4.4.3	Quadratic Discriminant Analysis	152
4.4.4	Naive Bayes	153
4.5	A Comparison of Classification Methods	158
4.5.1	An Analytical Comparison	158
4.5.2	An Empirical Comparison	161
4.6	Generalized Linear Models	164
4.6.1	Linear Regression on the Bikeshare Data	164
4.6.2	Poisson Regression on the Bikeshare Data	167
4.6.3	Generalized Linear Models in Greater Generality	170
4.7	Lab: Classification Methods	171
4.7.1	The Stock Market Data	171
4.7.2	Logistic Regression	172
4.7.3	Linear Discriminant Analysis	177
4.7.4	Quadratic Discriminant Analysis	179
4.7.5	Naive Bayes	180
4.7.6	K -Nearest Neighbors	181
4.7.7	Poisson Regression	185

4.8	Exercises	189
5	Resampling Methods	197
5.1	Cross-Validation	198
5.1.1	The Validation Set Approach	198
5.1.2	Leave-One-Out Cross-Validation	200
5.1.3	k -Fold Cross-Validation	203
5.1.4	Bias-Variance Trade-Off for k -Fold Cross-Validation	205
5.1.5	Cross-Validation on Classification Problems	206
5.2	The Bootstrap	209
5.3	Lab: Cross-Validation and the Bootstrap	212
5.3.1	The Validation Set Approach	213
5.3.2	Leave-One-Out Cross-Validation	214
5.3.3	k -Fold Cross-Validation	215
5.3.4	The Bootstrap	216
5.4	Exercises	219
6	Linear Model Selection and Regularization	225
6.1	Subset Selection	227
6.1.1	Best Subset Selection	227
6.1.2	Stepwise Selection	229
6.1.3	Choosing the Optimal Model	232
6.2	Shrinkage Methods	237
6.2.1	Ridge Regression	237
6.2.2	The Lasso	241
6.2.3	Selecting the Tuning Parameter	250
6.3	Dimension Reduction Methods	251
6.3.1	Principal Components Regression	252
6.3.2	Partial Least Squares	259
6.4	Considerations in High Dimensions	261
6.4.1	High-Dimensional Data	261
6.4.2	What Goes Wrong in High Dimensions?	262
6.4.3	Regression in High Dimensions	264
6.4.4	Interpreting Results in High Dimensions	266
6.5	Lab: Linear Models and Regularization Methods	267
6.5.1	Subset Selection Methods	267
6.5.2	Ridge Regression and the Lasso	274
6.5.3	PCR and PLS Regression	279
6.6	Exercises	282
7	Moving Beyond Linearity	289
7.1	Polynomial Regression	290
7.2	Step Functions	292
7.3	Basis Functions	294

7.4	Regression Splines	295
7.4.1	Piecewise Polynomials	295
7.4.2	Constraints and Splines	295
7.4.3	The Spline Basis Representation	297
7.4.4	Choosing the Number and Locations of the Knots	298
7.4.5	Comparison to Polynomial Regression	300
7.5	Smoothing Splines	301
7.5.1	An Overview of Smoothing Splines	301
7.5.2	Choosing the Smoothing Parameter λ	302
7.6	Local Regression	304
7.7	Generalized Additive Models	306
7.7.1	GAMs for Regression Problems	307
7.7.2	GAMs for Classification Problems	310
7.8	Lab: Non-linear Modeling	311
7.8.1	Polynomial Regression and Step Functions	312
7.8.2	Splines	317
7.8.3	GAMs	318
7.9	Exercises	321
8	Tree-Based Methods	327
8.1	The Basics of Decision Trees	327
8.1.1	Regression Trees	328
8.1.2	Classification Trees	335
8.1.3	Trees Versus Linear Models	338
8.1.4	Advantages and Disadvantages of Trees	339
8.2	Bagging, Random Forests, Boosting, and Bayesian Additive Regression Trees	340
8.2.1	Bagging	340
8.2.2	Random Forests	343
8.2.3	Boosting	345
8.2.4	Bayesian Additive Regression Trees	348
8.2.5	Summary of Tree Ensemble Methods	351
8.3	Lab: Decision Trees	353
8.3.1	Fitting Classification Trees	353
8.3.2	Fitting Regression Trees	356
8.3.3	Bagging and Random Forests	357
8.3.4	Boosting	359
8.3.5	Bayesian Additive Regression Trees	360
8.4	Exercises	361
9	Support Vector Machines	367
9.1	Maximal Margin Classifier	368
9.1.1	What Is a Hyperplane?	368
9.1.2	Classification Using a Separating Hyperplane	369

