STATISTICAL LEARNING

CHAPTER 1: INTRODUCTION

INSTRUCTOR: SEOKHO LEE

HANKUK UNIVERSITY OF FOREIGN STUDIES

2015 SPRING

Course Information

- Textbook and references:
  - **An Introduction to Statistical Learning (ISL)** (pdf is freely available) http://www-bcf.usc.edu/~gareth/ISL/
  - **The Elements of Statistical Learning (ESL)** (pdf is freely available) http://statweb.stanford.edu/~tibs/ElemStatLearn/
  - **Pattern Recognition and Machine Learning (PRML)** (pdf is available from Google search) http://research.microsoft.com/en-us/um/people/cmbishop/prml/
- This course consists of lectures and labs.
  - **Lectures** cover all chapters in **ISL**, including additional advanced materials from **ESL** and some topics from **PRML**
  - In the **Labs**, students will learn and conduct by themselves R codes associated with materials

An Overview of Statistical Learning

- **Statistical Learning** refers to a vast set of tools for understanding data
- Two classes of statistical learning methods : **supervised learning** and **unsupervised learning**
- Supervised learning
  - Sometimes called **predictive learning** or **pattern recognition**
  - Consists of **input** (**features**, **attributes**, **explanatory variable**, **covariates**, etc.) and **output** (**response variable**)
  - Regression, Classification, etc.
- Unsupervised learning
  - Sometimes called **descriptive learning** or **knowledge discovery**
  - Consists of **output only** (or input only)
  - Clustering, PCA, etc.

Example: Wage data

- Goals of study
  - examine factors that relate to wages for a group of males from the Atlantic region of the United States
  - understand the association between an employee's age, education, calendar year on his wage
- Output
  - wage : quantitative, continuous
- Input
  - age, education, year
  - quantitative : age, year
  - qualitative (ordered categorical) : education
- Supervised learning
  - predict wage from this data set
  - regression problem

Example : Wage Data

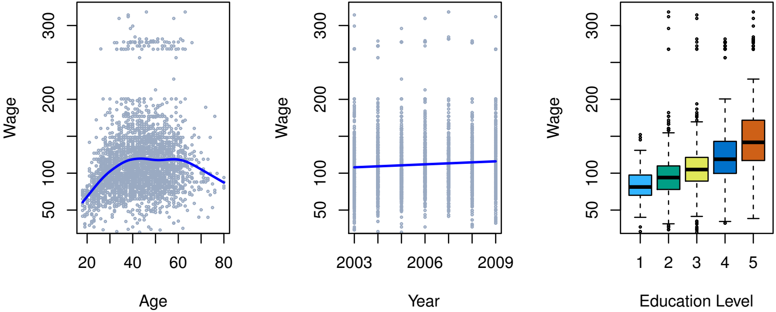

Figure 1.1: Wage data, which contains income survey information for males from the central Atlantic region of the United States. Left: wage as a function of age. On average, wage increases with age until about 60 years of age, at which point it begins to decline. Center: wage as a function of year. There is a slow but steady increase of approximately $10,000 in the average wage between 2003 and 2009. Right: Boxplot displaying wage as a function of education, with 1 indicating the lowest level (no high school diploma) and 5 the highest level (an advanced graduate degree). On average, wage increases with the level of education.

Example: Stock Market Data

- Goals of study
  - predict whether the index will increase or decrease on a given day using the past 5 days' percentage changes in the index
- Output
  - a day's performance (Up or Down) : discrete, qualitative
- Input
  - previous day's percentage changes in the stock market index (see figure) : continuous, quantitative
- Supervised learning
  - classification problem

Example : Stock Market Data

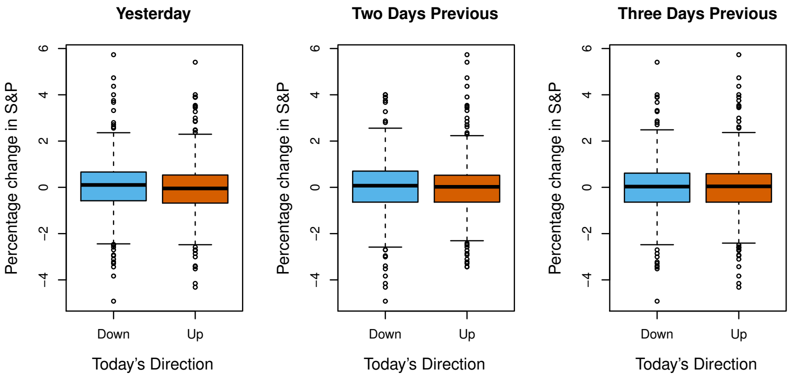

Figure 1.2: Left: Boxplots of the previous day's percentage change in the S&P index for the days for which the market increased or decreased, obtained from the Smarket data. Center and Right: Same as left panel, but the percentage changes for 2 and 3 days previous are shown.

