Agnik Saha

Roll: 21CS60A01

Running instructions:

Just Run the all cells of ipynb file

I have only used matplotlib for the visualization purpose. Apart from that, I have not used any machine learning special library.

Approach:

Step 1: Finding the optimal value of k

First I need to find out the optimal value of K.
In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

But I have used silhouetteCoefficient score for the optimal value of k in my code.

For k = 2, silhouetteCoefficient score is 0.24820106944568027,

For k = 3, silhouetteCoefficient score is 0.2848008165797193,

For k = 4, silhouetteCoefficient score is 0.27002942656571455,

For k = 5, silhouetteCoefficient score is 0.25432408467364304,

For k = 6, silhouetteCoefficient score is 0.2792678236415544

So I am taking k = 3 for the optimum value as it contains the highest value.

Step 2:
Randomly assign K cluster centers. We make sure that these are very distant from each other.

Step 3:
Calculate the distance of all the data points from all the K number of centers and allocate the points to the cluster based on the shortest distance. The model's inertia is the mean squared distance between each instance and its closest Kth point. The K points are also called centroids which we have randomly selected in step 1. Our goal is to have a model with lowest inertia.

Step 4:
Recompute the centroids (location) once all points are assigned to the nearby Kth centroid.

Step 5:
Repeat steps 2 and 3, until the locations of the centroid strop changing and the cluster allocation of the points become constant!

I have shown visualization of each and every step in the ipynb notebook taking only 2 attributes taking k = 6 for better visualization.