Chapter 1 notes

Karan Mann

Introduction

An Overview of Statistical Learning

Statistical learning refers to a set of tools used for understanding data. These tools are either supervised or unsupervised. Supervised learning methods use an output measurement to guide the learning process. On the other hand, unsupervised learning has no output measurement and we only have input measurements. Nonetheless, we can derive patterns and trends from this data.

Wage Data

This subsection of the text describes associations between an employee's wage, their age, the year, and their education level. Age, year and wage can act as predictors or features for a statistical model that predicts an employee's salary.

Stock Market Data

Regression problems refer to problems that involve outputs that are quantitative or continuous. However, certain problems require us to predict categorical or qualitative output. This subsection shows an example where the output is "increase" or "decreasing", referencing the day's stock index performance. These type of problems are classification problems.

Gene Expression Data

In some problems, namely unsupervised learning problems, we deal with only input variables. We may wish to observe similarities in people by grouping by their observed characteristics. This is known as a clustering problem. We are not trying to predict an output variable, but rather attempting to "cluster" together individuals with similarities. A dataset of 6830 gene expression measurements for 64 cancer cell lines is difficult because it is hard to visualize the data. However, we can reduce the dimensions to two using the first two principal components of the data. This dimension reduction has likely cost the loss of information, but it is now possible to examine the visual data for signs of clustering.

Notation and Simple Matrix Algebra

- n represents the number of distinct data points, or observations
- p denotes the number of variables available for use as predictors
- in some cases, p can be very large
- x_{ij} represents the value of the i^{th} observation for the j^{th} variable
- i will be used to index samples or observations and j will be used to index variables
- $\mathbf{X}$ will be used to denote a n x p matrix where x_{ij} is the (i,j) element
- the rows of $\mathbf{X}$ are denoted x_i^T from i = 1 to i = n
- x_i is a column vector of length p which contain the p variable measurements for the i^{th} observation
- $\mathbf{x}_j$ denotes the columns of $\mathbf{X}$ from j = 1 to j = p
- $\mathbf{x}_j$ is a column vector of length n
- we can view $\mathbf{X}$ as p column vectors of length n $\mathbf{x}_j$ or as n row vectors of length p x_i^T
- y_i denotes the i^{th} observation of the variable on which we wish to make predictions
- our observed data then consists of (x_i,y_i) where x_i is a p-vector
- a vector of length n is denoted as $\mathbf{a}$
- vectors that are not of length are denoted as a
- scalars will also be denoted as a
- matrices will be denoted as $\mathbf{A}$
- random variables will be denoted as A
- the product of two matrices $\mathbf{A}$ and $\mathbf{B}$ is $\mathbf{AB}$
- the (i,j) element of $\mathbf{AB}$ is obtained by multiplying the i^{th} row of A by the j^{th} column of $\mathbf{B}$
- in other words, $(\mathbf{AB})_{ij} = \sum_{k=1}^{d} a_{ik} b_{kj}$
- $\mathbf{AB}$ can only be computed if the number of columns of A are equal to the number of rows of B