CS1675 - Assignment 9

Zachary M. Mattis

April 2, 2019

I. Problem 1 - K-Means Clustering

S = {(0,0)(0,5)(7,0)(6,7)}

$$d(\vec{p}, \vec{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

a. Start: $\mu_1 = (0, 0), \mu_2 = (7, 0)$

After Convergence w/ Euclidean Algorithm:

$$\mu_1 = (0, 2.5) \qquad \mu_2 = (6.5, 3.5)$$

$$S_1 = \{(0, 0), (0, 5)\} \qquad S_2 = \{(7, 0), (6, 7)\}$$

b. Start: $\mu_1 = (3, 3), \mu_2 = (7, 0)$

After Convergence w/ Euclidean Algorithm:

$$\mu_1 = (2, 4) \qquad \mu_2 = (7, 0)$$

$$S_1 = \{(0, 0), (0, 5), (6, 7)\} \qquad S_2 = \{(7, 0)\}$$

II. Problem 2 - K-Means Clustering Experiments

a. K = 3

	S_1	S_2	S_3
μ	(3.94, 4.04)	(2.94, -4.97)	(0.86, 2.03)
Total	66	36	98
Plot Color	green	red	blue

Table 1: clustering_data.txt

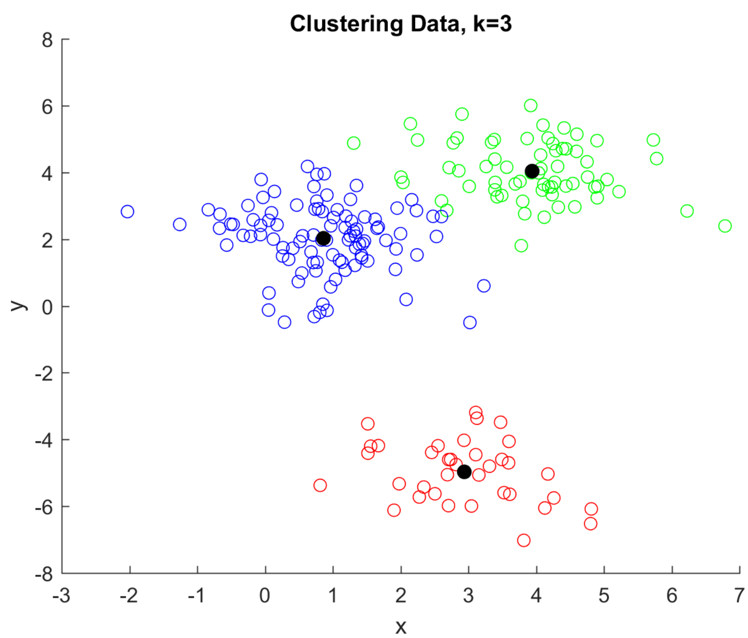

Figure 1: clustering_data.txt

b. K = 4

NOTE: Given the clustering of the data, running the algorithm on this dataset often yields different clusterings due to the starting points for the means. This concept is illustrated above from the differing resulting sets of Problem I, Parts A and B.

	S_1	S_2	S_3	S_4
μ	(4.04, 4.03)	(0.68, 2.73)	(2.94, -4.97)	(1.23, 1.04)
Total	63	61	36	40
Plot Color	red	green	blue	cyan

