Introduction to Data Mining and Statistical Machine Learning

Rebecca C. Steorts, Duke University

STA 325, Chapter 1 ISL

Agenda

- Notation (ISL)
- A further intro into the course
- A quick introduction to Chapter 1
- Please read this on your own
- There is no lab for Chapter 1.

Following ISL

- Please read all of Chapter 1.
- I am following the notation of the book.
- I will expect you to read and go through the chapters and labs on your own.
- We will go through this chapter very quickly.

Notation

- n: total number of observations.
- p: total number of features.
- Let x_{ij} be the value of the jth feature for the ith observation, where $i = 1, \ldots, n$ and $j = 1, \ldots p$.

Example: We have a $n = 100$ swimmers and we collect $p = 20$ features (variables) to help predict their rate of swimming.

Notation

- $\boldsymbol{X}$ denotes an $(n \times p)$ matrix whose (i, j)th element is x_{ij}.
- That is,

$$\boldsymbol{X}_{n \times p} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ x_{i1} & x_{i2} & \dots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

Notation

- At times, we will be interested in just the rows of $\boldsymbol{X}$, which we write as $(x_1, \ldots, x_n)$.
- x_i is a row vector of length p containing the p features for the ith observation.
- That is,

$$(x_i)_{p\times 1} = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

Note: vectors by default are represented by columns.

Notation

- At other times, we will be intersted in the columns of $\boldsymbol{X}$, which we write $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$.
- Each is a vector of length n, i.e.,

$$(\boldsymbol{x}_j)_{n\times 1} = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

Notation

Using this notation, $\boldsymbol{X}$, can be rewritten as

$$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots \boldsymbol{x}_p)$$

or

$$\boldsymbol{X}_{n \times p} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix},$$

where the ^{T} notation notes the transpose of a matrix of vector. (See page 11, ISL for a review).

Notation

- We use y_i to denote the ith observation of the variable on which we wish to make predictions (such as swimmers).
- Let

$$\boldsymbol{y}_{n\times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- Our observed data consists of

$$\{(x_1, y_1), \ldots (x_n, y_n)\},$$

where x_i is a vector of length p.

- If p=1, then x_i is just a scalar.

Additional Notation

- We follow additional notation, which can be found on page 11–12 of ISL
- Please make sure to read this on your own and follow this in your homeworks, exams, etc to avoid confusion.
- Each chapter of ISL has excellent R labs that you will be expected to work on your own. (There are solutions).
- If you want extra exercise to work through, please work the exercise for each chapter. I will not post solutions.

Setup

- $X_{n \times p}$: regression features or covariates (design matrix)
- $x_{p \times 1}$: ith row vector of the regression covariates
- $y_{n \times 1}$: response variable (vector)
- $\beta_{p \times 1}$: vector of regression coefficients

Goal: Estimation of $p(y \mid x)$.

Dimensions: $y_i - \beta^T x_i = (1 \times 1) - (1 \times p)(p \times 1) = (1 \times 1)$.

Health Insurance Example

- We want to predict whether or not a patient has health insurance based upon one covariate or predictor variable, income.
- Typically, we have many predictor variables, such as income, age, education level, etc.
- We store the predictor variables in a matrix $X_{n \times p}$.

Normal Regression Model

The Normal regression model specifies that

- $E[Y \mid x]$ is linear and
- the sampling variability around the mean is independent and identically (iid) from a normal distribution

$$Y_i = \beta^T x_i + e_i \tag{1}$$

$$e_1, \ldots, e_n \overset{iid}{\sim} Normal(0, \sigma^2)$$

Normal Regression Model (continued)

This allows us to write down

$$p(y_1, \dots, y_n \mid x_1, \dots x_n, \beta, \sigma^2) \tag{2}$$

$$= \prod_{i=1}^{n} p(y_i \mid x_i, \beta, \sigma^2) \tag{3}$$

$$(2\pi\sigma^2)^{-n/2} \exp\{\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2\} \tag{4}$$

Multivariate Setup

Let's assume that we have data points (x_i, y_i) available for all $i = 1, \ldots, n$.

- y is the response variable

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n \times 1}$$

- x_i is the ith row of the design matrix $X_{n \times p}$.

Consider the regression coefficients

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}_{p \times 1}$$

Multivariate Setup

$$y \mid X, \beta, \sigma^2 \sim MVN(X\beta, \sigma^2 I)$$
$$\beta \sim MVN(0, \tau^2 I)$$

The likelihood in the multivariate setting simpifies to

$$p(y_1, \ldots, y_n \mid x_1, \ldots x_n, \beta, \sigma^2) \tag{5}$$
$$(2\pi\sigma^2)^{-n/2} \exp\{\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2\} \tag{6}$$
$$(2\pi\sigma^2)^{-n/2} \exp\{\frac{-1}{2\sigma^2} (y - X\beta)^T (y - X\beta)\} \tag{7}$$

Summary

- Chapter 1 provides you with a roadmap to the course.
- We will follow the book for the most part.
- If time permits, we will cover some topics that are not in the book.
- For more advanced machine learning concepts, I highly recommend Cynthia Rudin's course on machine learning in the spring.