Data Mining
Assignment 2
[Clustering]
Deadline: 21st December 2020 10:20AM
Group Size 2 Persons.

Problem Statement:

In this assignment you have code the *k*-means clustering algorithm.

Once you have written the code for clustering, download the "Iris" and "Glass" datasets from the UCI ML repository[1].

Your code should have the ability to be executed for multiple values of *k*.

For the clustering formations obtained for each of the above value of *k*, compute the following two cluster validity indices.

Davies Bouldin Index (DBI): This index is used to compute the inter-cluster and intra-cluster relationship. Low DBI value shows that clustering formation is good. Its formulation is shown in Eq. (1), where *m* represents the number of clusters, σ_x represents the average distance between all the objects in the cluster with the midpoint of cluster σ_i, the distance between the two midpoints of two different clusters o_i and o_j is $d(o_i.o_j)$.

$$DBI = \frac{1}{M}\sum_{i=1}^{M} \max_{j \neq i}\left(\frac{\sigma_i+\sigma_j}{d(o_i.o_j)}\right) \quad (1)$$

Dunn Index (DI): It is another metric for evaluating cluster. It aims to identify the closeness of cluster having small variance among the members of cluster. In Eq. (2), the distance between two clusters is *d (k, l)*, $d'(i)$ is the distance among the data points in a cluster. DI is the measure of separation. Higher value of DI shows good clustering.

$$DI = \min_{1\leq k\leq n}\left\{\min_{1\leq l\leq n}\left\{\frac{d(k,l)}{\min\limits_{1\leq i\leq n} d'(i)}\right\}\right\} \quad (2)$$

Input:

1. It should be able provide the option to pick the file for clustering
2. Get value of *k* as input

Output:

1. Clustering results
2. Values of DBI and DI

Submission Guidelines:

1. Zero credit for no submission
2. Name your submission file <*yourregNo1_yourRegNo2*>_*HW2*
3. Email your assignment at halimzahid@gmail.com. Email subject must be "*yourregNo1_yourRegNo2*>_*HW2*"

[1] UCI Machine Learning Repository