CSE 572
Data Mining
Instructor- Prof. Arunabha Sen

Assignment 1

Total marks : 20

Submission Deadline : 02.07.2019 11:59 pm AZ time.

- For submission, you should submit codes and a PDF report containing the results in a zipped file (only one submission per group). The PDF file should contain names of all the members. The zipped file name should be in the following format :

  GroupName_GroupID.zip [eg.- DM_12.zip]

- Refer to the 'group formation signup' sheet in the blackboard for group ID and GroupName (Group name should be the First name of Member 1. Group ID can be obtained from the first column).

- For coding you can use both Matlab and Python. For Matlab and Python codes, include .m and .py files in the zipped folder respectively.

Accidental deaths by fatal drug overdose is a rising trend in the United States The **overdoses.csv** dataset contains information on such opioid related drug overdose fatalities. It has the 50 rows (one for each state) and the following four columns :

State : Names of states
Population : Population in a particular state
Deaths : Number of opioid casualties in that state
Abbrev : State abbreviation

Task 1) Correlation is any statistical association that refers to how close two variables are to having a linear relationship with each other. Pearson correlation coefficient is such a measure of the linear correlation between two variables X and Y. **[3]**

For the first task, calculate the Pearson correlation coefficient between the **Population** and **Deaths** columns (you may use python/ Matlab libraries).

Task 2) Construct a **bar-graph** representing the **Opioid Death Density (ODD)**, *Opioid Death Density = Number of deaths in the state/Population for that state*, for each state. There will be 50 bars (one for each state) with the height of each bar representing death density for that particular state. Give proper labels to the x and y-axes. **[5]**

Task 3) Construct a similarity matrix representing the closeness of state pairs with respect to their ODD- a state pair will have a similarity value of 1 if the difference in their ODD values is 0, and will have a value of 0 if difference in their ODD values is maximum among the ODD values of all the given pairs. **[12]**

For example, for states A, B, C and D, if their ODD values are 0.2, 0.4, 0.6 and 0.75 respectively, then , Similarity S(A,A) = S(B,B) = S(C,C) = S(D,D) = 1, and S(A,D) = S(D,A) = 0, and 0 < S(A,B) = S(B,A) < 1, etc.

An example similarity matrix structure is given below :

	AZ	**CA**	**NV**	**AR**	**NY**
AZ	1	0.8	0.2	0.6	0
CA	0.8	1	0.45	0.63	0.31
NV	0.2	0.45	1	0.82	0.56
AR	0.6	0.63	0.82	1	0.25
NY	0	0.31	0.56	0.25	1

For constructing such an similarity matrix, the similarity metric should follow these three rules:

(a) A state pair will have a similarity value S of 1 if the difference in their ODD values is 0,
(b) A pair will have a similarity value S of 0 if difference in their ODD values is maximum.
(c) Similarity between State-A and State-B should be equal to that between State-B and State-A, i.e., S(A,B) = S(B,A).

(You can form your own metric that satisfies these two criteria or use some standard similarity measures).

The output should be a 51 x 51 matrix, where the first row and first column contain the abbreviations of the state names (refer to column 4 of the dataset for abbreviations) and the remaining 50 x 50 cells will contain the similarity values respectively. The result should be saved in a file and included in the zippped file.