Date - 11-01-2024

K-Medoid Algorithm

- The k-medoid algorithm is a clustering algorithm that is an extension of the more well known k-means algorithm.
- Unlike k-means, which uses the mean (centroid) of a cluster to present it, k-medoid uses the actual data point within a cluster that minimizes the dissimilarity to other points in the cluster.
- This data point is called the "medoid"

Here are the basic steps of the k-Medoid Algorithm

1. Initialization:
   Select k initial data point as the initial medoids

2. Assignment:
   assign each data point to the nearest medoid based on a dissimilarity measure (commonly, it could be euclidean distance, manhattan distance, or any other appropriate metric).

3. Update Medoids:
   - For each cluster, calculate the total dissimilarity of each point to the other points in the cluster.
   - Select the data point with the lowest total dissimilarity as the new medoid for that cluster.

4. Repeat:
Repeat the assignment and update steps until convergence (no or minimal changes in cluster assignments)

5. output:
The final clusters and their respective medoids.

- K-medoid is more robust to outliers since it uses actual data points as medoids.
- it can be computationally more expensive than K-Means, especially when calculating dissimilarity for all pairs of points in the cluster during the medoid update step.

solved example:
Apply K-Medoid clustering algorithm to form two clusters.

Here manhattan distance to find the distance between data point and medoid.

i	x	y	c_1	c_2	Cluster	d	cluster
x_1	2	6	3	7	c_1	8	c_1
x_2	3	4	0	4	c_1	5	c_1
x_3	3	8	4	8	c_1	9	c_1
x_4	4	7	4	6	c_1	7	c_1
x_5	6	2	5	3	c_2	2	c_2
x_6	6	4	3	1	c_2	2	c_2
x_7	7	3	5	1	c_2	0	c_2
x_8	7	4	4	0	c_2	1	c_2
x_9	8	5	6	2	c_2	3	c_2
x_{10}	7	6	6	2	c_2	3	c_2

Step-1: Select any two medoids and find distance

c1 (3,4) c2 = (7,4)

manhattan distance = $|x_1 - x_2| + |y_1 - y_2|$

Mdist [(2,6), (3,4)] = |2-3| + |6-4| = 3

Mdist [(3,4), (3,4)] = |3-3| + |4-4| = 0

similary for other data point.

Now compare c_1 and c_2 distances for each data point and from decide the that point lies on which cluster c1 or c_2.

Step 2

Hence. clusters are

c1: {(2,6), (3,4), (3,8), (4,7)}

c2: {(6,2), (6,4), (7,3), (7,4), (8,5), (7,6)}

first calculate the individual cost then calculate the total cost:

$cost(c, x) = \sum_i |c_i - x_i|$

cardinality of 3-3 and 4-8

total cost = {cost((3,4), (2,6)) + cost((3,4), (3,8)) + cost((3,4), (4,7)) + cost((7,4), (6,2)) + cost((7,4), (6,4)) + cost((7,4), (7,3)) + cost((7,4), (8,5)) + cost((7,4), (7,6))}

= 3+4+4+2+3+1+1+2 = 20

Step.3 Randomly select one non-medoid point and recalculate the cost.

c_1 = (3,4) and c2 = (7,4)

O = (7,3)

Here swap c2 with O

new medoids

$c_1 = (3,4)$ and $O = (7,3)$

Hence we manhattan distance = $|x_1 - x_2| + |y_1 - y_2|$

similarly find manhattan distance that is we calculate in step-1 but with our new medoids.

Here New clusters are

$c_1 = \{(2,6), (3,4), (3,8), (4,7)\}$

$O = \{(6,2), (6,4), (7,3), (7,4), (8,5), (7,6)\}$

Now similar step-2 Here also calculate the costs. (total)

Total cost = $3+4+4+2+2+1+3+3 = 22$

Step-4

cost of swapping of medoid c_2 with O

S = Current Total cost - Previous Total cost

$S = 22 - 20 > 0$

Here swapping c_2 with O is not good idea

Hence final medoids are $c_1 = (3,4)$ and $c_2 = (7,4)$