National Textile University

NTU

INNOVATE AND LEAD

ASSIGNMENT NO# 04

Title:

K-Mean, K-Mode, K-Medoid

Submitted To:

Respected, Dr. Haseeb Ahmad

Submitted By:

Muhammad Azam	18-NTU-CS-1178
Muhammad Furqan Haider	18-NTU-CS-1183
Usama Sadiq	18-NTU-CS-1191

K-Modes

K-Modes clustering is one of the unsupervised Machine Learning algorithms that is used to cluster categorical variables.

KMeans uses mathematical measures (distance) to cluster continuous data. The lesser the distance, the more similar our data points are. Centroids are updated by Means.
But for categorical data points, we cannot calculate the distance. So, we go for KModes algorithm. It uses the dissimilarities (total mismatches) between the data points. The lesser the dissimilarities the more similar our data points are. It uses Modes instead of means.

Example

Student	Hair Color	Brand of shampoo	Gender
S1	Black	Sunsilk	Female
S2	White	Clear	Male
S3	Blonde	Dove	Female
S4	Black	Head & Shoulders	Male
S5	Blonde	Clear	Male
S6	Black	Dove	Female

Alright, we have the sample data now. Let us proceed by defining the number of clusters(K)=2

Step 1: Pick K observations at random and use them as leaders/clusters

I am choosing S2, S6 as leaders/clusters

<table>
<tr><th colspan="4">Leaders Clusters</th></tr>
<tr><th>Cluster</th><th>Hair Color</th><th>Brand of shampoo</th><th>Gender</th></tr>
<tr><td>Cluster 1(S2)</td><td>White</td><td>Clear</td><td>Male</td></tr>
<tr><td>Cluster 2(S6)</td><td>Black</td><td>Dove</td><td>Female</td></tr>
<tr><th colspan="4">Observations</th></tr>
<tr><th>Student</th><th>Hair Color</th><th>Brand of shampoo</th><th>Gender</th></tr>
<tr><td>S1</td><td>Black</td><td>Sunsilk</td><td>Female</td></tr>
<tr><td>S2</td><td>White</td><td>Clear</td><td>Male</td></tr>
<tr><td>S3</td><td>Blonde</td><td>Dove</td><td>Female</td></tr>
<tr><td>S4</td><td>Black</td><td>Head & Shoulders</td><td>Male</td></tr>
<tr><td>S5</td><td>Blonde</td><td>Clear</td><td>Male</td></tr>
<tr><td>S6</td><td>Black</td><td>Dove</td><td>Female</td></tr>
</table>

Step 2: Calculate the dissimilarities (no. of mismatches) and assign each observation to its closest cluster

Iteratively compare the cluster data points to each of the observations. Similar data points give 0, dissimilar data points give 1.

Student	Cluster 1(S2)	Cluster 2(S6)	Cluster
S1	3	1	Cluster 2(S6)
S2	0	3	Cluster 1(S2)
S3	3	1	Cluster 2(S6)
S4	2	2	Cluster 1(S2)
S5	1	3	Cluster 1(S2)
S6	3	0	Cluster 2(s6)

After step 2, the observations S2, S4, S5 are assigned to cluster 1; S1, S3,S6 are assigned to Cluster 2.

Step 3: Define new modes for the clusters

Student	Hair Color	Brand of shampoo	Gender	Cluster
S1	Black	Sunsilk	Female	Cluster 2
S2	White	Clear	Male	Cluster 1
S3	Blonde	Dove	Female	Cluster 2
S4	Black	Head & Shoulders	Male	Cluster 1
S5	Blonde	Clear	Male	Cluster 1
S6	Black	Dove	Female	Cluster 2

Mode is simply the most observed value.

Cluster 1 observations (S2, S4, S5) has blonde as the most observed hair color, Clear as the most observed brand of shampoo, and Male as the most observed gender.

Same in case of cluster 2

Below are our new leaders after the update.

New Leaders			
Cluster	Hair Color	Brand of shampoo	Gender
Cluster 1	Blonde	Clear	Male
Cluster 2	Black	Dove	Female

Repeat steps 1–3

After obtaining the new leaders, again calculate the dissimilarities between the observations and the newly obtained leaders.

