K-means Clustering in Data Mining

- K-means clustering is simple unsupervised learning algorithm developed by **J. MacQueen in 1967 and then J.A Hartigan and M.A Wong in 1975.**

- In this approach, the data objects ('n') are classified into 'k' number of clusters in which each observation belongs to the cluster with nearest mean.

- It defines 'k' sets (the point may be considered as the center of a one or two dimensional figure), one for each cluster $k \leq n$. The clusters are placed far away from each other.

- Then, it organizes the data in appropriate data set and associates to the nearest set. If there is no data pending, first step is complicated to perform, in this case an early grouping is done. It is necessary to re-calculate 'k' new set as barycenters of the clusters from previous step.

- After having these 'k' new sets, the same data set points and the nearest new sets are bound together.

- Finally, a loop is generated. As a result of this loop, the 'k' sets change their location step by step until no more changes are made.

Solved examples of K-means:

Method 1:

Using K-means clustering, cluster the following data into two clusters and show each step.

  {2, 4, 10, 12, 3, 20, 30, 11, 25}

Solution:

Given: {2, 4, 10, 12, 3, 20, 30, 11, 25}

Step 1: Assign alternate value to each cluster randomly.

Step 2: k_1= {2, 10, 3, 30, 25}, Mean value= 14

   k_2= {4, 12, 20, 11, 25}, Mean value = 11.75

Step 3: Again assign the values,

   k_1 = {20, 30, 25}, Mean value = 25

   k_2 = {2, 4, 10, 12, 3, 11}, Mean value = 7

Step 4: Again assign the values,

   k_1 = {20, 30, 25}, Mean value = 25

   k_2 = {2, 4, 10, 12, 3, 11}, Mean value = 7

Method 2:

Step 1: Randomly assign the means: $m_1 = 3$, $m_2 = 4$

Step 2: Group the numbers close to mean $m_1 = 3$ are grouped into cluster k_1 and

$m_2 = 4$ are grouped into cluster k_2

Step 3: $k_1 = \{2, 3\}$, $k_2 = \{4, 10, 12, 20, 30, 11, 25\}$, $m_1 = 2.5$, $m_2 = 16$

Step 4: $k_2 = \{2, 3, 4\}$, $k_2 = \{10, 12, 20, 30, 11, 25\}$, $m_1 = 3$ $m_2 = 18$

Step 5: $k_1 = \{2, 3, 4, 10\}$, $k_2 = \{12, 20, 30, 11, 25\}$, $m_1 = 4.75$, $m_2 = 19.6$

Step 6: $k_1 = \{2, 3, 4, 10, 11, 12\}$, $k_2 = \{20, 30, 25\}$, $m_1 = 7$, $m_2 = 25$

Step 7: $k_1 = \{2, 3, 4, 10, 11, 12\}$, $k_2 = \{20, 30, 25\}$, $m_1 = 7$, $m_2 = 25$

Step 8: Stop. The clusters in step 6 and 7 are same.

Final answer: $k_1 = \{2, 3, 4, 10, s11, 12\}$ and $k_2 = \{20, 30, 25\}$