Page No.
Date

Assignment A4

Title: KMeans Clustering

Problem statement:

We have given a collection of 8 points $P_1 = [0.1, 0.6]$, $P_2 = [0.15, 0.71]$, $P_3 = [0.08, 0.9]$, $P_4 = [0.16, 0.85]$, $P_5 = [0.2, 0.3]$, $P_6 = [0.25, 0.5]$ $P_7 = [0.24, 0.1]$, $P_8 = [0.3, 0.2]$. Perform KMeans clustering with initial centroids as $m_1 = P_1 =$ cluster #1 $= c_1$ and $m_2 = P_8 =$ cluster #2 $= c_2$

Answer the following

1) Which cluster does P_6 belong to.
2) What is the population of cluster around m_2?
3) What is updated value of m_1 and m_2?

Objective: To understand how k-means clustering algorithm works on the given dataset.

Outcome:-

Students will be able to:
- implement k-means clustering algorithm.

Software and Hardware requirements

i3/i5/i7 64 bit processor os Linux 64 bit os
Editor - gedit / Eclipse.
Software - Jupyter Notebook / Python.

Page No.
Date

Theory:

K means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithm make inferences from datasets using only input vectors without referring to known or labelled outcomes. We define a target number k which refers to the number of centroids you need in dataset. A centroid is the imaginery or real location representing the centre of the cluster. Every data point is allocated to each of the clusters through reducing the in clusters sum of squares.

In other words, the K means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

Squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

where,

$\|x_i - v_j\|$ is the Euclidean distance between x_i and v_j.

Page No.
Date

Algorithmic steps for K-means clustering.

Let $x = \{x_1, x_2, x_3, \ldots x_n\}$ be the set of data points and $V = \{v_1 \ldots v_c\}$ be the set of centres.

1) Randomly select 'c' cluster centre.
2) Calculate the distance between each data point and cluster centre.
3) Assign the data point to the cluster centre whose distance from the cluster centre is minimum of all the cluster centres.
4) Recalculate the new cluster centre using.

$$v_i = \frac{1}{c_i} \sum_{j=1}^{c_i} x_i$$

where c_i represents the number of data points in ith cluster.

5) Recalculate the distance between each data point and new obtained cluster centres.
6) If no data point was reassigned then stop otherwise repeat from step 3.

Direct and Statistical testing Methods

1) Direct :- It consists of optimising a criterion, such as cluster sums of squares or the average. The corresponding methods are named elbow method and Silhouette method.

2) Statistical testing Methods:- Consists of comparing evidence against null hypothesis. An example is the gap statistic.

Advantages of K-Means clustering.

1) Fast, robust, and easier to understand.
2) Relatively efficient O(tkdn)
   n is # objects
   k is # of clusters
   d is # dimension of each object
   t is # iterations

Normally k, t, d << n

3) Gives best result when data set is distinct or well separated from each other.

Disadvantages

1) Random chosing of the cluster center cannot lead to fruitful result.

2) Applicable only when mean is defined i.e fails for categorical data.

3) Algorithm fails for non-linear dataset.

Result-

1) which cluster does P6 belong to - m_2
2) What is the population of cluster around m_2 ? ..., 3
3) What is the updated value of m_1 and m_2

[(0.148, 0.712), (0.246, 0.200)]

Conclusion :

Successfully implemented the k-means clustering algorithm for the given problem Statement.