

**ASSIGNMENT# Data Science**

**Title:**

**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**

|  |  |  |  |
| --- | --- | --- | --- |
| **Leaders Clusters** | | | |
| **Cluster** | **Hair Color** | **Brand of shampoo** | **Gender** |
| Cluster 1(S2) | White | Clear | Male |
| Cluster 2(S6) | Black | Dove | Female |
| **Observations** | | | |
| **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 |

**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.

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

|  |  |  |  |
| --- | --- | --- | --- |
| **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**

|  |  |  |
| --- | --- | --- |
|  | **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:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **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.