**Experiment 3**

**Aim:**

To study and implement K-means Clustering

**Objective:**

To perform classification of given unlabelled data into ‘*K*’ number of clusters

**Software platform:**

Matlab

**Theory:**

K-means clustering is a type of unsupervised learning, which is used when you have unlabelled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the *K*-means clustering algorithm are:

* The centroids of the *K* clusters, which can be used to label new data
* Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyse the groups that have formed organically. The value of ‘k’ needs to be chosen according to number of classes in which we want the data samples to be classified.

Each centroid of a cluster is a collection of feature vales which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

**Algorithm:**

1. Initialize the cluster centers and the data samples to be classified.
2. Find the Euclidian distance between every data sample and the cluster points.

**d= ( (x1-x2)2 + (y1-y2)2 )1/2**

1. Classify the data samples into different classes according to the distances computed.
2. Calculate the new cluster points according to the classes. Find average of x and y coordinates in the specific class to find the new cluster points.
3. Repeat steps 2 to 4 till the cluster points remain same as the cluster points in previous iteration.
4. Plot the Classified data showing different classes.
5. Extend the same code for large number of random data points.

**Conclusion:**

**FAQ’S:**

1. What do you mean by a cluster?
2. Solve the following problem using K-means Clustering.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X | 1 | 1.5 | 3 | 5 | 3.5 | 4.5 | 3.5 |
| Y | 1 | 2 | 4 | 7 | 5 | 5 | 4.5 |

Initial cluster centers are C1 = (1, 1), C2 = (5, 7).

1. Give advantages and limitations of K-means Clustering technique. Also state its applications.