**K- MEANS CLUSTERING ALGORITHM**

The dataset contains several variables of different types. Sometimes we will be in need of grouping variables based on some criteria. Classification and clustering are the widely used methodology across all machine learning algorithms. Classification works for supervised learning where we have a dependent variable whereas Clustering works for unsupervised learning where we don’t have any knowledge of dependent variable.

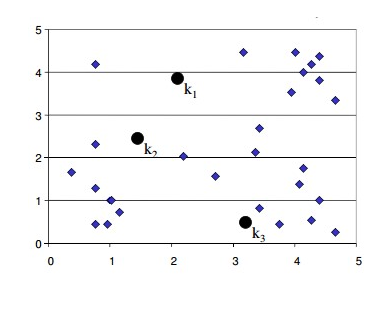
Clustering is used to group data points or variables into groups of similar types and maintains great distance between each group. K-Means clustering is used to cluster data points into k-clusters maintaining intergroup similarity and there will be more distance between each groups thereby avoiding mixing of data points among each clusters. K-means works only for data that are of continuous in nature. Hence categorical data points have to be converted to continuous type before applying the algorithm. And the data point units should be similar. For example if we have weight variable in KG and height variable in CM, those units have to be standardized before applying to the algorithm.

K-means algorithm uses Euclidean Distance for calculating the distance between two data points in two dimensional spaces. The distance formula is as follows:

Euclidean Distance =  Formula-1

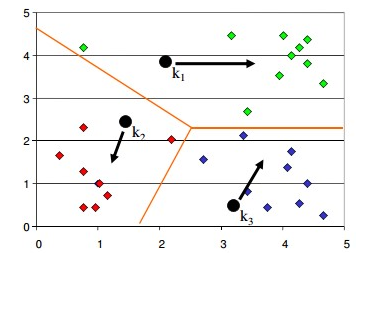
The same formula can be used for multi-dimensional spaces too by adding extra terms to the equation. In K-means, K represents the number of clusters that has to be grouped to the given data point. The K-means algorithm works as follows:

**Step 1:** Starts by assigning the K value say number of clusters. After that assigns three center points to the given data space and assigns the data points that are close to those centers.

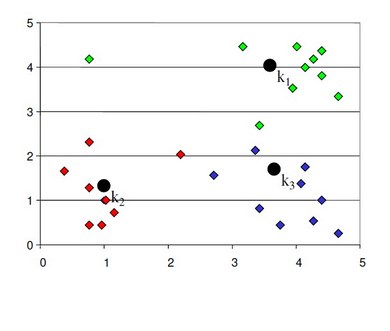


The above example assigns K as 3 hence; it divides the data space into three centers K1, K2, K3.

**Step 2:** In this step, the algorithm calculates the mean from the assigned centers and moves the centers to the calculated means.



**Step 3:** Now the algorithm assigns the data points that are closer to the newly created centers.



**Step 4:** Step 2 and 3 are repeated until no members changes its association.

Hence K-Means is an iterative process of clustering; which keeps iterating until it reaches the best solution or clusters in our problem space.