**SEMINAR TOPIC:K-means clustering**

**Abstract:**

K-means (MacQueen, 1967)is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group page is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Clustering is used to identify and classify into groups of similar objects in a multivariate data sets. It locates the centroid of the group of data points in the dataset to classify.

When a data set of items, with certain features and value is to be categorized, we can use k-means algorithm. This unsupervised learning algorithm will categorize the items into k similar groups. To calculate that similarity, we will use the Euclidean distance as measurement.

Advantage of K-means clustering is easy implementation with computationally fast and efficient with large number of variables and It works well with distinct boundary data sets. The disadvantage of K-means is Difficulty in predicting the exact k-value for unknown data set and Initial seeds have a strong influence on the final resulting cluster.