<table>
<tr><th colspan="4">New Leaders</th></tr>
<tr><th>Cluster</th><th>Hair Color</th><th>Brand of shampoo</th><th>Gender</th></tr>
<tr><td>Cluster 1</td><td>Blonde</td><td>Clear</td><td>Male</td></tr>
<tr><td>Cluster 2</td><td>Black</td><td>Dove</td><td>Female</td></tr>
<tr><th colspan="4">Observations</th></tr>
<tr><th>Student</th><th>Hair Color</th><th>Brand of shampoo</th><th>Gender</th></tr>
<tr><td>S1</td><td>Black</td><td>Sunsilk</td><td>Female</td></tr>
<tr><td>S2</td><td>White</td><td>Clear</td><td>Male</td></tr>
<tr><td>S3</td><td>Blonde</td><td>Dove</td><td>Female</td></tr>
<tr><td>S4</td><td>Black</td><td>Head & Shoulders</td><td>Male</td></tr>
<tr><td>S5</td><td>Blonde</td><td>Clear</td><td>Male</td></tr>
<tr><td>S6</td><td>Black</td><td>Dove</td><td>Female</td></tr>
</table>

Student	Cluster 1	Cluster 2	Cluster
S1	3	1	Cluster 2
S2	1	3	Cluster 1
S3	1	1	Cluster 2
S4	2	2	Cluster 1
S5	0	3	Cluster 1
S6	3	0	Cluster 2

The observations S2, S4, S5 are assigned to cluster 1; S1, S3, S6 are assigned to Cluster 2.

We stop here as we see there is no change in the assignment of observations.

K-Medoids

K-Medoids (also called as Partitioning Around Medoid) algorithm was proposed in 1987 by Kaufman and Rousseeuw. A medoid can be defined as the point in the cluster, whose dissimilarities with all the other points in the cluster is minimum.

Example

Points	X	Y
1	6	7
2	3	4
3	4	8
4	7	5
5	5	4
6	2	3

Step 1:

Let the randomly selected 2 medoids, so select k = 2 and let C1 = (4, 8) and C2 = (2, 3) are the two medoids.

Step 2: Calculating cost.

The dissimilarity of each non-medoid point with the medoids is calculated and tabulated:

Points	X	Y	Dissimilarity from C1	Dissimilarity from C2
1	6	7	3	8
2	3	4	5	2
3	4	8		
4	7	5	6	7
5	5	4	5	4
6	2	3		

Each point is assigned to the cluster of that medoid whose dissimilarity is less.
The points 1, 4 go to cluster C1 and 2, 5 go to cluster C2.
The Cost = (3 + 6) + (2 + 4) = 15

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

Let the randomly selected point be (5, 4). The dissimilarity of each non-medoid point with the medoids C1 (4, 8) and C2 (5, 4) is calculated and tabulated.

Points	**X**	**Y**	**Dissimilarity from C1**	**Dissimilarity from C2**
1	6	7	3	4
2	3	4	5	2
3	4	8		
4	7	5	6	3
5	5	4		
6	2	3	7	3

Each point is assigned to the cluster of that medoid whose dissimilarity is less.
The points 1 go to cluster C1 and 2, 4,6 go to cluster C2.
The new Cost = (3) + (2 + 3+3) = 11

Swap Cost = New Cost – Previous Cost = 11 – 15 and -4 <0

As the swap cost is less than zero, then we swap the clusters. Hence (4, 8) and (5, 4) are the final medoids.

Advantages:

1. It is simple to understand and easy to implement.
2. K-Medoid Algorithm is fast and converges in a fixed number of steps.
3. PAM is less sensitive to outliers than other partitioning algorithms.

Disadvantages:

1. The main disadvantage of K-Medoid algorithms is that it is not suitable for clustering non-spherical (arbitrary shaped) groups of objects. This is because it relies on minimizing the distances between the non-medoid objects and the medoid (the cluster Centre) – briefly, it uses compactness as clustering criteria instead of connectivity.
2. It may obtain different results for different runs on the same dataset because the first k medoids are chosen randomly.

K-Means

K-Means Clustering is an Unsupervised Learning algorithm, used to group the unlabeled dataset into different clusters/subsets.

Steps in K-Means:

step1: choose k value for ex: k=3

step2: initialize centroids randomly

step3: calculate Euclidean distance from centroids to each data point and form clusters that are close to centroids

step4: find the centroid of each cluster and update centroids

step:5 repeat step3

Example

Cluster the following eight points (with (x, y) representing locations) into three clusters:

A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).

The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as-

P(a, b) = |x2 – x1| + |y2 – y1|

Use K-Means Algorithm to find the three cluster centers after the second iteration.

Solution

We follow the above discussed K-Means Clustering Algorithm-

Iteration-01:

We calculate the distance of each point from each of the center of the three clusters.

The distance is calculated by using the given distance function.

The following illustration shows the calculation of distance between point A1(2, 10) and each of the center of the three clusters-

Calculating Distance Between A1(2, 10) and C1(2, 10)-

P (A1, C1)

= |x2 – x1| + |y2 – y1|

= |2 – 2| + |10 – 10|

= 0

Calculating Distance Between A1(2, 10) and C2(5, 8)-

P (A1, C2)

= |x2 – x1| + |y2 – y1|

= |5 – 2| + |8 – 10|

= 3 + 2

= 5

Calculating Distance Between A1(2, 10) and C3(1, 2)-

P (A1, C3)

= |x2 – x1| + |y2 – y1|

= |1 – 2| + |2 – 10|

= 1 + 8

= 9

In the similar manner, we calculate the distance of other points from each of the center of the three clusters.

