A Tidy Introduction To Statistical Learning

Beau Lucas

2017-12-19

Contents

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utilize the `tidyverse` ecosystem to tackle the exercises and labs, as the `R` code found in the original textbook is outdated.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, ..., 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 alwayas be variability
  - We cannot reduce the error introduced by

2.1.1.2 Inference

How does Y respond to changes in $X_1, X_2, ..., X_p$?

2.1.2 How do we estimate f?

- Use *training data* to train method
- x_ij is value of jth predictor for observation i, y_i is value of response variable

  - $i = 1, 2, ..., n$, $j = 1, 2, ..., p$
- 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 $_0, _1, ..., _p$
   - 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 intepretable 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

Most machine learning methods can be split into *supervised* or *unsupervised* categories. Most of this textbook involves supervised learning methods, in which a model that captures the relationship between predictors and response measurements is fitted. The goal is to accurately predict the response variables for future observations, or to understand the relationship between the predictors and response.

Unsupervised learning takes place when we have a set of observations and a vector of measurements x_i, but no response y_i. We can examine the relationship between the variables or between the observations. A popular method of unsupervised learning is cluster analysis, in which observations are grouped into distinct groups based on their vector of measurements x_i. An example of this would be a company segmenting survey respondents based on demographic data, in which the goal is to ascertain some idea about potential spending habits without possessing this data.

Clustering has some drawbacks. It works best when the groups are significantly distinct from each other. In reality, it is rare for data to exhibit this characteristic. There is often overlap between observations in different groups, and clustering will inevitably place a number of observations in the wrong groups. Further more, visualization of clusters breaks down as the dimensionality of data increases. Most data contains at least several, if not dozens, of variables.

It is not always clear-cut whether a problem should be handled with supervised or unsupervised learning. There are some scenarios where only a subset of the observations have response measurements. This is a

semi-supervised learning problem, in which a statistical learning method that can utilize all observations is needed.

2.1.5 Regression Versus Classification Problems

Variables can be categorized as either *quantitative* or *qualitative*. Both qualitative and quantatitive predictors can be used to predict both types of response variables. The more important part of choosing an appropriate statistical learning method is the type of the response variable.

2.2 Assessing Model Accuracy