Name: Krishnaraj Thadesar
Roll: PA-10
Batch: A-1
HW1: 1032210888

AIMLT - 6

Title: Write a program to implement any one clustering algorithm like k-means clustering.

Aim: Write a program to implement k-means clustering.

Objective: To study k means clustering.

Algorithm: k-means clustering algorithm.

Platform: Linux x86_64

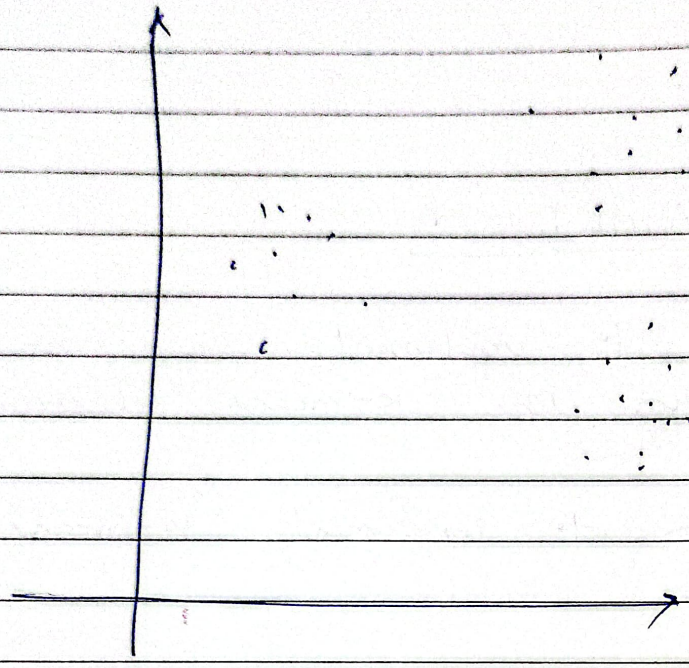
Theory

→ K Means clustering

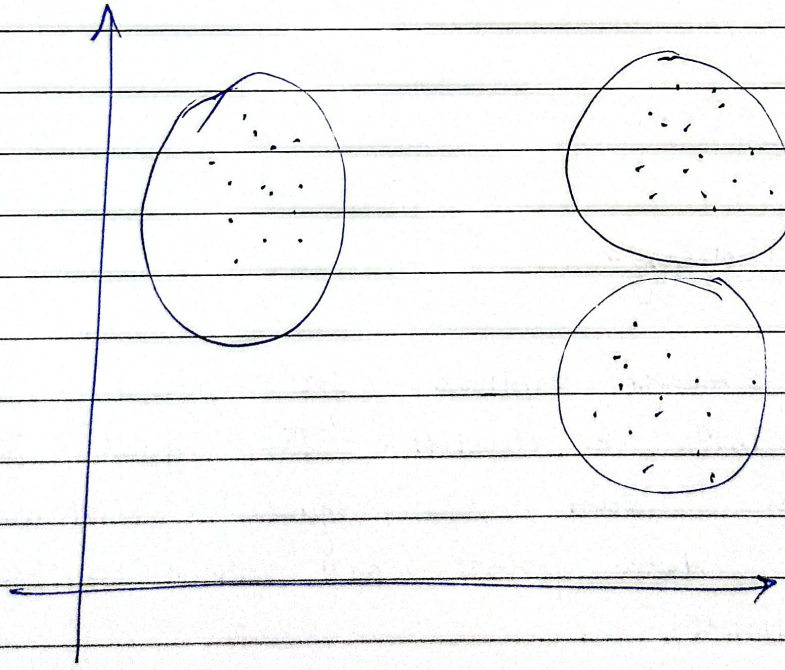
D → It is a ~~means~~ method for grouping n observations into k clusters. The goal is to minimize the sum of squared distances between the data points and their corresponding cluster centroids.

→ K defines no of predefined clusters that need to be created in the process.

(2.) Steps of the K means clustering Alg.



K – Means



(A) Step – 1 : Select the no. of k to decide value of clas.

→ Step : 2 : Selut Random k-points of centroids

→ Step : 3 : Assign each data point to this closest centroid of each cluster

→ Step 4 : Calculate the variance and plan a new centroid .

→ Step - 5 : Repeat the 3rd step , which means reassign each datapoint to the new closest centroid of each cluster.

→ Step 6 : If any reassignment occurs , then go to step 4 or else go to finish.

→ Step - 7 : The model is ready.

(*) Formula :

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

k : no. of clusters
n : no. of cases
distance function
$x_i^{(j)}$: centroid for cluster J.

(3) Objective function of k-means algorithm.

→ 1. The objective function of k-means algorithm is to minimize the total squared errors between the training samples.

2. This is equivalent to minimizing the dis

of the pooled within covariance matrix.

(iii) Algorithm aims to minimize average squared Euclidean distance of documents from their cluster centers.

(iv) A cluster center is defined as the mean or centroid.

(v) K-means clustering minimises within cluster variance.

FAQES

Q.1. How to determine value of K using Elbow method?

→ In the elbow method, we are actually varying the value of K from 1-10. For each value of K, we are calculating WCSS (within cluster sum of squares). WCSS is the sum of the squared distance between each point & the centroid in a cluster. When we plot the WCSS with the K-value, the plot looks like an elbow.

→ As number of clusters increase, the WCSS value will start to decrease.

→ WCSS value is the largest when value of K is 1.

→ When we analyze the graph, we will see that graph will rapidly change at a point, and thus creating an elbow shape.

→ For this point, the graph moves almost parallel to the x-axis. The k-value corresponding to this point, is the optimal value of k, or an optimal no. of clusters.

Q.2. Describe the initialization step in k-means algorithm; why is choice of initial centroids important?

1. As k-means clustering aims to cover an optimal set of cluster centers (centroids) and cluster membership based on distance from them centroids via successive iterations.

2. Now optimal positioning of initial centroids, fewer iterations required.

3. Methods in initialization steps.
1. Random data points: Select k centroids randomly from dataset - This method is volatile & may result in centroids poorly positioned across data space.

2. k-mean++: Initially plans a centroid at a random data point and selects subsequent centroids based on a probability proportional to the squared distance from a nearest point existing centroid.

3. Naive shording: calculates a composite summation, value reflecting all attributi values of each instance. Sorts the dataset based on this value & divides it horizontally into k- chorde.