**COMP 642 Assignment --- MODULE 7**

1. Implement k-means from scratch

You are given a dataset X whose rows represent different data points, you are asked to perform a k-means clustering on this dataset using the Manhattan Distance, k is chosen as 3.

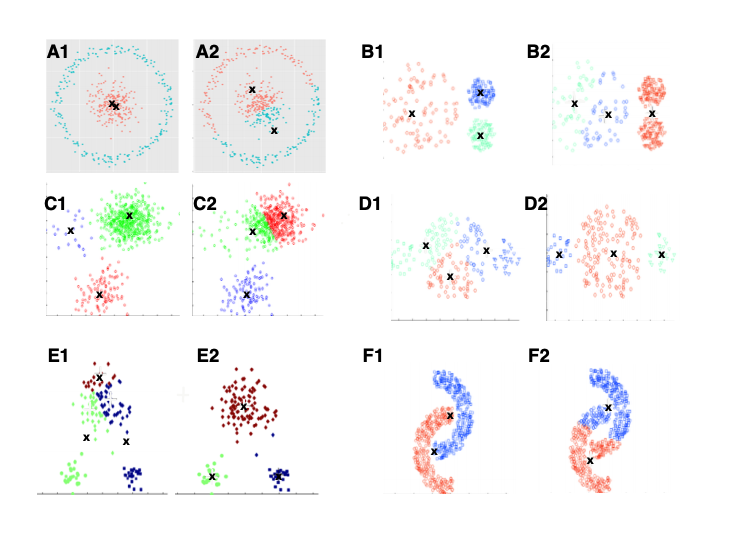
Note:

The Manhattan Distance of and is calculated by

.

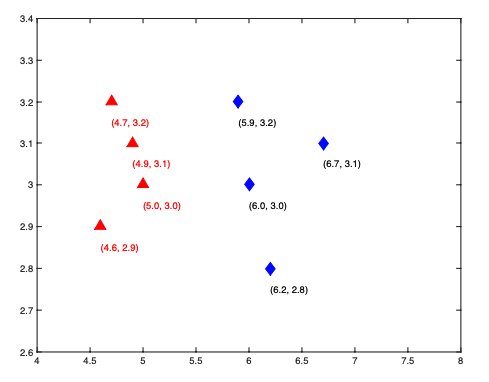
1. Since first column and second column are not on the same scale. Before running K-means, this dataset needs to be preprocessed, Show the preprocessed dataset. (Answer in the format of [x1, x2], round your results to two decimal places, same as problems b and c)
2. Suppose the initial centroids of the clusters are . What’s the center of the second cluster after two iterations?
3. What’s the center of the third cluster when the clustering converges?
4. How many iterations are required for the clusters to converge?
5. Determine the clustering result of k-means

There are 6 different datasets noted as A,B,C,D,E,F. Each dataset is clustered using two different methods, and one of them is K-means. All results are shown in Figure 2. You are required to determine which result is more likely to be generated by K-means method. (Hint: check the state when K-means converges; Centers for each cluster have been noted as X; Since x and y axis are scaled proportionally, you can determine the distance to centers geometrically). The distance measure used here is the Euclidean distance.



1. Dataset A
2. Dataset B
3. Dataset C
4. Dataset D
5. Dataset E
6. Dataset F
7. Hierarchical Clustering

Suppose there are two clusters A (red) and B (blue), each has four members and plotted in Figure below, compute the distance between two clusters using Euclidean distance.



1. What is the distance between the two farthest members (Complete-link) (round to four decimal places here, and next 2 problems)?
2. What is the distance between the two farthest members (Single-link)?
3. What is the average distance between all pairs (Average-link)?
4. Among all three distances above, which one is robust to noise?
5. Fill out the code cells in hw\_7.ipynb and answer the questions. Include the question answers in your homework document submission as well as in the Jupyter notebook.