Table 2: clustering_data.txt

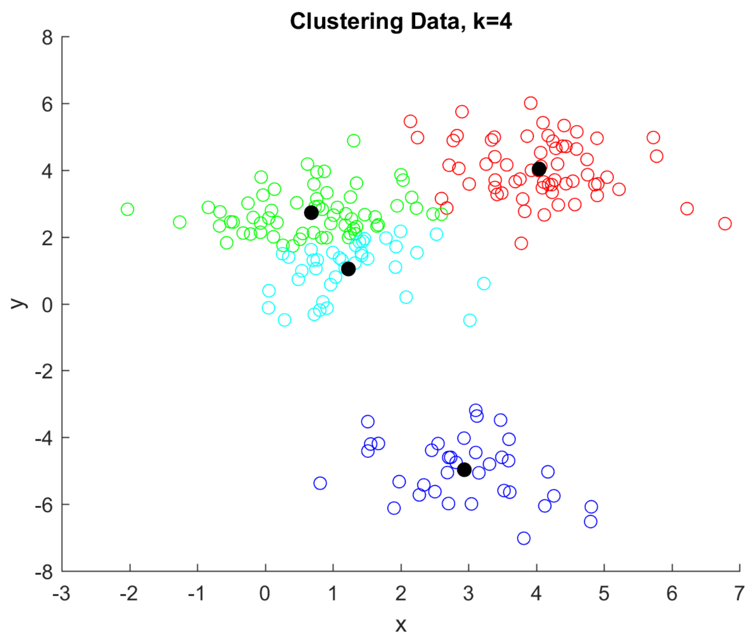

Figure 2: clustering_data.txt

c. Different Starting Means

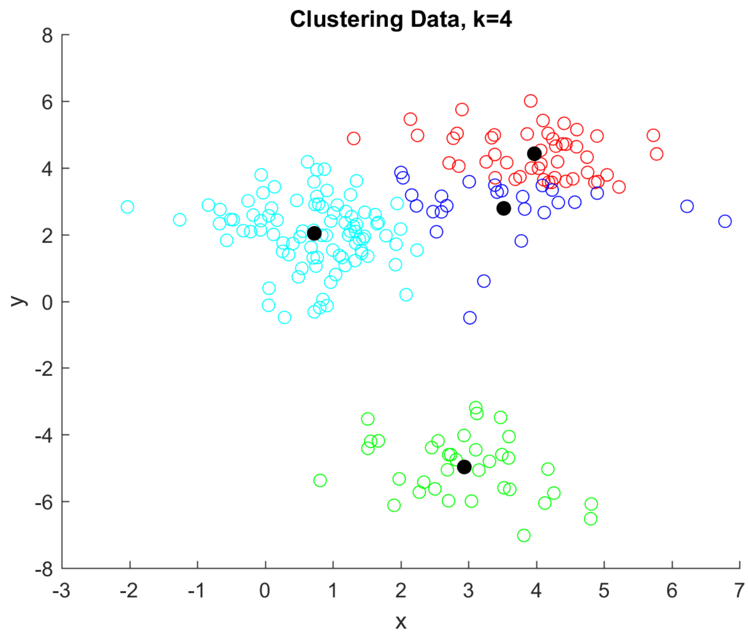

Figure 3: clustering_data.txt

d. K-Means Optimizations

Since the K-Means algorithm optimizes a distance metric, such as Euclidean Distance, it makes sense to compare the effectiveness of two different clustering via this measurement. For instance, the total distance between n data points for each given k^{th} set is given by eq. 1. By summing individual distance for each set $S \in k$, the overall effectiveness of each algorithm instance can be compared, as seen in eq. 2. Whichever algorithm instance produces the minimum total distance can be viewed as most effective.

$$d(\vec{x}, \mu_k) = \sqrt{\sum_{i=1}^{n} (\mu_k - x_i)^2} \tag{1}$$

$$\text{total distance} = \sum_{j=1}^{k} d(\vec{x}, \mu_k) \tag{2}$$

e. Random K-Means Initialization

	Run 1	Run 2	Run 3
Cluster Sizes	40,36,63,61	38,36,63,63	63,52,36,49
Total Distance	281.9109	281.9388	282.4565

