K-Means Algorithm

Definition:

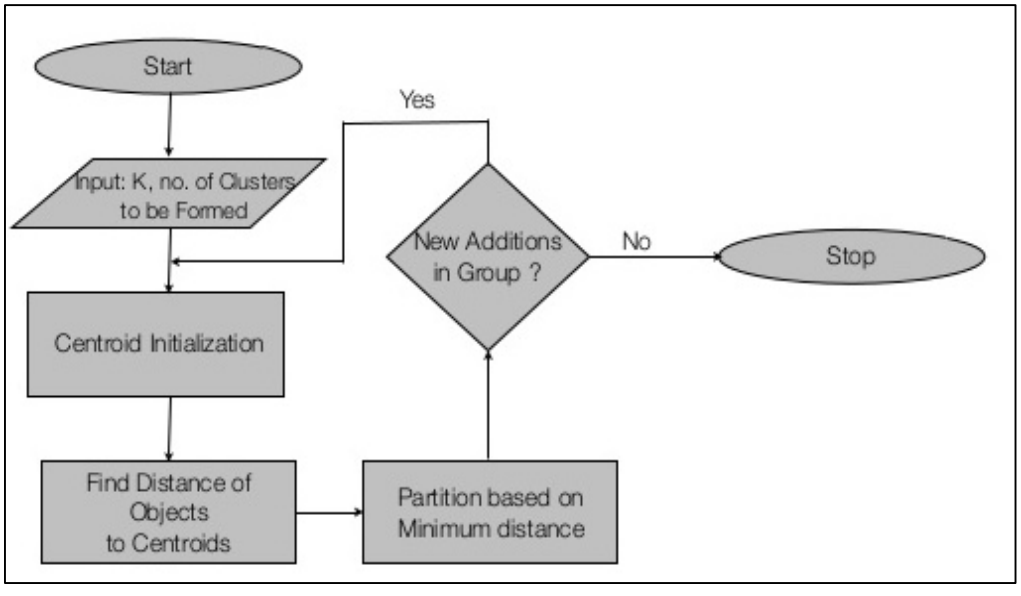
K-Means Algorithm is a **centroid-based** partitioning technique that uses the centroid of a cluster to represent that cluster. The number of data objects inside the cluster is computed on the bases of a distance measure (dissimilarity measure). K-Means minimizes aggregate intra-cluster distance.

Input Parameters to the Algorithm:

k – represents the number of clusters to formed from the dataset.
Distance metric – Euclidean Distance, Manhattan Distance, etc.

Process Flow of K-Means Algorithm:

- In the very first iteration, k data objects (seeds) are randomly selected. These are the **cluster centers**.
- Using the distance measure, the distance between each data object and the cluster centers is calculated.
- The data objects are then grouped based on the **minimum distance**.
- This is the end of the first iteration. At this point, the initial version of grouping has been done.
- The k -means algorithm then iteratively improves the within-cluster variation. For each cluster, it computes a new mean (and thus new cluster centers) using the objects assigned to the cluster in the previous iteration.
- The iterations continue until the assignment is stable, that is, the clusters formed in the current round are the same as those formed in the previous round. In the last iteration, the distance of each data point from the center does not change.
- The process of iteratively reassigning objects to clusters to improve the partitioning is referred to as **iterative relocation**. Eventually, no reassignment of the objects in any cluster occurs and so the process terminates.
- Finally, the resulting clusters are returned by the clustering process.



Disadvantages:

- K-Means is sensitive to noise and outlier data points. A small number of such data points can substantially influence the mean value. In order to control the sensitivity to outliers, instead of taking the mean value of the data objects, picking an actual value that represents the cluster can be beneficial.

- The necessity for users to specify k, the number of clusters, in advance can be seen as a disadvantage. Analytical techniques to determine the best k values by comparing results obtained for different k values have to be used.

- The K-Means method can be applied only when the mean of a set of objects is defined. This may not be the case in some applications such as when data is categorical and not numeric. Here, the **K-Modes** method, a variant of the K-Means method which extends the K-Means paradigm to cluster categorical data by replacing means of clusters with modes, is used.

Measurement of Error:

The quality of cluster can be measured by the within cluster variation, which is the sum of squared error between all objects in the cluster and the centroid (cluster center). In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This error value is then averaged to find the **Mean Square Error**.

Notes:

- If the distance is equal from both the cluster centers, the data point is randomly assign to any cluster.
- The size of the clusters does not need to be the same.
- There is no guarantee that nearby points end up in the same cluster.
- The clusters are separable in the way that mean value converges towards the cluster center.
- K-Means and K-Modes methods can be integrated to cluster data with mixed numeric and categorical values.
- In order to detect outliers, after clustering, for each cluster and for each data object, make a measure of the variance. Find out the cluster having the maximum variance and the cluster having the second highest variance. If the gap between these variances is substantial, then the cluster with the maximum variance contains the outlier. In this case, the cluster containing the outlier can be further partitioned according to the mean of the cluster. Another method is to find the diameter of a cluster (difference between the smallest value and the largest value in the cluster) and check if any of the diameters is substantially greater than the rest. These are called split and merge techniques.

K-Means Algorithm implementation in R:
https://github.com/DataAstrologer/DataScience/blob/master/R/Machine%20Learning%20Algorithm%20Templates/Unsupervised%20Learning/Clustering/K-Means.Rmd

K-Means Algorithm implementation in Python:
https://github.com/DataAstrologer/DataScience/blob/master/Python/Machine%20Learning%20Algorithm%20Templates/Unsupervised/Clustering/K-Means.ipynb