STAT 154: Lab 1

Yuansi Chen and Raaz Dwivedi

Jan 28, 2019

1 Three-circle notation

Use the three-circle notation to describe the statistical learning problem involved in the following scenarios.

1. We want to build a facial recognition system in Evans that recognizes an undergraduate student when he/she enters the building. We collected one picture when they first applied to the university and another one at their freshman year.
2. We are interested in predicting the gold price in the US market. Hence we collect everyday data for all of 2012.
3. We want to understand what characteristics of a company affect CEO salary. We collect a set of data on the top 500 firms in the US. Can you imagine a better way of data collection?
4. Your experience on statistical learning?

2 Basic vector and matrix manipulations in R

If you are a superstar in Stat133 and very familiar with R, skip to the next section.

1. Consider the vector x:

```
x <- 1:9
```

   Use the vector x with the function **matrix()** to create a 3×3 matrix B

2. Create the transpose of B with only the function **matrix()**
3. Print the diagonal of B with **diag()**
4. Create a diagonal matrix with all ones on the diagonal
5. Consider the following vectors v_1, v_2, v_3:

```
v1 <- c(2, 3, 4, 5)
v2 <- c(1.1, 1.2, 1.9, 2.0)
v3 <- c(100, 500, 1000, 5000)
```

Column-bind the three vectors to form a matrix C

6. Row-bind the three vectors to form another matrix E
7. Compute the matrix product CE with **%*%**.
8. Compute the matrix

$$C(C^\top C)^{-1}C^\top$$

9. Write a function **vnorm()** which takes a vector as input and outputs its Euclidean norm.
10. Write a function **mtrace()** which takes a square matrix as input and outputs its trace. If the matrix is not a square, throw an error message with **stop()**.

3 Basic data loading in R

Download the "Auto.csv" dataset from the ISL textbook website.

1. Load the data as R dataframe

```
setwd("/Users/yuansichen/UCB/Teaching/2019_Spring/Problems/stat154/labs/")
Auto <- read.csv("Auto.csv", na.strings="?")
```

2. How many rows and columns are there?
3. Use **na.omit()** to create a dataframe AutoClean without NA values
4. How many rows and columns are in AutoClean now?
5. Use **summary()** to summarize the column info of Auto.

Get familiar with **ggplot2**.

1. Plot histogram of cars in each year
2. Plot the a histogram of horsepower for cars before year 1976 and anothor histogram for cars after year 1976 in a same plot with two different colors.
3. Plot horsepower as a function of weight
4. Plot a scatterplot of mpg vs weight with each point annotated with its car name
5. Download the package **GGally**. Generate a quick matrix of pair-wise scatterplots for the first 8 columns in Auto using **ggpairs**.

4 Topics on board

You are expected to be familiar with the following concepts.

1. Expectation
2. Variance
3. Median, quantile
4. Vector space related
5. PSD matrix
6. Eigenvalues and eigenvectors