

Springer Texts in Statistics

Gareth James  
Daniela Witten  
Trevor Hastie  
Robert Tibshirani

# An Introduction to Statistical Learning

with Applications in R



Springer

# Contents

<b>Preface</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Statistical Learning</b>	<b>15</b>
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 . . . . .	42
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 . . . . .	49
2.4 Exercises . . . . .	52

<b>3</b>	<b>Linear Regression</b>	<b>59</b>
3.1	Simple Linear Regression . . . . .	61
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 . . . . .	82
3.3.1	Qualitative Predictors . . . . .	82
3.3.2	Extensions of the Linear Model . . . . .	86
3.3.3	Potential Problems . . . . .	92
3.4	The Marketing Plan . . . . .	102
3.5	Comparison of Linear Regression with $K$ -Nearest Neighbors . . . . .	104
3.6	Lab: Linear Regression . . . . .	109
3.6.1	Libraries . . . . .	109
3.6.2	Simple Linear Regression . . . . .	110
3.6.3	Multiple Linear Regression . . . . .	113
3.6.4	Interaction Terms . . . . .	115
3.6.5	Non-linear Transformations of the Predictors . . . . .	115
3.6.6	Qualitative Predictors . . . . .	117
3.6.7	Writing Functions . . . . .	119
3.7	Exercises . . . . .	120
<b>4</b>	<b>Classification</b>	<b>127</b>
4.1	An Overview of Classification . . . . .	128
4.2	Why Not Linear Regression? . . . . .	129
4.3	Logistic Regression . . . . .	130
4.3.1	The Logistic Model . . . . .	131
4.3.2	Estimating the Regression Coefficients . . . . .	133
4.3.3	Making Predictions . . . . .	134
4.3.4	Multiple Logistic Regression . . . . .	135
4.3.5	Logistic Regression for $>2$ Response Classes . . . . .	137
4.4	Linear Discriminant Analysis . . . . .	138
4.4.1	Using Bayes' Theorem for Classification . . . . .	138
4.4.2	Linear Discriminant Analysis for $p = 1$ . . . . .	139
4.4.3	Linear Discriminant Analysis for $p > 1$ . . . . .	142
4.4.4	Quadratic Discriminant Analysis . . . . .	149
4.5	A Comparison of Classification Methods . . . . .	151
4.6	Lab: Logistic Regression, LDA, QDA, and KNN . . . . .	154
4.6.1	The Stock Market Data . . . . .	154
4.6.2	Logistic Regression . . . . .	156
4.6.3	Linear Discriminant Analysis . . . . .	161

4.6.4	Quadratic Discriminant Analysis . . . . .	163
4.6.5	$K$ -Nearest Neighbors . . . . .	163
4.6.6	An Application to Caravan Insurance Data . . . . .	165
4.7	Exercises . . . . .	168
<b>5</b>	<b>Resampling Methods</b>	<b>175</b>
5.1	Cross-Validation . . . . .	176
5.1.1	The Validation Set Approach . . . . .	176
5.1.2	Leave-One-Out Cross-Validation . . . . .	178
5.1.3	$k$ -Fold Cross-Validation . . . . .	181
5.1.4	Bias-Variance Trade-Off for $k$ -Fold Cross-Validation . . . . .	183
5.1.5	Cross-Validation on Classification Problems . . . . .	184
5.2	The Bootstrap . . . . .	187
5.3	Lab: Cross-Validation and the Bootstrap . . . . .	190
5.3.1	The Validation Set Approach . . . . .	191
5.3.2	Leave-One-Out Cross-Validation . . . . .	192
5.3.3	$k$ -Fold Cross-Validation . . . . .	193
5.3.4	The Bootstrap . . . . .	194
5.4	Exercises . . . . .	197
<b>6</b>	<b>Linear Model Selection and Regularization</b>	<b>203</b>
6.1	Subset Selection . . . . .	205
6.1.1	Best Subset Selection . . . . .	205
6.1.2	Stepwise Selection . . . . .	207
6.1.3	Choosing the Optimal Model . . . . .	210
6.2	Shrinkage Methods . . . . .	214
6.2.1	Ridge Regression . . . . .	215
6.2.2	The Lasso . . . . .	219
6.2.3	Selecting the Tuning Parameter . . . . .	227
6.3	Dimension Reduction Methods . . . . .	228
6.3.1	Principal Components Regression . . . . .	230
6.3.2	Partial Least Squares . . . . .	237
6.4	Considerations in High Dimensions . . . . .	238
6.4.1	High-Dimensional Data . . . . .	238
6.4.2	What Goes Wrong in High Dimensions? . . . . .	239
6.4.3	Regression in High Dimensions . . . . .	241
6.4.4	Interpreting Results in High Dimensions . . . . .	243
6.5	Lab 1: Subset Selection Methods . . . . .	244
6.5.1	Best Subset Selection . . . . .	244
6.5.2	Forward and Backward Stepwise Selection . . . . .	247
6.5.3	Choosing Among Models Using the Validation Set Approach and Cross-Validation . . . . .	248

