**Homework 7**

**Name: Megha Godwal**

**UID: U01695574**

**Clustering**

1. Consider the following eight points in 2-dimensional space: (2,10); (2,5); (8,4); (5,8); (7,5); (6,4); (1,2); (4,9). Suppose we plan to use the Euclidean distance metric and that we are interested in clustering these points into 3 clusters.

1. Plot the data by hand on paper to see what might be appropriate clusters.

Ans :

A picture containing building, train, sitting, white

Description automatically generated

1. Beginning with the points (2,10), (5,8) and (1,2) as initial cluster centers, form the three initial clusters.

Ans:

Initial cluster centers / centroids: A1 = (2, 10), A2 = (5, 8), A3 = (1, 2)

Initial Clusters: C1 = {(2, 10)}, C2 = {(5, 8)}, C3 = {(1, 2)}

Using the Euclidean distance between points to form clusters:-

• (2, 5):

D((2, 5), A1) = 5

D((2, 5), A2) = 4.24

D((2, 5), A3) = 3.16 (smallest)

(2, 5) is in cluster 3

• (8, 4):

D((8, 4), A1) = 6

D((8, 4), A2) = 5 (smallest)

D((8, 4), A3) = 7.28

(8, 4) is in cluster 2

• (7, 5):

D((7, 5), A1) = 7.07

D((7, 5), A2) = 3.60 (smallest)

D((7, 5), A3) = 6.70

(7, 5) is in cluster 2

• (6, 4):

D((6, 4), A1) = 7.21

D((6, 4), A2) = 4.12 (smallest)

D((6, 4), A3) = 5.38

(6, 4) is in cluster 2

• (4, 9):

D((4, 9), A1) = 2.24

D((4, 9), A2) = 1.41 (smallest)

D((4, 9), A3) = 7.62

(4, 9) is in cluster 2

The 3 initial clusters are :

C1 = {(2,10)}

C2 = {(5,8), (8,4) , (7,5), (6,4), (4,9)}

C3 = {(1,2), (2,5)}

The 3 initial clusters are:-

C1 = {(2, 10)}

C2 = {(5, 8), (8, 4), (7, 5), (6, 4), (4, 9)}

C3 = {(1, 2), (2, 5)}

A picture containing building, white, table, sitting

Description automatically generated

1. Use the 3-means clustering algorithm to get the final three clusters. What are the resulting centers and resulting clusters?

Ans:

Using the 3 – means clustering algorithm to find the final 3 clusters :

Clusters after 1st iteration are :

C1={(2,10)}

C2 = {(5,8),(8,4), (7,5), (6,4), (4,9)}

C3 = {(1,2), (2,5)}

Updated centers :

A1 = (2,10)

A2 = ((5+8+7+6+4)/5, (8+4+5+4+9)/5) = (6,6)

A3 = ((1+2)/2, (2+5)/2) = (1.5,3.5)

Clusters after 2

nd

iteration are:-

C1 = {(2, 10), (4, 9)}

C2 = {(5, 8), (8, 4), (7, 5), (6, 4)}

C3 = {(1, 2), (2, 5)}

Updated centers:

A1 = (3, 9.5)

A2 = (6.5, 5.25)

A3 = (1.5, 3.5

Clusters after 2

nd

iteration are:-

C1 = {(2, 10), (4, 9)}

C2 = {(5, 8), (8, 4), (7, 5), (6, 4)}

C3 = {(1, 2), (2, 5)}

Updated centers:

A1 = (3, 9.5)

A2 = (6.5, 5.25)

A3 = (1.5, 3.5

Clusters after 2

nd

iteration are:-

C1 = {(2, 10), (4, 9)}

C2 = {(5, 8), (8, 4), (7, 5), (6, 4)}

C3 = {(1, 2), (2, 5)}

Updated centers:

A1 = (3, 9.5)

A2 = (6.5, 5.25)

A3 = (1.5, 3.5

Clusters after 2

nd

iteration are:-

C1 = {(2, 10), (4, 9)}

C2 = {(5, 8), (8, 4), (7, 5), (6, 4)}

C3 = {(1, 2), (2, 5)}

Updated centers:

A1 = (3, 9.5)

A2 = (6.5, 5.25)

A3 = (1.5, 3.5

Clusters after 2nd iteration are :

C1={(2,10), (4,9)}

C2 = {(5,8),(8,4), (7,5), (6,4)}

C3 = {(1,2), (2,5)}

Updated centers :

A1= (3,9.5)

A2 = (6.5, 5.25)

A3 = (1.5, 3.5)

Clusters after 3rd iteration are :

C1={(2,10), (5,8),(4,9)}

C2 = {(8,4), (7,5), (6,4)}

C3 = {(1,2), (2,5)}

Updated centers :

A1= (3.66,9)

A2 = (7,4.33)

A3 = (1.5, 3.5)

1. Mark the initial and final clusters on your graph and comment on what you see.

Ans: Initial and Final Clusters :

A close up of a fence

Description automatically generated

In the above, the outlines represent the initial clusters and the shaded regions represent the final clusters. With every iteration, we see that the clusters get precise and updated and in turn the centroids get updated too. Cluster 1 and Cluster 2 points changed however Cluster 3 points were the same through all the iterations from initial to final stage.

2. k-means and EM clustering are available in the Weka Explorer. Both of them produce measures of the goodness of the clustering. For the simplified Iris and Glass datasets, graph the values of these measures for k-means and EM clustering from k = 1 to k = 5 clusters for both datasets (4 graphs total). What value of k seems best in each case?

Ans :

• Iris Dataset:-

K-means Clustering-

Values of k tested: 1, 2, 3, 4, 5

Best case: k = 3

• Iris Dataset:-

EM Clustering-

Values of k tested: 1, 2, 3, 4, 5

Best case: k = 3

• Glass Dataset:-

K-means Clustering-

Values of k tested: 1, 2, 3, 4, 5

Best case: k = 4

• Glass Dataset:-

EM Clustering-

Values of k tested: 1, 2, 3, 4, 5

Best case: k = 3

The 3 initial clusters are:-

C1 = {(2, 10)}

C2 = {(5, 8), (8, 4), (7, 5), (6, 4), (4, 9)}

C3 = {(1, 2), (2, 5)}

The 3 initial clusters are:-

C1 = {(2, 10)}

C2 = {(5, 8), (8, 4), (7, 5), (6, 4), (4, 9)}

C3 = {(1, 2), (2, 5)}

The 3 initial clusters are:-

C1 = {(2, 10)}

C2 = {(5, 8), (8, 4), (7, 5), (6, 4), (4, 9)}

C3 = {(1, 2), (2, 5)}

The 3 initial clusters are:-

C1 = {(2, 10)}

C2 = {(5, 8), (8, 4), (7, 5), (6, 4), (4, 9)}

C3 = {(1, 2), (2, 5)}