**NAME: SRUJAN JADHAV**

**CWID: 10445782**

**Assignment no. 5**

1. *Suppose that we have the following data:  
   a b c d e f g h i j  
   (2,0) (1,2) (2,2) (3,2) (2,3) (3,3) (2,4) (3,4) (4,4) (3,5)*

*Identify the cluster by applying the k-means algorithm, with k = 2. Try using initial cluster centers as far apart as possible.*

**Answer:**

Given data: K=2

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| a | b | c | d | e | f | g | h | i | j |
| (2,0) | (1,2) | (2,2) | (3,2) | (2,3) | (3,3) | (2,4) | (3,4) | (4,4) | (3,5) |

As per question, farthest points are (1, 2) and (3, 5)

So, m1 = (1, 2) & m2 = (3, 5)

Calculate the distance of each point from m1 & m2 and divide them in to clusters C1 and C2.

Euclidean distance formula: *dEuclidean* (x, y) = 2

|  |  |  |  |
| --- | --- | --- | --- |
| Points | Distance from Centre1  (1, 2) | Distance from Centre2  (3, 5) | Cluster |
| a (2,0) | 2.236 | 5.099 | C1 |
| b (1,2) | 0 | 3.605 | C1 |
| c (2,2) | 1 | 3.162 | C1 |
| d (3,2) | 2 | 3 | C1 |
| e (2,3) | 1.414 | 2.236 | C1 |
| f (3,3) | 2.236 | 2 | C2 |
| g (2,4) | 2.236 | 1.414 | C2 |
| h (3,4) | 2.828 | 1 | C2 |
| I (4,4) | 3.605 | 1.414 | C2 |
| j (3,5) | 3.605 | 0 | C2 |

Points in cluster C1 are (a, b, c, d, e) and that in C2 are (f, g, h, i, j)

SSE = =

Hence, SSE= 20.997

= 3.605/20.997 = 0.171

New Centroid: C1 = = (2, 1.8)

New Centroid: C2 = = (3, 4)

**2ND Iteration:**

|  |  |  |  |
| --- | --- | --- | --- |
| Points | Distance from Centre1  (2, 1.8) | Distance from Centre2  (3, 4) | Cluster |
| a (2,0) | 1.8 | 4.123 | C1 |
| b (1,2) | 1.183 | 2.828 | C1 |
| c (2,2) | 0.2 | 2.236 | C1 |
| d (3,2) | 1.183 | 2 | C1 |
| e (2,3) | 1.2 | 1.414 | C1 |
| f (3,3) | 1.562 | 1 | C2 |
| g (2,4) | 2.2 | 1 | C2 |
| h (3,4) | 2.416 | 0 | C2 |
| I (4,4) | 2.973 | 1 | C2 |
| j (3,5) | 3.352 | 1 | C2 |

SSE = =

SSE = 11.518

= 2.416/11.518 = 0.209

New Centroid: C1 = = (2, 1.8)

New Centroid: C2 = = (3, 4)

When the value of the centroid of two clusters doesn’t change then K-means algorithm terminates.

**Cluster 1 = {a, b, c, d, e}**

**Cluster 2 = {f, g, h, i, j}**