Chapter 1 Notes: Introduction

Statistical Learning with R

Jon Geiger

January 27, 2022

Statistical Learning is split up into **supervised learning** and **unsupervised learning**.

- *Supervised learning* “involves building a statistical modeling for predicting, or estimating, an output based on one or more inputs.”
- *Unsupervised learning* involves “inputs but no supervising output,” allowing us to learn structure from the data.

Some Key Datasets

`Wage` Data:

- Includes a number of factors relating to wages for a specific group of men from the Atlantic region of the U.S.
- Involves predicting a *quantitative* output, useful in *regression*

`Smarket` (Stock Market) Data:

- Contains daily movements in the S&P 500 in the five-year period between 2001 and 2005
- The goal is *classification*, or to predict whether the prices will increase or decrease on a given day.

`NCI60` Gene Expression Data:

- Contains 6,830 gene expression measurements for each of 64 cancer cell lines.
- This is a *clustering* problem, and we can analyze *principal components* of the data.

Purpose of the Book

“The purpose of *An Introduction to Statistical Learning* (ISL) is to facilitate the transition of statistical learning from an academic to a mainstream field.”

ISL is based on four principles:

1. Many statistical learning methods are relevant and useful in a wide range of academic and non-academic disciplines, beyond just the statistical sciences.

2. Statistical learning should not be viewed as a series of black boxes.
3. While it is important to know what job is performed by each cog, it is not necessary to have the skills to construct the machine inside the box.
4. We presume that the reader is interested in applying statistical learning methods to real-world problems.

Notation

- $\mathbf{X}$ is an $n \times p$ matrix whose (i, j)th element is x_{ij}
  - Rather than $m \times n$ for m observations of n variables, we can think of $\mathbf{X}$ as a matrix (or spreadsheet) with n rows and p columns, or n observations of p variables.
- Vectors are column vectors by default.
- x_i is the vector of **rows** of $\mathbf{X}$, with length p:

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

- $\mathbf{x}$ is the vector of the **columns** of $\mathbf{X}$, with length n:

$$\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

- y_i denotes the ith observation of the variable on which we wish to make predictions. We can write the set of all n observations in vector form as:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- A vector of length n will be denoted in lower-case bold, such as:

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$