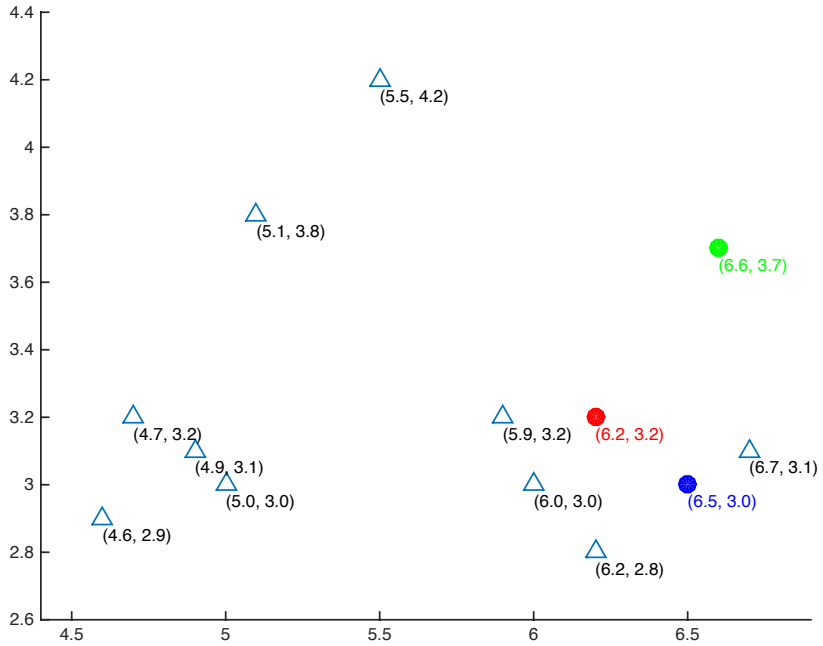
Written Assignment 4

Deadline: March 08, 2024

Instruction: You may discuss these problems with classmates, but please complete the write-ups individually. Remember the collaboration guidelines set forth in class: you may meet to discuss problems with classmates, but you may not take any written notes (or electronic notes, or photos, etc.) away from the meeting. Your answers must be **typewritten**, except for figures or diagrams, which may be hand-drawn. Please submit your answers (pdf format only) on **Canvas**.

Q1. K-means Clustering (25 points)

Given the matrix $\mathbf{X}$ whose rows represent different data points, you are asked to perform a k-means clustering on this dataset using the Euclidean distance as the distance function. Here k is chosen as 3. The Euclidean distance d between a vector x and a vector y both in $\mathbb{R}^p$ is defined as $d = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$. All data in $\mathbf{X}$ were plotted in the following figure.

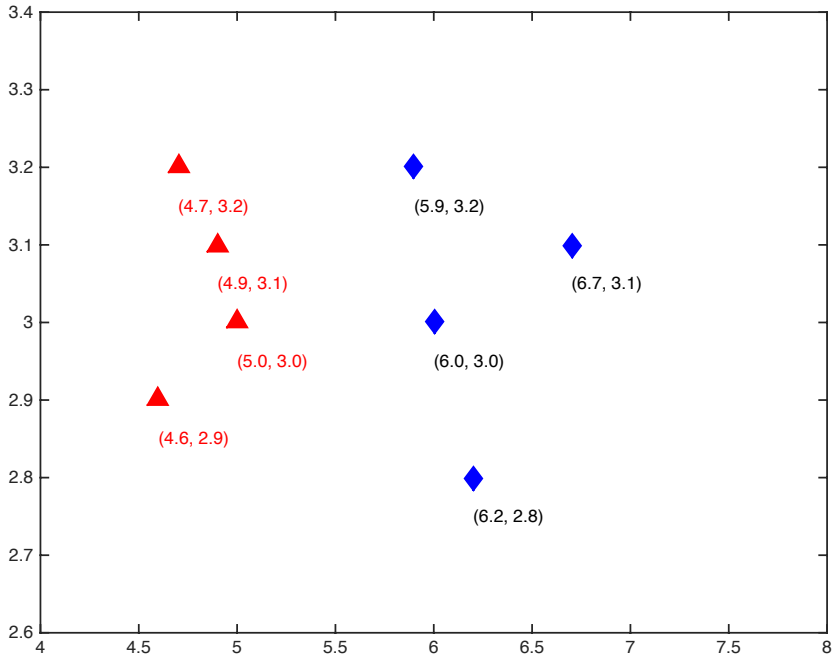


The center of 3 clusters were initialize as $\mu_1 = (6.2, 3.2)$ (red), $\mu_2 = (6.6, 3.7)$ (green), and $\mu_3 = (6.5, 3.0)$ (blue).

1. What is the center of the first cluster (red) after one iteration?
2. What is the center of the second cluster (green) after two iterations?
3. What is the center of the third cluster (blue) when the clustering converges?
4. How many iterations are required for the clusters to converge?

Q2. Hierarchical Clustering (15 points)

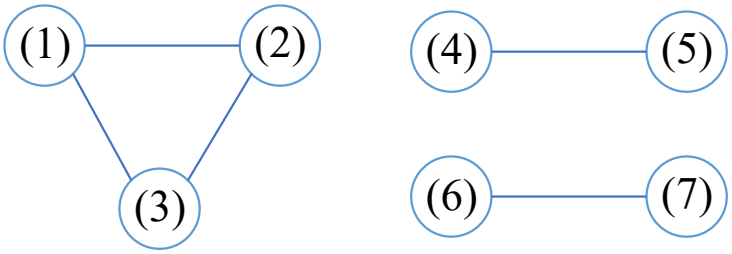
In the following figure, there are two clusters A (red) and B (blue), each has four members and plotted in the figure. The coordinates of each member are labeled in color accordingly. Compute the distance between two clusters using Euclidean distance.



1. What is the distance between the two clusters if we use the complete linkage?
2. What is the distance between the two clusters if we use the single linkage?
3. What is the distance between the two clusters if we use the average linkage?

Q3. Spectral Clustering (30 points)

Let's consider the following similarity graph in which there are six nodes which correspond to seven data samples in our dataset $(x^{(1)}, x^{(2)}, \cdots, x^{(6)})$. For example, node (1) represents the sample $x^{(1)}$. Every edge has a weight of 1.



1. Write down the Laplacian matrix of this graph.
2. What are the first three eigenvectors of the Laplacian matrix (which corresponds to the smallest eigenvalues)?
3. Write down the spectral embedding representations of the data samples using these three eigenvectors.

Q4. Principal Component Analysis (PCA) (30 points)

Consider 4 data points in the 2-d space: $(-1, -1)$, $(0.5, -0.5)$, $(1, 1)$, and $(-0.5, 0.5)$.

1. What is the first principal component?
2. If we project all points into the 1-d subspace by the first principal component, what is the new representation of the four data points?