

TITLE : Assignment on k Means Clustering

PROBLEM STATEMENT : We have given a collection of 8 points. $P1=[0.1,0.6]$, $P2=[0.15,0.71]$, $P3=[0.08,0.9]$ $P4=[0.16, 0.85]$, $P5=[0.2,0.3]$, $P6=[0.25,0.5]$, $P7=[0.24,0.1]$, $P8=[0.3,0.2]$. Perform the k-mean 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 updated value of m1 and m2?

OBJECTIVE :

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

S/W and H/W : Python, Colab, Core i5, 8 GB RAM, 1 TB HDD, Keyboard, mouse

Theory :

1. 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.
2. You'll define a target number k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster.
3. Every data point is allocated to each of the clusters through reducing the in-cluster 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.
4. K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group.
5. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum.
6. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

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

where,

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

' c_i ' is the number of data points in i th cluster.

' c ' is the number of cluster centers.

Algorithmic steps for k-means clustering

Let $X = \{x_1, x_2, x_3, \dots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \dots, v_c\}$ be the set of centers.

- 1) Randomly select ' c ' cluster centers.
- 2) Calculate the distance between each data point and cluster centers.
- 3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
- 4) Recalculate the new cluster center using:
- 5) Recalculate the distance between each data point and new obtained cluster centers.
- 6) If no data point was reassigned then stop, otherwise repeat from step 3).

7. It halts creating and optimizing clusters when either:

The centroids have stabilized — there is no change in their values because the clustering has been successful.

The defined number of iterations has been achieved.

The Elbow method:

The Elbow method looks at the total WSS as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS.

The optimal number of clusters can be defined as follow:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k . For instance, by varying k from 1 to 10 clusters.
2. For each k , calculate the total within-cluster sum of square (wss).
3. Plot the curve of wss according to the number of clusters k .
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

Note that, the elbow method is sometimes ambiguous. An alternative is the average silhouette method which can be also used with any clustering approach.

Disadvantages of Kmeans clustering algorithm:

- 1) The learning algorithm requires apriori specification of the number of cluster centers.
- 2) The use of Exclusive Assignment - If there are two highly overlapping data then k-means will not be able to resolve that there are two clusters.
- 3) The learning algorithm is not invariant to non-linear transformations i.e. with different representation of data we get different results (data represented in form of cartesian co-ordinates and polar co-ordinates will give different results).
- 4) Euclidean distance measures can unequally weight underlying factors.
- 5) The learning algorithm provides the local optima of the squared error function.
- 6) Randomly choosing of the cluster center cannot lead us to the fruitful result.
- 7) Applicable only when mean is defined i.e. fails for categorical data

Applications

k-means algorithm is very popular and used in a variety of applications such as market segmentation, document clustering, image segmentation and image compression, etc. The goal usually when we undergo a cluster analysis is either:

1. Get a meaningful intuition of the structure of the data we're dealing with.
2. Cluster-then-predict where different models will be built for different subgroups if we believe there is a wide variation in the behaviors of different subgroups. An example of that is clustering patients into different subgroups and build a model for each subgroup to predict the probability of the risk of having heart attack.

Conclusion : Successfully implemented kmeans clustering and answer to questions are

- 1) P6 belongs to Cluster 1
- 2) Population is 3
- 3) Updated values of centroids is [0.24666667 0.2]