Exercises:

Answer All question in an answer sheet (word) and submit only ONE document in the pdf format to Smartv3. Please state your name and matric number clearly on the front page.

1. Consider the 2-dimensional data set with 6 data points in the following. Show three iterations of the k-means algorithms when k = 2, and the initial centroids are data point 1 and 2.

ID	A1	A2
1	2	5
2	3	8
3	4	9
4	4	10
5	1	4
6	1	3

Answer:

Remember *k* denotes the number of clusters.

First Iteration

1. Find distance between centroid and each data points in the table except data point 1 and 2 because they have been used as initial centroid of clusters, based on euclidean distance.

   Cluster 1, Centroid c1 = 2,5 (i.e. data point 1)

   Dist(c1,3)=$\sqrt{(2-4)^2+(5-9)^2}$ = 4.47

   Dist(c1,4)= $\sqrt{(2-4)^2+(5-10)^2}$= 5.39

   Dist(c1,5)= $\sqrt{(2-1)^2+(5-4)^2}$= 1.41

   Dist(c1,6)= $\sqrt{(2-1)^2+(5-3)^2}$ = 2.24

   Cluster 2, Centroid c2 = 3,8 (i.e. data point 2)

   Dist(c2,3)=$\sqrt{(3-4)^2+(8-9)^2}$ = 1.41

   Dist(c2,4)= $\sqrt{(3-4)^2+(8-10)^2}$= 2.24

   Dist(c2,5)= $\sqrt{(3-1)^2+(8-4)^2}$= 4.47

   Dist(c2,6)= $\sqrt{(3-1)^2+(8-3)^2}$ = 5.39

2. Compare the distance between each data point and each centroid of clusters. Assign data point to its closest centroid of cluster.

Cluster 1 contains data points 1,5,6.

Cluster 2 contains data points 2,3,4.

3. Compute the means of each individual attribute value for all data points in each cluster as the new centroid of the cluster for next clustering iteration.

   New Centroid Cluster 1= (2+1+1)/3 , (5+4+3)/3=(1.33,4)

   New Centroid Cluster 2= (3+4+4)/3 , (8+9+10)/3=(3.67,9)

Second Iteration

4. Find distance between new centroid and each data points in the table including data point 1 and 2 because they are not used as centroid at this time.

   Cluster 1, Centroid c1 = 1.33,4

   Dist(c1,1)=$\sqrt{(1.33-2)^2+(4-5)^2}$ = 1.20

   Dist(c1,2)=$\sqrt{(1.33-3)^2+(4-8)^2}$ = 4.33

   Dist(c1,3)=$\sqrt{(1.33-4)^2+(4-9)^2}$ = 5.67

   Dist(c1,4)= $\sqrt{(1.33-4)^2+(\mathbf{4}-10)^2}$= 6.57

   Dist(c1,5)= $\sqrt{(1.33-1)^2+(\mathbf{4}-4)^2}$= 0.33

   Dist(c1,6)= $\sqrt{(1.33-1)^2+(4-3)^2}$ = 1.05

   Cluster 2, Centroid c2 = 3.67,9

   Dist(c1,1)=$\sqrt{(3.67-2)^2+(9-5)^2}$ = 4.33

   Dist(c1,2)=$\sqrt{(3.67-3)^2+(9-8)^2}$ = 1.20

   Dist(c2,3)=$\sqrt{(3.67-4)^2+(9-9)^2}$ = 0.33

   Dist(c2,4)= $\sqrt{(3.67-4)^2+(9-10)^2}$= 1.05

   Dist(c2,5)= $\sqrt{(3.67-1)^2+(9-4)^2}$= 5.67

   Dist(c2,6)= $\sqrt{(3.67-1)^2+(9-3)^2}$ = 6.57

5. Compare the distance between each data point and each centroid of clusters. Assign data point to its closest centroid of cluster.

   Cluster 1 contains data points 1,5,6.

   Cluster 2 contains data points 2,3,4.

*Note: Although the question request for 3 iterations, but the clustering process can be terminated at second iteration because cluster membership stabilizes (i.e. the members in cluster 1 for second iteration are same with the members in cluster 1 from first iteration while the members in cluster 2 for second iteration are same with the members in cluster 2 from first iteration).