Q1.) **K-means**

Given a data set with the following points, please manually run the K-means algorithm with two iterators

(the initial iteration and one more iteration) and report the discovered clusters for K=2.

x1=(1, 1), x2=(5, 5), x3 =(5, 6), x4=(7, 8), x5=(8, 6), x6=(8, 7)

Run the algorithm by utilizing x1 and x2 as the initial cluster centroids. At each iteration, show the

initial centroids, the clusters formed at that iteration, the new centroids, and the SSE value. Use Euclidian

distance to calculate distances.

Ans.) First, find the distance between each point by using Euclidian distance,

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | X1 | X2 | X3 | X4 | X5 | X6 |
| **X1** | 0 | 5.656 | 6.403 | 9.219 | 8.602 | 9.219 |
| **X2** | 5.656 | 0 | 1 | 3.605 | 3.162 | 3.605 |
| **X3** | 6.403 | 1 | 0 | 2.828 | 3 | 3.162 |
| **X4** | 9.219 | 3.605 | 2.828 | 0 | 2.236 | 1.414 |
| **X5** | 8.602 | 3.162 | 3 | 2.236 | 0 | 1 |
| **X6** | 9.219 | 3.605 | 3.162 | 1.414 | 1 | 0 |

Given the points x1, x2 be initial cluster centroids for the **first iteration**,

Step-1: Let the points x1 and x2 be the cluster centroids. Then calculate the distance between each point and the centroids.

|  |  |  |
| --- | --- | --- |
|  | X1 | X2 |
| **X1** | 0 | 5.65 |
| **X2** | 5.65 | 0 |
| **X3** | 6.40 | 1 |
| **X4** | 9.21 | 3.60 |
| **X5** | 8.60 | 3.16 |
| **X6** | 9.21 | 3.60 |

Step-2: The clusters formed at the first Iteration are:

C1 🡪 0.

C2 🡪 0, 1, 3.60, 3.16, 3.60.

Step-3: The SSE (Sum of Squared Error) value is:

We know that for SSE = sum of square of clusters distances.

02 + 02 + 12 + 3.602 + 3.162 + 3.602 = 36.9.

Step-4: The new centroids for the first iteration by using clusters C1 & C2 are:

We know that

Centroid for cluster C1 is (x1, x1).

Centroid for cluster C2 is (x2, x2), (x3, x2), (x4, x2), (x5, x2), (x6, x2).

Centroid for cluster C1 is (1, 1) and the centroid for cluster C2 is (6.6, 6.4).

Let the clusters C1 & C2 be the centroids for the **second iteration**,

Step-1: Let the clusters C1 & C2 be the centroids. Then calculate the distance between each point and the centroids.

|  |  |  |
| --- | --- | --- |
|  | C1 (1,1) | C2 (6.6, 6.4) |
| **X1** | 0 | 7.77 |
| **X2** | 5.65 | 2.12 |
| **X3** | 6.40 | 1.64 |
| **X4** | 9.21 | 1.64 |
| **X5** | 8.60 | 1.45 |
| **X6** | 9.21 | 1.52 |

Step-2: The clusters formed at the second Iteration are:

C3 🡪 0.

C4 🡪 2.12, 1.64, 1.64, 1.45, 1.52.

Step-3: The SSE (Sum of Squared Error) value is:

We know that for SSE = sum of square of clusters distances.

02 + 2.122 + 1.642 + 1.642 + 1.452 + 1.522 = 14.26.

Step-4: The new centroids for the first iteration by using clusters C3 & C4 are:

We know that

Centroid for cluster C3 is (x1, c1).

Centroid for cluster C4 is (x2, c2), (x3, c2), (x4, c2), (x5, c2), (x6, c2).

Centroid for cluster C3 is (1, 1) and the centroid for cluster C4 is (6.6, 6.4).

Q2.) **Hierarchical**

Use the distance matrix in the following table to do hierarchical clustering.

(a) Use MIN to calculate cluster distances.

At each step show the updated matrix with distances between clusters. Show your results by drawing the final dendrogram.

Ans.)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | P1 | P2 | P3 | P4 | P5 |
| **P1** | 1.00 | 0.10 | 0.41 | 0.55 | 0.35 |
| **P2** | 0.10 | 1.00 | 0.64 | 0.47 | 0.98 |
| **P3** | 0.41 | 0.64 | 1.00 | 0.44 | 0.85 |
| **P4** | 0.55 | 0.47 | 0.44 | 1.00 | 0.76 |
| **P5** | 0.35 | 0.98 | 0.85 | 0.76 | 1.00 |

To calculate the MIN, select the smallest points and merge the points.

Since 0.10 is smallest merge points p1 & p2, after merging the matrix will be as follows,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | p1Up2 | P3 | P4 | P5 |
| **p1Up2** | 0.10 | 0.41 | 0.47 | 0.35 |
| **P3** | 0.41 | 1.00 | 0.44 | 0.85 |
| **P4** | 0.47 | 0.44 | 1.00 | 0.76 |
| **P5** | 0.35 | 0.85 | 0.76 | 1.00 |

Since 0.35 is smallest merge points p1, p2, p5, after merging the matrix will be as follows,

|  |  |  |  |
| --- | --- | --- | --- |
|  | p1Up2Up5 | P3 | P4 |
| **p1Up2Up5** | 0.10 | 0.41 | 0.47 |
| **P3** | 0.41 | 1.00 | 0.44 |
| **P5** | 0.47 | 0.44 | 1.00 |

Since 0.41 is smaller merge points p1, p2, p5, p3, after merging the matrix will be as follows,

|  |  |  |
| --- | --- | --- |
|  | p1Up2Up5Up3 | P4 |
| **p1Up2Up5Up3** | 0.10 | 0.44 |
| **P4** | 0.44 | 1.00 |

Since 0.44 is smaller merge points p1, p2, p5, p3, p4, after merging the matrix will be as follows,

|  |  |
| --- | --- |
|  | p1Up2Up5Up3Up4 |
| **p1Up2Up5Up3Up4** | **0.10** |

Chart, box and whisker chart

Description automatically generated

Dendrogram of MIN

(b) Use AVERAGE to calculate cluster distances.

