

A Tidy Introduction To Statistical Learning

Beau Lucas

2017-11-17

Contents

Preface	5
1 Introduction	7
1.1 An Overview of Statistical Learning	7
1.2 Data Sets Used in Labs and Exercises	7
1.3 Book Website	7
2 Statistical Learning	9
2.1 What is Statistical Learning?	9

Preface

This book will serve as a source of notes and exercise solutions for *An Introduction to Statistical Learning*. My approach will be centered around the **tidyverse**. This is not a replacement for the book, which should be read front to back by all machine learning enthusiasts.

Chapter names will line up, and certain subheadings will also match. Sometimes my notes will contain text lifted straight from the book without modification. This is not an attempt to plagiarize or claim their writing as my own. My goal is for this bookdown project to be a quick stop for machine learning enthusiasts to reference high-level ideas from ISLR in a modern media format.

Chapter 1

Introduction

1.1 An Overview of Statistical Learning

Statistical learning is focused on supervised and unsupervised modeling and prediction.

1.2 Data Sets Used in Labs and Exercises

All data sets used in this book can be found in `ISLR` and `MASS` packages, with some also being found in the base R distribution.

We will utilize the `tidyverse` ecosystem to tackle the exercises and labs, as the R code found in the original textbook is outdated.

```
require(ISLR)
require(MASS)
require(tidyverse)
```

1.3 Book Website

The website and free PDF for the book can be found here:

www.statlearning.com Youtube - Statistical Learning

Chapter 2

Statistical Learning

2.1 What is Statistical Learning?

Methods to estimate functions that connect inputs to outputs.

If there exists a quantitative response variable Y and p different predictors (X_1, X_2, \dots, X_p) , we can write this relationship as:

$$Y = f(X) +$$

2.1.1 Why Estimate f ?

2.1.1.1 Prediction

We can predict Y using:

$$\hat{Y} = \hat{f}(X)$$

Accuracy of Y is dependant on:

- *reducible error*
 - \hat{f} will never be perfect estimate of f , and model can always be potentially improved
 - Even if $\hat{f} = f$, prediction would still have some error
- *irreducible error*
 - Because Y is also a function of random, there will always be variability
 - We cannot reduce the error introduced by

2.1.1.2 Inference

How does Y respond to changes in X_1, X_2, \dots, X_p ?

2.1.2 How do we estimate f ?

- Use *training data* to train method
- x_{ij} is value of j th predictor for observation i , y_i is value of response variable
 - $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$
- Using training data, apply statistical learning method estimate unknown function f
- Most statistical learning methods can be characterized as either *parametric* or *non-parametric*

2.1.2.1 Parametric Methods

Two-step model-based approach:

1. Make an assumption about functional form of f , such as “ f is linear in X ”
2. Perform procedure that uses training data to train the model
 - In case of linear model, this procedure estimates parameters $\theta_0, \theta_1, \dots, \theta_p$
 - Most common approach to fit linear model is *(ordinary) least squares*

This is *parametric*, as it reduces the problem of estimating f down to one of estimating a set of parameters. Problems that can arise: - Model will not match the true unknown form of f - If model is made more *flexible*, which generally requires estimating a greater number of parameters, *overfitting* can occur

2.1.2.2 Non-parametric Methods

Non-parametric methods do not make assumptions about the form of f . An advantage of this is that they have the potential to fit a wider range of possible shapes for f . A disadvantage is that, because there are no assumptions about the form of f , the problem of estimating f is not reduced to a set number of parameters. This means more observations are needed compared to a parametric approach to estimate f accurately.

2.1.3 The Trade-Off Between Prediction Accuracy and Model Interpretability

Restrictive models are much more interpretable than flexible ones. Flexible approaches can be so complicated that it is hard to understand how predictors affect the response.

If inference is the goal, simple and inflexible methods are easier to interpret. For prediction, accuracy is the biggest concern. However, flexible models are more prone to overfitting.

2.1.4 Supervised Versus Unsupervised Learning