Assignment 1

Due Date/Time: 9/20/2021, 11:59 PM

Total Points: 100

You will implement the K-means clustering and Fuzzy C-means clustering from scratch using a programming language of your choice. Follow software design principles and document (comment) your code clearly explaining what you did and why you did what you did. In your report, include a README that states how your code is supposed to be run to obtain the expected results.

You will use observational data collected from caregivers of people with dementia on their sleep quality. There are 20 variables in the dataset and you will be given which variables to use to cluster the data points. The dataset is included in the Assignment with filename *SleepQuality.csv*

Each row in the data represents variables recorded per observation. Use the Euclidean distance for computing the distance between any two samples in the dataset.

You will run your clustering algorithms with different combinations of variables as specified in each question.

1. **K-means clustering with different number of clusters (30 points)**

   a. Run K-means on the entire *SleepQuality* dataset with the following 2 variables: '**swsLengthM**', and '**epochCapacity**' with number of clusters K = 2. Plot your clusters using a 3D scatter plot and report (print) the centroid locations. Based on this plot, what are your thoughts on the generated clusters?

b. Test with different numbers of clusters K, running from K = 2 to K = 10 using the same variables in 1a. According to the scatter plots, which number of clusters do you think is the most appropriate? Justify your response.

c. Implement Dunn index (DI) cluster validity measure from scratch. Repeat the experiments in problem 1b and compute the corresponding DI indices. Which one do you believe is the best number of clusters according to dunn indices? Does this agree with your initial observation in problem 1b?

2. <u>K-means clustering with different variables and sample size (30 points)</u>

a. Based on the best number of clusters you obtained in problem 1c and the 2 variables, does adding the '**lengthEdaStorm**' variable (total 3 variables) improve clustering results? Use scatter plots or any other equivalent method to justify your response.

b. Based on the model in problem 2a, does adding the '**epochPeakCounter**' and '**stormPeak**' variables (total 5 variables) improve the clustering results? Plot the results and compute the dunn index to justify your response.

c. Randomly sample 100 observations and 50 observations from the data and re-run 2a and 2b, for each sample size. Plot the clustering results and compute dunn index for each sample size and compare the results with 100 and 50 observations vs the entire dataset. Justify what you observe.

d. **(Bonus)**: What happens to the relative positioning of the centroids as you sample fewer observations (100, 50, 25) from the data? Do the centroids go farther apart or do they get closer after your clustering algorithm has converged? Justify why. Plot your findings (sample size (x-axis) vs Dunn Index (y-axis)). (**Bonus: 10 points**)

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 variables you selected in problem 2 for the entire observations. 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. Select one more variable by exploring the data and add this variable into the model in problem 3a. Justify why you selected this variable. Does adding this new variable improve the clustering results? If so, why or why not? If you play with different variables for 3c, please mention that as well as the variables you experimented with and why you chose that particular additional variable.

Submit a zipped file containing your code(s) 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.