At each step show the updated matrix with distances between clusters. Show your results by drawing the final dendrogram.

Ans.)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | P1 | P2 | P3 | P4 | P5 |
| **P1** | 1.00 | 0.10 | 0.41 | 0.55 | 0.35 |
| **P2** | 0.10 | 1.00 | 0.64 | 0.47 | 0.98 |
| **P3** | 0.41 | 0.64 | 1.00 | 0.44 | 0.85 |
| **P4** | 0.55 | 0.47 | 0.44 | 1.00 | 0.76 |
| **P5** | 0.35 | 0.98 | 0.85 | 0.76 | 1.00 |

To calculate the AVERAGE, select the pair of two points and calculate the average between the points and merge the points accordingly.

Since 0.10 is smaller, then calculate the average between points and merge the points p1 & p2, then the matrix will be as follows,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | p1Up2 | P3 | P4 | P5 |
| **p1Up2** | 0.55 | 0.52 | 0.51 | 0.66 |
| **P3** | 0.52 | 1.00 | 0.44 | 0.85 |
| **P4** | 0.51 | 0.44 | 1.00 | 0.76 |
| **P5** | 0.66 | 0.85 | 0.76 | 1.00 |

Since 0.44 is smaller, then calculate the average between the points and merge the points p3 & p4, then the matrix will be as follows,

|  |  |  |  |
| --- | --- | --- | --- |
|  | p1Up2 | p3Up4 | P5 |
| **p1Up2** | 0.55 | 0.52 | 0.66 |
| **P3Up4** | 0.52 | 0.72 | 0.80 |
| **P5** | 0.66 | 0.80 | 1.00 |

Since 0.52 is smaller, then calculate the average between the points and merge the points p1, p2, p3, p4, then the matrix will be as follows,

|  |  |  |
| --- | --- | --- |
|  | p1Up2Up3Up4 | P5 |
| **p1Up2Up3Up4** | 0.57 | 0.73 |
| **P5** | 0.73 | 1.00 |

Since 0.73 is smaller, then calculate the average between the points and merge the points p1, p2, p3, p4, p5, then the matrix will be as follows,

|  |  |
| --- | --- |
|  | P1Up2Up3Up4Up5 |
| **P1Up2Up3Up4Up5** | 0.76 |

Chart, box and whisker chart

Description automatically generated

Dendrogram of AVERAGE

Q3.) **MST-based clustering method**

Given a data set with the following points, please manually run the Minimum Spanning Tree (MST) based

clustering algorithm.

x1=(1, 1), x2=(5, 5), x3 =(5, 6), x4=(7, 8), x5=(8, 6), x6=(8, 7)

(a) Draw the minimum spanning tree you get from the dataset (i.e., after finish running Step 1 of

the algorithm). Use Euclidean distance for distance calculation. Succinctly explain your calculation.

Ans.) select any random point, lets take x2=(5, 5), now find the distance between x2 and the remaining points,

(x2,x1) = 5.65, (x2,x3) = 1, (x2,x4) = 3.60, (x2,x5) = 3.16, (x2,x6) = 3.60.

Now that the shortest path is 1, i.e., the closest pair of points (x2,x3) such that we should find the distance between x3 and the remaining points,

(x3,x1) = 6.40, (x3,x2) = 1, (x3,x4) = 2.82, (x3,x5) = 3, (x3,x6) = 3.16.

Now that the shortest path is 2.82, i.e., the closest pair of points (x3,x4) such that we should find the distance between x4 and the remaining points,

(x4,x1) = 9.21, (x4,x2) = 3.60, (x4,x3) = 2.82, (x4,x5) = 2.23, (x4,x6) = 1.41.

Now that the shortest path is 1.41, i.e., the closest pair of points (x4,x6) such that we should find the distance between x6 and the remaining points,

(x6,x1) = 9.21, (x6,x2) = 3.60, (x6,x3) = 3.16, (x6,x4) = 1.41, (x6,x5) = 1.

Now that the shortest path is 1, i.e., the closest pair of points (x6,x5) such that we should find the distance between x5 and the remaining points,

(x5,x1) = 8.60, (x5,x2) = 3.16, (x5,x3) = 3.00, (x5,x4) = 2.23, (x5,x6) = 1.

Chart, scatter chart

Description automatically generated

Minimum Spanning Tree

(b) What are the clusters after you run two iterations of Step 3 of the algorithm? Please explain

succinctly.

Ans.) According to the step-3 of the algorithm, the largest distance link should be breaking by creating a new cluster. So, the largest distance link is breaking to form new cluster in the first iteration, they are:

x1, (x2,x3).

In the second iteration, the largest distance link is broken further until a singleton cluster remains. Hence the following clusters are obtained after breaking the link, they are:

x1, (x2,x3), (x4,x5,x6).