Page :
Date : / /

ASSIGNMENT A4

Title : K-Means 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 k-means clustering with initial centroids as m1 = P1 = Cluster #1 = C1 and m2 = P8 = Cluster #2 = C2.

Answer the following
1. Which cluster does P6 belong to ?
2. What is the population of cluster around m2.
3. What is the updated value of m1 and m2.

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

Outcome : Successfully implemented k-means clustering algorithm

Requirements : python3, jupyter, pandas, numpy, sklearn

Theory :
K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known or labelled outcomes.

A target number k is defined, which refers to the number of centroids needed in the dataset. A centroid is the imaginary of

real location representing the center 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. The means in the k-means refers to the averaging of the data, i.e. finding the centroid.

Finally, this algorithm aims at minimizing an objective function known as squared error function given by:

Objective function $\leftarrow J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - c_j \|^2$, where

$x_i^{(j)}$: case i
c_j: centroid for cluster j
$\| x_i^{(j)} - c_j \|$: Distance function

k = no. of clusters
n = no. of cases

Algorithm

Let $X = \{x1, x2, x3, \ldots, x_n\}$ be the set of data points and $V = \{v1, v2, \ldots, v_c\}$ be the set of centers

1. Randomly select 'c' cluster centres
2. Repeat the following until the results of the iteration are new:
   (a) Calculate the distance between each data point and cluster centers.
   (b) Assign the data point to the cluster center whose distance from the cluster center is the minimum of the cluster centers.
   (c) Recalculate the new cluster using $v_i = (1/c_i) \sum_{i=1}^{c_i} x_i$, where c_i represents the number of points in the i^{th} cluster

(d) Recalculate the distance b/w each data point and the newly obtained cluster centers.

It halts creating and optimizing clusters when either the centroids have stabilized or the defined number of iterations has been achieved.

Since clustering algorithms including k-means use distance-based measurements to determine the similarity between data points, its recommended to standardize the data to have a mean of zero and a standard deviation of one since almost always the features in any dataset would have different units of measurements such as age vs income.

Given k-means iterative nature and the random initialization of centroids at the start of the algorithm, different initializations may lead to different clusters since k-means algorithm may be stuck in a local optimum, and may not converge to global optimum. Therefore, its recommended to run the algorithm using different initializations of centroids and pick the results of the run that yielded the lower sum of squared distance.

Test Cases

Test Case	Expected Output	Actual Output
Determine which cluster [0.2, 0.3] belongs to	1	1
Determine population of 2nd centroid's cluster	4	4
Centroid values after applying k-means clustering	[0.2475, 0.275] and [0.1225, 0.765]	[0.2475, 0.275] and [0.1225, 0.765]

Conclusion: Successfully performed k-means clustering on the given data points and calculated clusters for the same