Table 3: Random Initialization Analysis

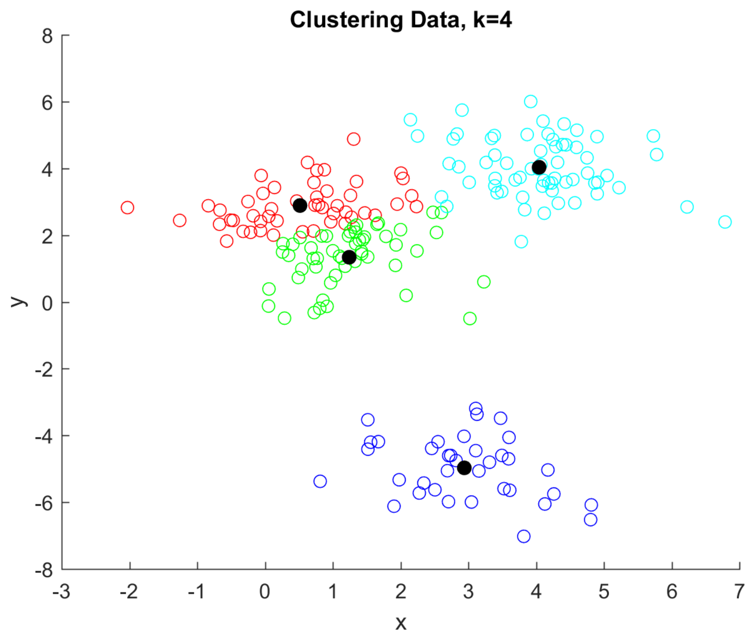

Figure 4: Optimal K=4

III. Hierarchical Clustering

a. Hierarchy of Clusters

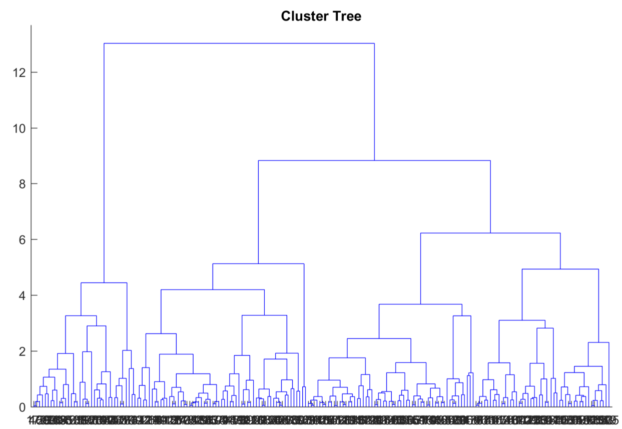

Figure 5: Dendrogram

b. Cluster Tree

The clusters for both my K-Means algorithm implementation and the built-in Matlab linkage() and cluster() functions produced the same clustering of data. This makes sense, as each technique optimizes the a distance metric to produce the best result set.

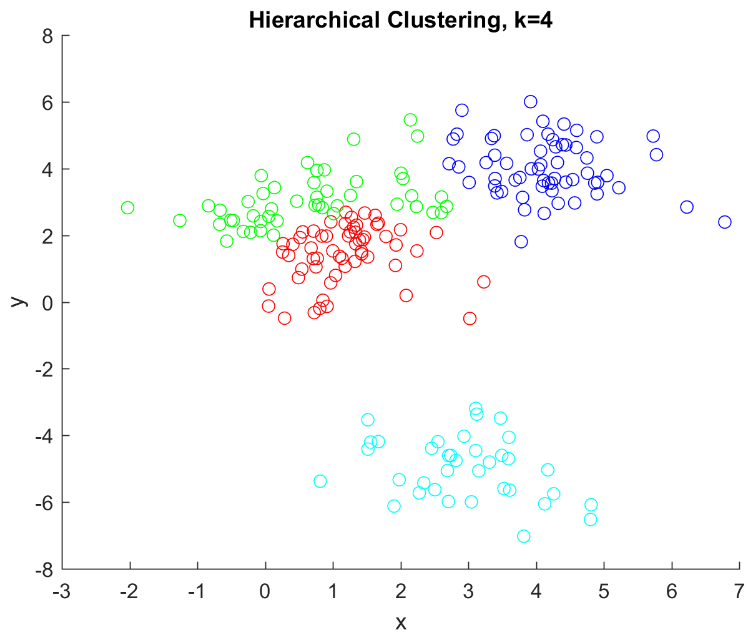

Figure 6: Cluster Tree