Example : Stock Market Data

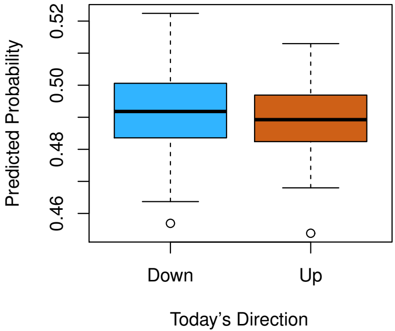

Figure 1.3: We fit a quadratic discriminant analysis model to the subset of the Smarket data corresponding to the 2001-2004 time period, and predicted the probability of a stock market decrease using the 2005 data. On average, the predicted probability of decrease is higher for the days in which the market does decrease. Based on these results, we are able to correctly predict the direction of movement in the market 60% of the time.

Example: Gene Expression Data

- Goals of study
    - We are interested in determining whether there are groups, or clusters, among the cell lines based on their gene expression measurements
- Output
    - No output exists
- Input
    - gene expression measurements : quantitative, continuous
    - 64 cancer cell lines (samples) from 6,830 gene expression measurements (variables)
- Unsupervised learning
    - clustering, principal component analysis

Example: Gene Expression Data

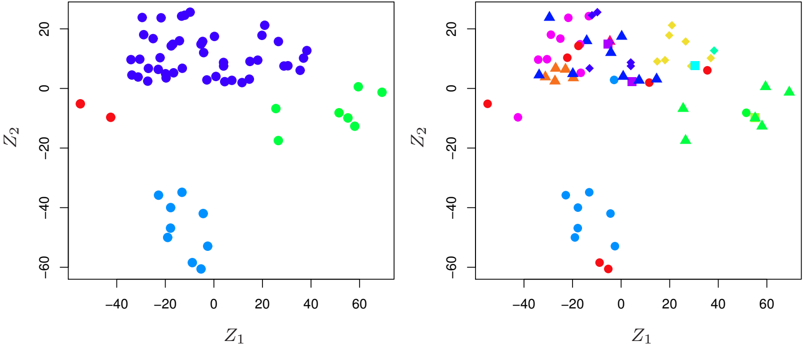

Figure 1.4: Left: Representation of the NCI60 gene expression data set in a two-dimensional space, Z_1 and Z_2. Each point corresponds to one of the 64 cell lines. There appear to be four groups of cell lines which we have represented using different colors. Right: Same as left panel except that we have represented each of the 14 different types of cancer using a different colored symbol. Cell lines corresponding to the same cancer type tend to be nearby in the two-dimensional space.

Book Website

For datasets used in this class, you can refer to the book website:
http://www-bcf.usc.edu/~gareth/ISL/

Name	Description
Auto	Gas mileage, horsepower, and other information for cars.
Boston	Housing values and other information about Boston suburbs.
Caravan	Information about individuals offers caravan insurance.
Carseats	Information about car seat sales in 400 stores.
College	Demographic characteristics, tuition, and more for USA colleges.
Default	Customer default records for a credit card company.
Hitters	Records and salaries for baseball players.
Khan	Gene expression measurements for four cancer types.
NCI60	Gene expression measurements for 64 cancer cell lines.
OJ	Sales information for Citrus Hill and Minute Maid orange juice.
Portfolio	Past values of financial assets, for use in portfolio allocation.
Smarket	Daily percentage returns for S&P 500 over a 5-year period.
USArrests	Crime statistics per 100,000 residents in 50 states of USA.
Wage	Income survey data for males in central Atlantic region of USA.
Weekly	1,089 weekly stock market returns for 21 years.

Table 1.1: A list of data sets needed to perform the labs and exercises in this textbook. All data sets are available in the ISLR library, with the exception of Boston (part of MASS) and USArrest (part of the base R distribution).