9.1.3	The Maximal Margin Classifier	371
9.1.4	Construction of the Maximal Margin Classifier . . .	372
9.1.5	The Non-separable Case	373
9.2	Support Vector Classifiers	373
9.2.1	Overview of the Support Vector Classifier	373
9.2.2	Details of the Support Vector Classifier	375
9.3	Support Vector Machines	379
9.3.1	Classification with Non-Linear Decision Boundaries	379
9.3.2	The Support Vector Machine	380
9.3.3	An Application to the Heart Disease Data	383
9.4	SVMs with More than Two Classes	385
9.4.1	One-Versus-One Classification	385
9.4.2	One-Versus-All Classification	385
9.5	Relationship to Logistic Regression	386
9.6	Lab: Support Vector Machines	388
9.6.1	Support Vector Classifier	389
9.6.2	Support Vector Machine	392
9.6.3	ROC Curves	394
9.6.4	SVM with Multiple Classes	396
9.6.5	Application to Gene Expression Data	396
9.7	Exercises	398

10 Deep Learning 403

10.1	Single Layer Neural Networks	404
10.2	Multilayer Neural Networks	407
10.3	Convolutional Neural Networks	411
10.3.1	Convolution Layers	412
10.3.2	Pooling Layers	415
10.3.3	Architecture of a Convolutional Neural Network . .	415
10.3.4	Data Augmentation	417
10.3.5	Results Using a Pretrained Classifier	417
10.4	Document Classification	419
10.5	Recurrent Neural Networks	421
10.5.1	Sequential Models for Document Classification . .	424
10.5.2	Time Series Forecasting	427
10.5.3	Summary of RNNs	431
10.6	When to Use Deep Learning	432
10.7	Fitting a Neural Network	434
10.7.1	Backpropagation	435
10.7.2	Regularization and Stochastic Gradient Descent . .	436
10.7.3	Dropout Learning	438
10.7.4	Network Tuning	438
10.8	Interpolation and Double Descent	439
10.9	Lab: Deep Learning	443

10.9.1	A Single Layer Network on the Hitters Data	443
10.9.2	A Multilayer Network on the MNIST Digit Data .	445
10.9.3	Convolutional Neural Networks	448
10.9.4	Using Pretrained CNN Models	451
10.9.5	IMDb Document Classification	452
10.9.6	Recurrent Neural Networks	454
10.10	Exercises	458
11	Survival Analysis and Censored Data	461
11.1	Survival and Censoring Times	462
11.2	A Closer Look at Censoring	463
11.3	The Kaplan-Meier Survival Curve	464
11.4	The Log-Rank Test	466
11.5	Regression Models With a Survival Response	469
11.5.1	The Hazard Function	469
11.5.2	Proportional Hazards	471
11.5.3	Example: Brain Cancer Data	475
11.5.4	Example: Publication Data	475
11.6	Shrinkage for the Cox Model	478
11.7	Additional Topics	480
11.7.1	Area Under the Curve for Survival Analysis	480
11.7.2	Choice of Time Scale	481
11.7.3	Time-Dependent Covariates	481
11.7.4	Checking the Proportional Hazards Assumption . .	482
11.7.5	Survival Trees	482
11.8	Lab: Survival Analysis	483
11.8.1	Brain Cancer Data	483
11.8.2	Publication Data	486
11.8.3	Call Center Data	487
11.9	Exercises	490
12	Unsupervised Learning	497
12.1	The Challenge of Unsupervised Learning	497
12.2	Principal Components Analysis	498
12.2.1	What Are Principal Components?	499
12.2.2	Another Interpretation of Principal Components .	503
12.2.3	The Proportion of Variance Explained	505
12.2.4	More on PCA	507
12.2.5	Other Uses for Principal Components	510
12.3	Missing Values and Matrix Completion	510
12.4	Clustering Methods	516
12.4.1	K -Means Clustering	517
12.4.2	Hierarchical Clustering	521
12.4.3	Practical Issues in Clustering	530
12.5	Lab: Unsupervised Learning	532

12.5.1	Principal Components Analysis	532
12.5.2	Matrix Completion	535
12.5.3	Clustering	538
12.5.4	NCI60 Data Example	542
12.6	Exercises	548
13	Multiple Testing	553
13.1	A Quick Review of Hypothesis Testing	554
13.1.1	Testing a Hypothesis	555
13.1.2	Type I and Type II Errors	559
13.2	The Challenge of Multiple Testing	560
13.3	The Family-Wise Error Rate	561
13.3.1	What is the Family-Wise Error Rate?	562
13.3.2	Approaches to Control the Family-Wise Error Rate	564
13.3.3	Trade-Off Between the FWER and Power	570
13.4	The False Discovery Rate	571
13.4.1	Intuition for the False Discovery Rate	571
13.4.2	The Benjamini-Hochberg Procedure	573
13.5	A Re-Sampling Approach to p -Values and False Discovery Rates	575
13.5.1	A Re-Sampling Approach to the p -Value	576
13.5.2	A Re-Sampling Approach to the False Discovery Rate	578
13.5.3	When Are Re-Sampling Approaches Useful?	581
13.6	Lab: Multiple Testing	582
13.6.1	Review of Hypothesis Tests	582
13.6.2	The Family-Wise Error Rate	583
13.6.3	The False Discovery Rate	586
13.6.4	A Re-Sampling Approach	588
13.7	Exercises	591
Index		597