K- MODES CLUSTERING

K-Modes Clustering(Introductions):

ML Labs Pvt Ltd

Whenever we talk about unsupervised learning algorithm, the term which pops our mind first is K-Means Clustering!!! However, we often forget that K-means clustering works efficiently only for "numerical dataset". We don't get proper results for the "categorical data" because of the improper spatial representation. K-Means Clustering fails to find patterns in the categorical dataset. Hence, comes in picture K-Modes Clustering.

K-Modes Clustering(Introductions):

ML Labs Pvt Ltd

K-means clustering can’t handle non-numerical (categorical) data. But we can map categorical value to 1/0. However, this mapping can’t generate quality clusters for high-dimensional data. Then people requesting the K-Modes method by replacing the means of the clusters with modes, which is called k-modes clustering.

K-Modes Clustering:

ML Labs Pvt Ltd

Most of the real-world datasets are in categorical form. Let's say, if we are working on analyzing the social media, we have categorical data like gender (male or female), profession and so on. So, deal with all this categorical data or cluster the categorical variables we use K Modes Clustering.

K-Modes Clustering:

ML Labs Pvt Ltd

It is widely used algorithm for grouping the categorical data because it is easy to implement and efficiently handles large amount of data. . It defines clusters based on the number of matching categories between data points. (This contrasts with the more well-known k-means algorithm, which clusters numerical data based on Euclidean distance.)

How is it used???

ML Labs Pvt Ltd

The k-modes clustering algorithm is an extension of k-means clustering algorithm. The k-means algorithm is the most widely used Centre based partitional clustering algorithm. Huang extends the k-means clustering algorithm to k-modes clustering algorithm to group the categorical data.

How is it used???

ML Labs Pvt Ltd

The modifications done in the k-means are -

(i) using a simple matching dissimilarity measure for categorical objects,

(ii) replacing means of clusters by modes, and

(iii) using a frequency-based method to update the modes.

How is it used???

ML Labs Pvt Ltd

Statistics behind!!

Let X, x11, x12,…,xnm be the data set consists of n number of objects with m number of attributes.

The main objective of the k-modes clustering algorithm is to group the data objects X into K-clusters by minimize the cost function Eq.(1) below.

How is it used???

$$P(W,Q)=\sum_{l=1}^{k}\sum_{i=1}^{n}w_{il}d_{sim}(x_i,q_l) \quad (1)$$

where, w_{il} is an $N\times K$ matrix where each element belongs to 0 or 1. N is the total number data objects and K is the number of clusters. $d_{sim}(x_i,q_l)$ is the simple dissimilarity measure and it is defined in the following Eq.(2).

$$d_{sim}(x_i,q_i)=\sum_{j=1}^{m}\delta(x_{ij},z_{lj}) \quad (2)$$

where, $\delta(x_{ij}, q_{lj})$ is calculated using the following Eq.(3)

$$\delta(x_{ij},z_{lj})=\begin{cases}1 & if\ x_{ij}=z_{lj}\\ 0 & if\ x_{ij}\neq z_{lj}\end{cases} \quad (3)$$

HOW THE K-MODE CLUSTERING ALGORITHM WORKS?

ML Labs Pvt Ltd

The k-modes clustering algorithm is described as,

Input: Data objects X, Number of clusters K.

Step 1: Randomly select the K initial modes from the data objects such that Cj, j = 1,2,…,K

Step 2: Find the matching dissimilarity between the each K initial cluster modes and each data objects using the Eq.(2).

HOW THE K-MODE CLUSTERING ALGORITHM WORKS?

ML Labs Pvt Ltd

Step 3: Evaluate the fitness using the Eq.(1)

Step 4: Find the minimum mode values in each data object i.e., finding the objects nearest to the initial cluster modes.

Step 5: Assign the data objects to the nearest cluster centroid modes.

HOW THE K-MODE CLUSTERING ALGORITHM WORKS?

ML Labs Pvt Ltd

Step 6: Update the modes by apply the frequency-based method on newly formed clusters.

Step 7: Recalculate the similarity between the data objects and the updated modes.

Step 8: Repeat the step 4 and step 5 until no changes in the cluster ship of data objects. Output: Clustered data objects

K-Modes: Clustering Categorical Data

ML Labs Pvt Ltd

- K-Means cannot handle non-numerical (categorical) data
- Mapping categorical value to 1/0 cannot generate quality clusters for high-dimensional data
- K-Modes: An extension to K-Means by replacing means of clusters with modes
- Dissimilarity measure between object X and the center of a cluster Z

K-Modes: Clustering Categorical Data

ML Labs Pvt Ltd

- This dissimilarity measure (distance function) is frequency-based
- Algorithm is still based on iterative object cluster assignment and centroid update.
- A fuzzy K-Modes method is proposed to calculate a fuzzy cluster membership value for each object to each cluster
- mixture of categorical and numerical data: Using a K-Prototype method