Assignment 1

Due Date: 2/13/2023

Total Points: 100

In this exercise, you will implement k-means clustering and Fuzzy C-means clustering. You have to implement the k-means clustering and Fuzzy C-means clustering ***from scratch*** using the programming language of your choice (**without using a toolbox from R, Matlab, Python or any other programming language, please make sure you attach your code in the folder**!). Use the **Euclidean distance** for computing the distance between any two samples in the dataset. For implementing some of the principles of programming, try to modularize the code as much as possible and consider testing your algorithm on a smaller known dataset before starting the assignment. In this assignment, we will use data relating high school students' knowledge of flu to detect groups of students based on their knowledge. A dataset, Flu.csv, is provided to help testing your clustering algorithms. Please address the subparts in each section to receive full credit. Also, analysis is a crucial aspect of the assignment, so for each subpart try to answer the question in more detail.

1. K-means clustering with different number of clusters (40 points):

a. Apply k-means clustering on the Flu dataset with 3 features: 'Risk', 'NoFaceContact', 'Sick', given the number of clusters k = 2. Visualize your clusters using a 3D scatter plot.

b. Test with different numbers of clusters k, from k = 2 to k = 10. Which one you believe is the best number of clusters? Justify your response. (Hint: you may compare the 3D scatter plots with different numbers of clusters.)

c. Implement Dunn index (DI) validity measure. Repeat experiments in problem 1b and calculate corresponding DI indices. Which one you believe is the best number of clusters using the validity measure? Does it agree with your initial observation in problem 1b?

2. K-means clustering with different features (20 points):

a. Based on the best number of clusters you obtained in problem 1c and the 3 features, does adding the 'Sick' (total 4 features) improve the clustering results? Use validity measures to justify your response.

b. Based on the model in problem 2a, does adding the ‘HndWshQual’ (total 5 features) improve the clustering results? Use validity measures to justify your response.

3. Fuzzy C-means clustering (40 points):

a. Implement Fuzzy C-means and apply it with the best number of clusters you selected in problem 1 and the best combination of features you selected in problem 2. Was there any difference in the clusters as compared to the k-means clusters? (Compare using visualization tools, using centroid values, OR using some labels and observing the differences).

b. Harden the cluster assignment of Fuzzy C-means and use DI index to compare it with the k-means clustering result. Which clustering algorithm do you think produces better clusters and why?

c. Add one more feature into the model in problem 3a. Does adding this new feature improve the clustering results? If so, why or why not? Note: If you play with different features for 3 c, please mention that as well as the features you experimented and why you chose that particular additional feature.

Please make sure to submit a zipped file containing the code (separate files) and report (in pdf) in the Dropbox folder titled “Assignment 1- LastName” on Pilot.

Academic Integrity: Please note that the code and report you submit should be your work, and yours alone. If plagiarism is detected, it will be dealt with strictly and in accordance with Wright State guidelines.