Assignment no : 4

Title: Assignment based 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 following1] Which cluster does P6 belongs to?2] What is the population of cluster around m2?3] What is the updated value of m1 and m2?

Objective: To understand and implement k-means algorithm

Outcomes:

- Students will know how the K-means algorithm works.
- Implementation of K-means clustering

Hardware and Software Requirements:

- Windows/Linux OS
- Python
- Anaconda
- Jupyter notebook
- Dell desktop

Theory:

What is Clustering?

Clustering is dividing data points into homogeneous classes or clusters:
Points in the same group are as similar as possible
Points in different group are as dissimilar as possible
When a collection of objects is given, we put objects into group based on similarity.

Application of Clustering:

Clustering is used in almost all the fields. You can infer some ideas from Example 1 to come up with lot of clustering applications that you would have come across.
Listed here are few more applications, which would add to what you have learnt.

- Clustering helps marketers improve their customer base and work on the target areas. It helps group people (according to different criteria's such as willingness, purchasing power etc.) based on their similarity in many ways related to the product under consideration.
- Clustering helps in identification of groups of houses on the basis of their value, type and geographical locations.

- Clustering is used to study earth-quake. Based on the areas hit by an earthquake in a region, clustering can help analyse the next probable location where an earthquake can occur.

Clustering Algorithms:
A Clustering Algorithm tries to analyse natural groups of data on the basis of some similarity. It locates the centroid of the group of data points. To carry out effective clustering, the algorithm evaluates the distance between each point from the centroid of the cluster.
The goal of clustering is to determine the intrinsic grouping in a set of unlabelled data.
What is K-means Clustering?
K-means (Macqueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining.

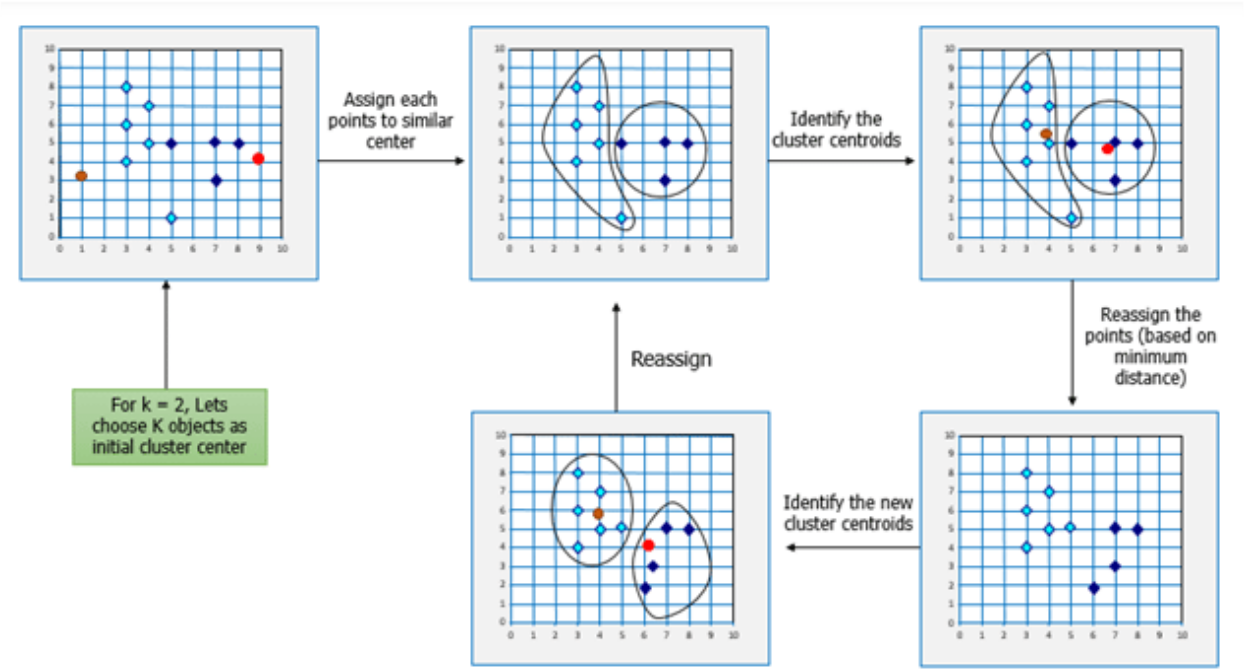
K-means Clustering – Example 1:
A pizza chain wants to open its delivery centres across a city. What do you think would be the possible challenges?

