Assignment 1

Due Date: 9/17/2019

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 (a synthetic Gaussian blobs dataset, Synthetic_test_data.csv, is provided to help testing you clustering algorithms). In this assignment, we will use the BSOM dataset to detect groups in Medical School exam scores. 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 BSOM dataset with 3 features: 'all_NBME_avg_n4', 'all_PIs_avg_n131', 'HD_final', given the number of clusters k = 3. Visualize your clusters using a 3D scatter plot.

b. Test with different number 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 number of clusters.)

c. Implement Davies-Bouldin (DB) validity measure. Repeat experiments in problem 1b and calculate corresponding DB 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 ‘all_irats_avg_n34’ (total 4 features) improve the clustering results? Using validity measures to justify your response.

b. Based on model in problem 2a, does adding the ‘HA_final’ (total 5 features) improve the clustering results? Using 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 DB index to compare it with the k-means clustering result. Which clustering algorithm you think produce better clusters and why?

c. Add one more feature into the to the model in problem 3a. Does adding this new feature improve the clustering results? If so, why or why not?

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

Academic Integrity

Discussion of course contents with other students is an important part of the academic process and is encouraged. However, it is expected that course programming assignments, homework assignments, and other course assignments will be completed on an individual basis (unless specified otherwise). Students may discuss general concepts with one another, but may not, under any circumstances, work together on the actual implementation of any course assignment. If you work with other students on "general concepts" be certain to acknowledge the collaboration and its extent in the assignment. Unacknowledged collaboration will be considered dishonest. "Code sharing" (including code from previous quarters) is strictly disallowed. "Copying" or significant collaboration on any graded assignments will be considered a violation of the university guidelines for academic honesty.

If the same work is turned in by two or more students (outside the teams), all parties involved will be held equally accountable for violation of academic integrity. You are responsible for ensuring that other students do not have access to your work: do not give another student access to your account, do not leave printouts in the recycling bin, pick up your printouts promptly, do not leave your workstation unattended, etc. If you suspect that your work has been compromised notify me immediately. If you have any questions about collaboration or any other issues related to academic integrity, please see me immediately for clarification. In addition to the policy stated in this syllabus, students are expected to comply with the Wright State University Code of Student Conduct (http://www.wright.edu/students/judicial/conduct.html) and in particular the portions pertaining to Academic Integrity (http://www.wright.edu/students/judicial/integrity.html) at all times.