CMP4336 – Introduction to Data Mining

Homework 1

Deadline: August 24, 2020 till 23:59 (strict deadline, no extension!)

The dataset given in the following link consists of 45211 instances.

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Write a program that performs the following tasks on the above-mentioned dataset:

1) Replace the missing values using one of the methods we have discussed in the lecture hour.
2) Calculate the mean, standard deviation, mode, and skewness of all numerical attributes and report them.
3) Find the mode of each categorical variable.
4) Plot the probability density function of numerical variables and histogram of categorical variables.
5) Using y (has the client subscribed a term deposit?) attribute as the class variable, plot the scatter plots of each pair of numerical attributes.
6) Compute the distance matrix using Euclidean distance. The size of the distance matrix will N x N where N is the number of samples in the dataset and include the distances between each pair of samples.
7) Compute the distance matrix using Mahalonobis distance. The size of the distance matrix will N x N where N is the number of samples in the dataset and include the distances between each pair of samples.
8) Choose one of the discretization methods we have discussed in the lecture and discretize all numerical attributes using that method.

Guidelines

1. Use Python.
2. Submit **a single pdf** file which includes the **required output for each of the tasks given above** and **the source code** you have written. Submissions that include more than one pdf file will NOT be evaluated.
3. Submission will be made through itslearning, NOT e-mail.