SNU Fourth Industrial Revolution Academy

Basic Math for Big Data

Homework 3

Due: August 3, 10:00 AM

Reminders

- T.A.: Chiwan Park (chiwanpark@snu.ac.kr)
- The points of this homework add up to 135.
- This has to be done individually like all the homeworks.
- Please answer clearly; illegible handwriting may get no points.
- Whenever you are making an assumption, please state it clearly.
- If you have a question about assignments, please upload your question in FIRA portal.

Submissions

- You can submit your homework in the class or via email (only PDFs are accepted).
- Do not submit the homework in a photography form.

Question 1 [13 points]

Table 1 shows the probability distribution for the number of years to earn a Doctor of Philosophy (Ph.D.) degree. Answer the following questions.

Table 1. Probability distribution

x	$P(X = x)$
3	0.05
4	0.40
5	
6	0.15
7	0.10

(a) Fill in the missing value on Table 1. [3 points]

(b) What does it mean that the values 0, 1, and 2 are not included for X on the probability distribution? [5 points]

(c) How many years do we expect to earn a Ph.D. degree? [5 points]

Question 2 [20 points]

Consider a biased coin that comes up heads with a probability $p > 0.5$. The probability of more than 25 heads in 45 tosses is approximately equal to $P(z > \frac{25-27}{3.29})$ where z follows a standard normal distribution. Answer the following questions. Note that all your answers should be rounded off to three decimal points.

(a) What is the value of p? [10 points]

(b) Find the probability of 65 heads in 100 tosses. [10 points]

Question 3 [25 points]

The probability density function of X, the decaying period (measured in seconds) of a certain type of radioactive molecule, is given by:

$$f(x) = \begin{cases} \frac{10}{x^2} & x \geq 10 \\ 0 & x < 10 \end{cases}$$

Answer the following questions.

(a) Find $P(X > 20)$. [10 points]

(b) Assume that there are 6 molecules at the beginning. What is the probability that we observe at least 3 molecules after 15 seconds? (Note that the decaying of radioactive molecule occurs independently.) [15 points]

Question 4 [10 points]

Compute $E(X)$ if X has a probability density function given by:

(a) $f(x) = \begin{cases} c(1-x^2) & -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$. [5 points]

(b) $f(x) = \begin{cases} \frac{5}{x^2} & x > 5 \\ 0 & x \leq 5 \end{cases}$. [5 points]

Question 5 [16 points]

A scientist starts to wait for a signal from his/her co-worker at 7:30 AM. Assume that the signal will arrive at some time uniformly distributed between 7:30 AM and 8 AM. Answer the following questions.

(a) What is the probability that the scientist should wait longer than 10 minutes? [8 points]

(b) Assume the signal has not arrived at 7:45 AM. What is the probability that the signal will not arrive until 7:55 AM? [8 points]

Question 6 [16 points]

Figure 1 shows 5 data points $\{(-2,-2),(0,-1),(0,0),(0,1),(2,2)\}$ on a xy-plane. We use a simple linear regression model $y = ax + b$ to describe the relationship between x and y. Answer the following questions.

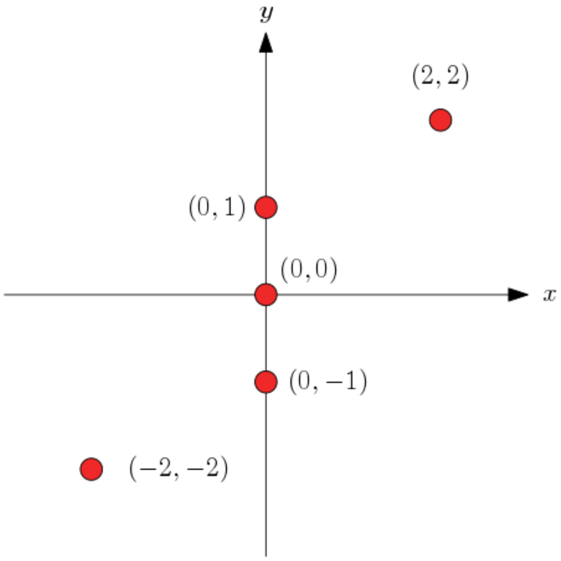

Figure 1. 5 data points on a xy-plane

(a) Derive the equations to find a and b using the least-squares method, and find the values of a and b. [8 points]

(b) Draw the best-fit line on Figure 1. [4 points]

(c) Calculate the squared error between the data points and the model. [4 points]

Question 7 [35 points]

For this programming-based question, you will write one R script (subway.R) to analyze Seoul subway traffic data. The dataset and skeleton code are contained in a zip file that you can download the following link: https://datalab.snu.ac.kr/fira/subway.zip.

The zip file contains a comma-separated-value (CSV) file (subway.csv) containing the number of passengers entering and exiting subway stations during January 2015. Each line contains the number of passengers of a subway station for a day. There are five columns in the file:

- date: the date of the observation in YYYYMMDD format (e.g. 20150101).
- line: the line number of subway station (e.g. 2호선)
- station_name: the name of subway station (e.g. 강남)
- get_in: the number of passengers entering the station for the date
- get_off: the number of passengers exiting the station for the date

Fill in the R script to answer the following questions. After that, send an email with your script to the T.A.'s email address (chiwanpark@snu.ac.kr).

(a) What is the average number of passengers of Sinchon (신촌) station with line number 2? [5 points]

(b) Find top 10 subway stations in terms of the average number of passengers. [10 points] (Hint: you may use sort function in R)

(c) Find top 3 subway lines in terms of the average number of passengers. [10 points]

(d) Draw scatter plots for the number of passengers of Nakseongdae (낙성대) station and Incheon Int'l Airport (인천국제공항) station during January 2015. You may see two different patterns in the plots. Give a reason why the patterns are different from each other. [10 points]