C Methods to initialize centroids

1- Classic K means random selection method

- We have to do hyper parameter tunning to get the precise value of K.

2. K means ++

- An upgrade over K means.
- We select and initialize first centroid randomly and rest of centroids based on maximum square distance.
- Aim is to push centroids as far as possible from each other.

ⓐ Randomly pick centroid first.

ⓑ Calculate distance between first centroid and all the data points.

$$\boxed{d(c_j, x_i) = \|x_i - c_j\|^2}$$

which is distance of data point (x_i) from farthest centroid (c_j)

ⓒ Select the data point (x_i) with max distance as new centroid.

ⓓ Repeat untill relevant K found.

- Complexity: O(log K)

3. Naive sharding

- Depends on the composite summation value of all the attributes for particular instance or row in dataset
- Aim is to calculate the composite value of attribute and use it to sort instancess of the data.
- Once the data set is sorted, divide it into K shards horizontally.
- Then attributes from each shards will be summed and mean will be calculated.
- The shard attribute mean value collection will be identified as the set of centroids that can be used for initialization.

D Evaluation metrics

1. Dunn index : ratio of minimum inter cluster distance and maximum intra cluster distance.

- Used to identify dense well seperated groups.
- Higher the dunn index (DI) better the seperation.

$$\boxed{DI = \frac{\min[d(i,j)]}{\max[d(k)]}}$$

where,

→ d(i,j) is distance between cluster i and j, which is minimum of all inter cluster distances.

→ d(k) is intra cluster distance of cluster K, which is maximum of all intra cluster distances.

2. Silhoute coefficent

- Silhouete coefficent measures quality of a cluster by checking how similar the data point within cluster is compared to other cluster.
- The discrete value range in +1 to -1.
- +1 means, data points very unsimilar to data points in other cluster
- 0 means, data points very close to decision boundary.
- -1 means, data points very similar to data point in another cluster
- Shiloute coefficent (SC)
  - ↳ +1, great seperataion
  - ↳ -1, worst sepperation.
- Average silhoute score or coefficient (SC_{avg}) used to measure clustering model performance.

$$\boxed{SC_{avg} = \frac{b_i - a_i}{\max(b_i - a_i)}}$$

where,

b_i : avg intra cluster distance

a_i : avg inter cluster distance.

- Calculate SC for each cluster and then calculate SC avg.

3. Rand Index :

- Higher the rand Index (RI) better the clustering.

$$\boxed{RI = TP + TN / TP + FP + FN + TN}$$