Next,

We draw a table showing all the results.

Using the table, we decide which point belongs to which cluster.

The given point belongs to that cluster whose center is nearest to it.

Given Points	**Distance from center (2, 10) of Cluster-01**	**Distance from center (5, 8) of Cluster-02**	**Distance from center (1, 2) of Cluster-03**	**Point belongs to Cluster**
A1(2, 10)	0	5	9	C1
A2(2, 5)	5	6	4	C3
A3(8, 4)	12	7	9	C2
A4(5, 8)	5	0	10	C2
A5(7, 5)	10	5	9	C2
A6(6, 4)	10	5	7	C2
A7(1, 2)	9	10	0	C3
A8(4, 9)	3	2	10	C2

From here, new clusters are-

Cluster-01:

First cluster contains points-

A1(2, 10)

Cluster-02:

Second cluster contains points-

A3(8, 4)

A4(5, 8)

A5(7, 5)

A6(6, 4)

A8(4, 9)

Cluster-03:

Third cluster contains points-

A2(2, 5)

A7(1, 2)

Now,

We re-compute the new cluster clusters.

The new cluster center is computed by taking mean of all the points contained in that cluster.

For Cluster-01:

We have only one point A1(2, 10) in Cluster-01.

So, cluster center remains the same.

For Cluster-02:

Center of Cluster-02

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

$$= (6, 6)$$

For Cluster-03:

Center of Cluster-03

$$= ((2 + 1)/2, (5 + 2)/2)$$

$$= (1.5, 3.5)$$

This is completion of Iteration-01.

Iteration-02:

We calculate the distance of each point from each of the center of the three clusters.

The distance is calculated by using the given distance function.

The following illustration shows the calculation of distance between point A1(2, 10) and each of the center of the three clusters-

Calculating Distance Between A1(2, 10) and C1(2, 10)-

$$\begin{gathered} P\ (A1, C1) \\ = |x2 - x1| + |y2 - y1| \\ = |2 - 2| + |10 - 10| \\ = 0 \end{gathered}$$

Calculating Distance Between A1(2, 10) and C2(6, 6)-

$$\begin{gathered} P\ (A1, C2) \\ = |x2 - x1| + |y2 - y1| \\ = |6 - 2| + |6 - 10| \\ = 4 + 4 \\ = 8 \end{gathered}$$

Calculating Distance Between A1(2, 10) and C3(1.5, 3.5)-

$$\begin{gathered} P\ (A1, C3) \\ = |x2 - x1| + |y2 - y1| \\ = |1.5 - 2| + |3.5 - 10| \\ = 0.5 + 6.5 \\ = 7 \end{gathered}$$

In the similar manner, we calculate the distance of other points from each of the center of the three clusters.

Next,

We draw a table showing all the results.

Using the table, we decide which point belongs to which cluster.

The given point belongs to that cluster whose center is nearest to it.

Given Points	**Distance from center (2, 10) of Cluster-01**	**Distance from center (6, 6) of Cluster-02**	**Distance from center (1.5, 3.5) of Cluster-03**	**Point belongs to Cluster**
A1(2, 10)	0	8	7	C1
A2(2, 5)	5	5	2	C3
A3(8, 4)	12	4	7	C2
A4(5, 8)	5	3	8	C2
A5(7, 5)	10	2	7	C2
A6(6, 4)	10	2	5	C2
A7(1, 2)	9	9	2	C3
A8(4, 9)	3	5	8	C1

From here, new clusters are-

Cluster-01:

First cluster contains points-

A1(2, 10)

A8(4, 9)

Cluster-02:

Second cluster contains points-

A3(8, 4)

A4(5, 8)

A5(7, 5)

A6(6, 4)

Cluster-03:

Third cluster contains points-

A2(2, 5)

A7(1, 2)

Now,

We re-compute the new cluster clusters.

The new cluster center is computed by taking mean of all the points contained in that cluster.

For Cluster-01:

Center of Cluster-01

= ((2 + 4)/2, (10 + 9)/2)

= (3, 9.5)

For Cluster-02:

Center of Cluster-02

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

= (6.5, 5.25)

For Cluster-03:

Center of Cluster-03

= ((2 + 1)/2, (5 + 2)/2)

= (1.5, 3.5)

This is completion of Iteration-02.

After second iteration, the center of the three clusters is-

C1(3, 9.5)

C2(6.5, 5.25)

C3(1.5, 3.5)

Advantages

- Easy to implement.
- With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small).
- k-Means may produce Higher clusters than hierarchical clustering.

Disadvantages

- Difficult to predict the number of clusters (K-Value).
- Initial seeds have a strong impact on the final results.