- They need to analyse the areas from where the pizza is being ordered frequently.
- They need to understand how many pizza stores have to be opened to cover delivery in the area.
- They need to figure out the locations for the pizza stores within all these areas in order to keep the distance between the store and delivery points minimum.

Resolving these challenges includes a lot of analysis and mathematics. We would now learn about how clustering can provide a meaningful and easy method of sorting out such real life challenges. Before that let's see what clustering is.
K-means Clustering Method:
If k is given, the K-means algorithm can be executed in the following steps:

- Partition of objects into k non-empty subsets
- Identifying the cluster centroids (mean point) of the current partition.
- Assigning each point to a specific cluster
- Compute the distances from each point and allot points to the cluster where the distance from the centroid is minimum.
- After re-allocating the points, find the centroid of the new cluster formed.

The step by step process:



Now, let's consider the problem in Example 1 and see how we can help the pizza chain to come up with centres based on K-means algorithm.

Similarly, for opening Hospital Care Wards:
K-means Clustering will group these locations of maximum prone areas into clusters and define a cluster center for each cluster, which will be the locations where the Emergency Units will open. These clusters centers are the centroids of each cluster and are at a minimum distance from all the points of a particular cluster, henceforth, the Emergency Units will be at minimum distance from all the accident prone areas within a cluster.
Here is another example for you, try and come up with the solution based on your understanding of K-means clustering.

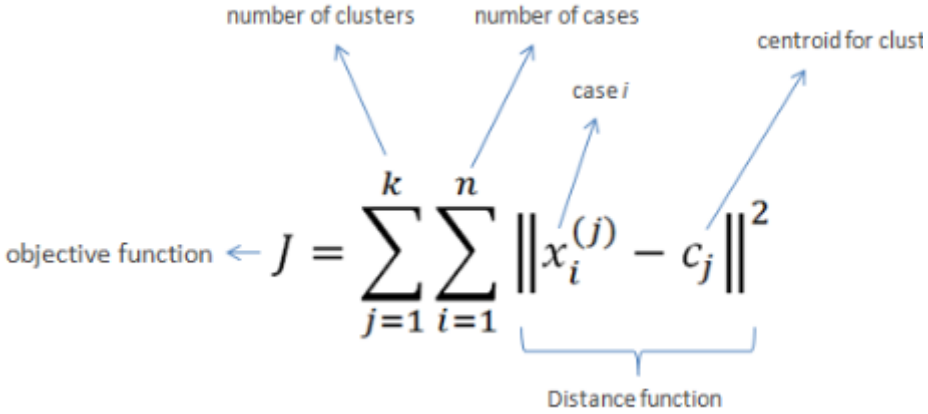
K-means Clustering – Example 2:
Let's consider the data on drug-related crimes in Canada. The data consists of crimes due to various drugs that include, Heroin, Cocaine to prescription drugs, especially by underage people. The crimes resulting due to these substance abuse can be brought down by starting de-addiction centres in areas most affected by this kind of crime. With the available data, different objectives can be set. They are:

- Classify the crimes based on the abuse substance to detect prominent causes.
- Classify the crimes based on age groups.
- Analyze the data to determine what kinds of de-addiction centres are required.
- Find out how many de-addiction centres need to be set up to reduce drug related crime rate.

The K-means algorithm can be used to determine any of the above scenarios by analyzing the available data.
Following the K-means Clustering method used in the previous example, we can start off with a given k, followed by the execution of the K-means algorithm.

Mathematical Formulation for K-means Algorithm:

K-Means clustering intends to partition *n* objects into *k* clusters in which
belongs to the cluster with the nearest mean. This method produces exact
clusters of greatest possible distinction. The best number of clusters *k* lea
greatest separation (distance) is not known as a priori and must be compu
data. The objective of K-Means clustering is to minimize total intra-cluste
the squared error function:



Algorithm

1. Import the Required Packages
2. Create dataset using DataFrame
3. Find centroid points
4. plot the given points
5. for i in centroids():
6. plot given elements with centroid elements
7. import KMeans class and create object of it
8. using labels find population around centroid
9. Find new centroids

Conclusion

Thus we have learned K-means clustering and successfully implemented the algorithm.