COMP1433: Introduction to Data Analytics
COMP1003: Statistical Tools and Applications

Assignment

PolyU Spring 2023

Answers Submission Due: 23:59, Apr 9, 2023.

Important Notes.

- This is an *individual* assessment. So, no discussion (in any forms) is allowed among classmates.
- Following the syllabus, only the R language is allowed for answering the programming questions.
- Please submit the compressed folder (in zip or rar) with all the answers. Please also name the folder with your student ID, such as "21123456D.zip" or "21123456D.rar".
- In the compressed folder, for non-coding questions, please put your answers and detailed problem solving steps into a report in PDF version (to avoid messy formula). For the coding questions with the index i below, you may create a folder named as "Qi" (e.g., Q1), which contains the codes for Q_i and the readme.txt file to explain how to run the code.
- The codes should be well commented for easy reading, and indicate clearly in the comments which part is for which sub-question (if any). It is for the case that the implementation is imperfect (with bugs) and we need to somehow find scores from the codes to see if your algorithm is designed in a correct way.
- The compressed folder should be submitted to the *blackboard*.[1] The full mark is 100' and submission entry is: Assessments/Assignment. For Question 2 below, we have provided the input data in the form of the compressed folder "Assignment_Data" available in the same entry.

[1] `learn.polyu.edu.hk`

- No late submission is allowed and don't forget to double check if the submission is saved successfully before leaving.
- When handling the paths for file loading and saving, please use relative paths for the TAs to run your codes easily in a different environment. It can be assumed that the input data file is stored in the same folder as the codes used to read that data.
- It is assumed that the external library "ggplot2" have already been installed in the R system. For any other external libraries you need to use, please indicate them in the readme.txt file.
- Last but not least, best of the luck for this assignment! :)

Question 1. [**Coding Question**] We have learned the K-means clustering algorithm during the class. In this question, you are required to implement this algorithm from scratch (without using external packages or libraries) and use it to cluster the data samples in the iris dataset into 3 clusters based on their petal length and petal width. At each training iteration, it is required to calculate and record the mean distance of data points to their respective cluster centroid.

After obtaining the clustering results, it is required to generate figures to visualize the results. In the visualization, you are first required to draw a scatter plot for all the data samples (x-axis corresponds to the petal length while y-axis corresponds to the petal width), and color each sample in red, green, and blue, indicating the cluster it has been assigned to. Then, you are required draw another line plot with x-axis corresponds to the training iteration, and y-axis corresponds to the mean distance to the cluster centroids at that iteration.

In the algorithm implementation, no external package or library should be used, while for visualization purpose, you can use any graphics or visualization libraries you prefer (e.g., "ggplot2"). For the initialization, the cluster centroids of the three classes are (1.4, 0.1), (1.3, 0.2), and (1.7, 0.1), respectively. Please run your K-means clustering algorithm for 100 iterations. (30')

Question 2. [**Coding Question**] You are given a dataset containing information about employees in a company. The dataset is stored in the file "employees.csv", which can be downloaded from the blackboard. Each observation in this dataset has 4 attributes: Employee ID, Gender, Years of Experience, and Monthly Salary.

(a) Load the data from the file employees.csv into an R data frame called "employees". Create a new column called "Salary Per Year" that contains the calculated salary per year for each employee (Monthly Salary multiplied by 12). Add this column to the "employees" data frame. (8')

(b) Calculate the average salary per year for male and female employees separately, and print the results on the screen with the following format (6'):

The average salary per year for male employees is xxx

The average salary per year for female employees is xxx

(c) Create a histogram to visualize the distribution of years of experience among all employees. Label the x-axis as "Years of Experience" and the y-axis as "Frequency". (8')

(d) Create a scatter plot to visualize the relationship between years of experience and monthly salary among all employees in the dataset. Label the x-axis as "Years of Experience" and the y-axis as "Monthly Salary". Use different colors to distinguish male and female employees. (8')

Question 3. We have introduced the idea of Bernoulli distribution and geometric distribution during the lecture and tutorial. The geometric distribution describes the number of trails until the first success, what does the distribution looks like if we further generalize the situation. In this case, suppose X denote the number of independent Bernoulli trials with the same probability of success p, until we have r successes. This is known as the negative binomial distribution.

(a) [**Non-coding Question**] Please derive the probability mass function of X (i.e., P(X=k; r, p)) of the above negative binomial distribution. (Hint: recall the binomial distribution when get r successes out of X trials). Furthermore, please derive the mean and variance of X (i.e., $\mathbb{E}(X)$ and Var(X))? (10')

(b) [**Coding Question**] Simulate the negative binomial distribution with $r = 20$, $p = 0.5$ for 10,000 times (with random seed = 42) and compare the simulated probability mass function with the theoretical ones derived in question (a). The results should be presented in two scatter plots within the same figure and different colors should be used for these two plots. In this figure, x-axis represents the number of failures until get the 20th success, and it is assumed that x is ranged from 0 to 100. (10')

(c) [**Non-coding Question**] It is a well-attested fact that cats have 9 lives. *Montgomery* is a cat who lives in a house on a busy main road. On the other side of the road is a fish shop. Every Friday, *Montgomery* sprints across the road during the rush hour to steal a haddock, and then sprints back. The probability that he is being hit by a car in any week is 1/20, independently from week to week; if he is being hit, he loses one of his lives. Find his life expectancy and the standard deviation of his lifespan in weeks if he has 9 lives left. (i.e., $\mathbb{E}(X)$, std(X) where X denotes the number of weeks he will survive). (10')

(d) [**Non-coding and Coding Question**] What is the probability that *Montgomery* will survive for another 2 years (104 weeks) if he has 1 and 9 lives left, respectively. Please provide the theoretical results as well as use R simulation to verify your results. (10’)