ASSIGNMENT- 5

Poorvi Raut- 20009560

Chapter 8, page 162, problems 6 (2004 edition)

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.

Solution:

Given 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)

Step 1: k = 2 specifies number of clusters to partition

Step 2: Randomly assign k=2 cluster centers for example m1= (2,0) and m2= (3,5)

First Iteration:

Step 3: For each record find nearest cluster center by calculating the Euclidean distance between the points and cluster centers and determine the closest values to m1 and m2 and divide in clusters of k=2

Euclidean distance between points x (x1, x2) and y (y1, y2)= sqrt((x1-y1)^2 + (x2-y2)^2)

Point	a	b	c	d	e	f	g	h	i	j
Distance from m1	0	2.236	2	2.236	3	3.16	4	4.123	4.47	5.099
Distance from m2	5.09	3.6055	3.16	3	2.236	2	1.414	2	1.414	0
Cluster Membership	C1	C1	C1	C1	C2	C2	C2	C2	C2	C2

cluster m1 contains: {a,b,c,d} and cluster m2 contains {e,f,g,h,i,j}

cluster membership is assigned and now calculate SSE

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, m_i)^2$$

= 0+(2.236)^2+(2)^2+(2.236)^2+(2.236)^2+(2)^2+(1.414)^2+(2)^2+(1.414)^2+0 = 27

Recall clusters constructed where between-cluster variation (BCV) large, as compared to within-cluster variation (WCV)

$$\frac{\text{BCV}}{\text{WCV}} = \frac{d(m_1, m_2)}{\text{SSE}} = \frac{5}{27} = 0.185185.., \text{ where}$$
$$d(m_1, m_2) = \text{surrogate for BCV}$$
$$\text{SSE} = \text{surrogate for WCV}$$

Ratio BCV/WCV expected to increase for successive iterations.

Step 4: For *k* clusters, find cluster centroid, update location. Calculate the new cluster centers as the mean of the data points assigned to each cluster.

Cluster 1: Mean = ((2+1+2+3)/4 ,(0+2+2+2)/4)= (2,1.5)

Cluster 2: Mean = ((2+3+2+3+4+3)/6, (3+3+4+5+4+5)/6) = (2.83, 4)

Step 5: Repeats Steps 3 – 4 until convergence or termination

Second Iteration: Repeat steps 3 and 4

Again m1= (2,1.5) m2= (2.83,4). calculating the Euclidean distance between the points and cluster centers and determine the closest to new values to m1 and m2 and divide in clusters of k=2

Point	a	b	c	d	e	f	g	h	i	j
Distance from m1	1.5	1.118	0.5	1.118	1.5	1.802	2.5	2.69	3.20	3.64
Distance from m2	4.085	2.710	2.16	2.007	1.29	1.014	0.83	0.17	1.17	1.014
Cluster Membership	C1	C1	C1	C1	C2	C2	C2	C2	C2	C2

cluster m1 contains: {a,b,c,d} and cluster m2 contains {e,f,g,h,i,j}

cluster membership is assigned and now calculate SSE

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, m_i)^2$$

=(1.5)^2+(1.118)^2+(0.5)^2+(1.118)^2+(1.5)^2+(1.014)^2+(0.83)^2+(0.17)^2+(1.17)^2+(1.014)^2 = 11.39

Recall clusters constructed where between-cluster variation (BCV) large, as compared to within-cluster variation (WCV)

$$\frac{\text{BCV}}{\text{WCV}} = \frac{d(m_1, m_2)}{\text{SSE}} = \frac{2.634}{11.39} = 0.2312 \text{ where}$$
$$d(m_1, m_2) = \text{surrogate for BCV}$$
$$\text{SSE} = \text{surrogate for WCV}$$

Ratio BCV/WCV increases as compared to previous iteration 0.185 to 0.2312.

Step 4: For *k* clusters, find cluster centroid, update location. Calculate the new cluster centers as the mean of the data points assigned to each cluster.

Cluster 1: Mean = ((2+1+2+3)/4, (0+2+2+2)/4) = (2,1.5)

Cluster 2: Mean = ((2+3+2+3+4+3)/6, (3+3+4+5+4+5)/6) = (2.83, 4)

Step 5: Repeat steps 3 and 4 until convergence or termination. Since the mean values of clusters /centroids remain unchanged, the algorithm terminates.