Women's Institute of Technology & Innovation (WITI)
Introduction to Data Science & Machine Learning
End of Semester II Examination

Course code: CSD 114
Time Allowed: 4 hours
Year of study: 1

Cohort II 2022

Instructions

- *Attempt any FOUR Questions including Question one.*
- *In answering these questions, full marks are given for clear and thoughtful explanations, not just right answers.*

1. In general it is a good idea to visualize the relationship between variables before you compute a correlation. Using the exam1 dataset. Generate a visualization of the relationship between any two continouse variables of your choice. How would you describe the relationship, if any?

   (a) Compute the descriptive statistics.

   (b) Explain and visualise the regression analysis results for variables: Confirmed cases as the dependent variable and socioeconomic variables in the dataset as the independent variables.

2. Using the taxi dataset. Investigate whether the variable of total_ amount follows a normal distribution. Plot the CDF of the total _ amount variable in the dataset.

   (a) Compute the mean and standard deviation of continous variables in the dataset.

   (b) Create a normal distribution with the same mean and standard deviation as the data, then use it to compute the normal CDF for each value in the array.

   (c) Plot the normal CDF with a gray line.

3. The underlying objective function for K-Means Cluster algorithm tries to find cluster centers such that, if the data are partitioned into the corresponding clusters, distances between data points and their closest cluster centers become as small as possible.

   Given a set of datapoints $x_1, ..., x_n$ and a positive number k , find the clusters $C_1, ..., C_k$ that minimize

   $$J = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \, ||x_i - \mu_j||_2 \tag{1}$$

   where:
   $z_{ij} \in \{0, 1\}$ defines whether of not datapoint x_i belongs to cluster C_j
   μ_j denotes the cluster center of cluster C_j
   $|| \, ||_2$ denotes the Euclidean distance

   (a) Implement a very simple K-Means clustering algorithm (clustering belongs to unsupervised learning) to cluster flowers in the Iris dataset.

4. If we start with a principal sum, P, and earn compounded interest, the total accumulated value, V, at the end of time t is:

$$V = P\left(1 + \frac{r}{n}\right)^{nt} \tag{2}$$

where r is the annual interest rate and n is the compounding frequency. For example, if you deposit $2,100 in a bank paying an annual interest rate of 3.4% compounded quarterly, compute the balance after 7 years using the above variables.

   (a) Continuing the previous example, suppose you start with the same principle and the same interest rate, but interest is compounded twice per year, so n = 2. What would the total value be after 7 years?

   (b) If interest is compounded continuously, the value after time is given by the formula:

   $$V = P\ e^{rt} \tag{3}$$

   Translate this equation into Python and use it compute the value of the investment in the previous example with continuous compounding.

5. Recall that the linear regression problem can be solved using the Least Squares method by optimizing the normal equation below:

$$\hat{\Theta} = (X^T X)^{-1} X^T y \tag{4}$$

Implement the simple linear regression and multiple linear regression using the taxi_ 2 dataset:

GOOD LUCK!