Homework 3: Clustering

1. Apply the k-means clustering algorithm to the following set of points and initial centroids, for k = 3. You will compute and show your work for the first three iterations of the algorithm. Although you will stop after three iterations, note that the algorithm may require up to six iterations to fully converge.

Point	**(x, y)**
1	(2, 3)
2	(3, 2)
3	(2, -3)
4	(3, -2)
5	(1, -1.5)
6	(0, -1.5)

Table 1: Set of points

Centroid	**(x, y)**
C_1	(5, 0)
C_2	(0, 5)
C_3	(0, -5)

Table 2: Initial Centroids

(a) First Iteration
Step 1: Initialize centroids (given in Table 2)
Step 2: Compute euclidean distance between each data point and each centroid. Fill in distance columns below. You can use the function we created in class. Euclidean distance: $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Point (x, y)	**Distance to C_1**	**Distance to C_2**	**Distance to C_3**	**Assigned Centroid**
(2, 3)				
(3, 2)				
(2, -3)				
(3, -2)				
(1, -1.5)				
(0, -1.5)				

Step 3: Assign each point to the cluster of closest centroid. Fill in Assigned Centroid column.
Step 4: For each cluster, average the coordinates of all of the data points, finding new centroid. To find new (x,y): $\left(\frac{\sum x_i}{n}, \frac{\sum y_i}{n}\right)$

Centroid	n	$\sum x_i$	$\sum y_i$	**(x, y)**
C_1				
C_2				
C_3				

Step5: Repeat steps 2-4 until stable. If centroids don't change from previous iteration, stop. For this problem, we will have to continue this process a few more times.

(b) Second Iteration

Step 2: Compute euclidean distance between each data point and each centroid. Fill in distance columns below. You can use the function we created in class. Euclidean distance: $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Point (x, y)	**Distance to C_1**	**Distance to C_2**	**Distance to C_3**	**Assigned Centroid**
(2, 3)				
(3, 2)				
(2, -3)				
(3, -2)				
(1, -1.5)				
(0, -1.5)				

Step 3: Assign each point to the cluster of closest centroid. Fill in Assigned Centroid column.

Step 4: For each cluster, average the coordinates of all of the data points, finding new centroid. To find new (x,y): ($\frac{\sum x_i}{n}$, $\frac{\sum y_i}{n}$)

Centroid	n	$\sum x_i$	$\sum y_i$	**(x, y)**
C_1				
C_2				
C_3				

(c) Third Iteration

Step 2: Compute euclidean distance between each data point and each centroid. Fill in distance columns below. You can use the function we created in class. Euclidean distance: $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Point (x, y)	**Distance to C_1**	**Distance to C_2**	**Distance to C_3**	**Assigned Centroid**
(2, 3)				
(3, 2)				
(2, -3)				
(3, -2)				
(1, -1.5)				
(0, -1.5)				

Step 3: Assign each point to the cluster of closest centroid. Fill in Assigned Centroid column.

Step 4: For each cluster, average the coordinates of all of the data points, finding new centroid. To find new (x,y): ($\frac{\sum x_i}{n}$, $\frac{\sum y_i}{n}$)

Centroid	n	$\sum x_i$	$\sum y_i$	**(x, y)**
C_1				
C_2				
C_3				

(d) You've completed three iterations of this algorithm so far, but it will continue to iterate until it converges. In general, how do you know when the k-means algorithm should stop iterating? Explain.

2. For the following clustering:

Point	(-2, 2)	(-1, 2)	(-1, 1)	(1, 1)	(1, 2)	(2, 2)
Cluster	0	0	0	1	1	1

(a) Compute the silhouette coefficient.

Intracluster distances

point	same cluster	Euclidean distance

Intercluster distances

point	different cluster	Euclidean distance

The average of intracluster distances: a =

The average of intercluster distances: b =

The silhouette coefficient: $\frac{b-a}{max(a,b)}$ =

(b) Is this a "good" clustering? How do you know?

3. For the following clustering:

Point	(-2, 2)	(-1, 2)	(-1, 1)	(1, 1)	(1, 2)	(2, 2)
Cluster	0	0	0	0	1	1

(a) Compute the silhouette coefficient.

Intracluster distances

point	same cluster	Euclidean distance

Intercluster distances

point	different cluster	Euclidean distance

The average of intracluster distances: a =

The average of intercluster distances: b =

The silhouette coefficient: $\frac{b-a}{max(a,b)}$ =

(b) Is this a "good" clustering? How do you know?

4. For the following clustering:

Point	(-2, 2)	(-1, 2)	(-1, 1)	(1, 1)	(1, 2)	(2, 2)
Cluster	1	0	0	0	0	1

(a) Compute the silhouette coefficient.

Intracluster distances

point	same cluster	Euclidean distance

Intercluster distances

point	different cluster	Euclidean distance

The average of intracluster distances: a =

The average of intercluster distances: b =

The silhouette coefficient: $\frac{b-a}{max(a,b)}$ =

(b) Is this a "good" clustering? How do you know?