**Data Mining Assignment 2**

**Problem 2:**

In finding the value of K, I propose using **distortion** as termination condition for deciding the value of K. We use **elbow method** to finalize the value of K.

**Distortion**:

It is calculated as the average of the squared distances from the cluster centers of the respective clusters using the *Euclidean* distance metric.

We store sum of distortion experienced in all clusters for a value of K as distortion score of that K. This way, we store distortion scores of all values of K.

**Elbow Method:**

In this method, the distortion is expressed as function of number of clusters. If we plot this graph, first (if only one is elected) cluster has much distortion. As we keep on adding more clusters, the distance of each point from cluster centroid is reduced. This follows a power curve. We decide an optimum cluster number such that adding more clusters does not yield in significant reduction in the distortion experienced in the data. This value is called as “Elbow” and thus the name elbow method.

**Termination:**

Def findK(list\_of\_destrotion\_score\_of\_all\_K)

For x in list\_of\_destrotion\_score\_of\_all\_K:

If sum(distortion\_score\_till\_k) > 0.5(sum(list\_of\_destrotion\_score\_of\_all\_K)

DecideK = k

Return DecideK