6.6	Lab 2: Ridge Regression and the Lasso . . . . .	251
6.6.1	Ridge Regression . . . . .	251
6.6.2	The Lasso . . . . .	255
6.7	Lab 3: PCR and PLS Regression . . . . .	256
6.7.1	Principal Components Regression . . . . .	256
6.7.2	Partial Least Squares . . . . .	258
6.8	Exercises . . . . .	259
<b>7</b>	<b>Moving Beyond Linearity</b>	<b>265</b>
7.1	Polynomial Regression . . . . .	266
7.2	Step Functions . . . . .	268
7.3	Basis Functions . . . . .	270
7.4	Regression Splines . . . . .	271
7.4.1	Piecewise Polynomials . . . . .	271
7.4.2	Constraints and Splines . . . . .	271
7.4.3	The Spline Basis Representation . . . . .	273
7.4.4	Choosing the Number and Locations of the Knots . . . . .	274
7.4.5	Comparison to Polynomial Regression . . . . .	276
7.5	Smoothing Splines . . . . .	277
7.5.1	An Overview of Smoothing Splines . . . . .	277
7.5.2	Choosing the Smoothing Parameter $\lambda$ . . . . .	278
7.6	Local Regression . . . . .	280
7.7	Generalized Additive Models . . . . .	282
7.7.1	GAMs for Regression Problems . . . . .	283
7.7.2	GAMs for Classification Problems . . . . .	286
7.8	Lab: Non-linear Modeling . . . . .	287
7.8.1	Polynomial Regression and Step Functions . . . . .	288
7.8.2	Splines . . . . .	293
7.8.3	GAMs . . . . .	294
7.9	Exercises . . . . .	297
<b>8</b>	<b>Tree-Based Methods</b>	<b>303</b>
8.1	The Basics of Decision Trees . . . . .	303
8.1.1	Regression Trees . . . . .	304
8.1.2	Classification Trees . . . . .	311
8.1.3	Trees Versus Linear Models . . . . .	314
8.1.4	Advantages and Disadvantages of Trees . . . . .	315
8.2	Bagging, Random Forests, Boosting . . . . .	316
8.2.1	Bagging . . . . .	316
8.2.2	Random Forests . . . . .	320
8.2.3	Boosting . . . . .	321
8.3	Lab: Decision Trees . . . . .	324
8.3.1	Fitting Classification Trees . . . . .	324
8.3.2	Fitting Regression Trees . . . . .	327

8.3.3	Bagging and Random Forests . . . . .	328
8.3.4	Boosting . . . . .	330
8.4	Exercises . . . . .	332
<b>9</b>	<b>Support Vector Machines</b>	<b>337</b>
9.1	Maximal Margin Classifier . . . . .	338
9.1.1	What Is a Hyperplane? . . . . .	338
9.1.2	Classification Using a Separating Hyperplane . . . . .	339
9.1.3	The Maximal Margin Classifier . . . . .	341
9.1.4	Construction of the Maximal Margin Classifier . . . . .	342
9.1.5	The Non-separable Case . . . . .	343
9.2	Support Vector Classifiers . . . . .	344
9.2.1	Overview of the Support Vector Classifier . . . . .	344
9.2.2	Details of the Support Vector Classifier . . . . .	345
9.3	Support Vector Machines . . . . .	349
9.3.1	Classification with Non-linear Decision Boundaries . . . . .	349
9.3.2	The Support Vector Machine . . . . .	350
9.3.3	An Application to the Heart Disease Data . . . . .	354
9.4	SVMs with More than Two Classes . . . . .	355
9.4.1	One-Versus-One Classification . . . . .	355
9.4.2	One-Versus-All Classification . . . . .	356
9.5	Relationship to Logistic Regression . . . . .	356
9.6	Lab: Support Vector Machines . . . . .	359
9.6.1	Support Vector Classifier . . . . .	359
9.6.2	Support Vector Machine . . . . .	363
9.6.3	ROC Curves . . . . .	365
9.6.4	SVM with Multiple Classes . . . . .	366
9.6.5	Application to Gene Expression Data . . . . .	366
9.7	Exercises . . . . .	368
<b>10</b>	<b>Unsupervised Learning</b>	<b>373</b>
10.1	The Challenge of Unsupervised Learning . . . . .	373
10.2	Principal Components Analysis . . . . .	374
10.2.1	What Are Principal Components? . . . . .	375
10.2.2	Another Interpretation of Principal Components . . . . .	379
10.2.3	More on PCA . . . . .	380
10.2.4	Other Uses for Principal Components . . . . .	385
10.3	Clustering Methods . . . . .	385
10.3.1	$K$ -Means Clustering . . . . .	386
10.3.2	Hierarchical Clustering . . . . .	390
10.3.3	Practical Issues in Clustering . . . . .	399
10.4	Lab 1: Principal Components Analysis . . . . .	401

10.5	Lab 2: Clustering . . . . .	404
10.5.1	$K$ -Means Clustering . . . . .	404
10.5.2	Hierarchical Clustering . . . . .	406
10.6	Lab 3: NCI60 Data Example . . . . .	407
10.6.1	PCA on the NCI60 Data . . . . .	408
10.6.2	Clustering the Observations of the NCI60 Data . . .	410
10.7	Exercises . . . . .	413
	<b>Index</b>	<b>419</b>