

① Clustering

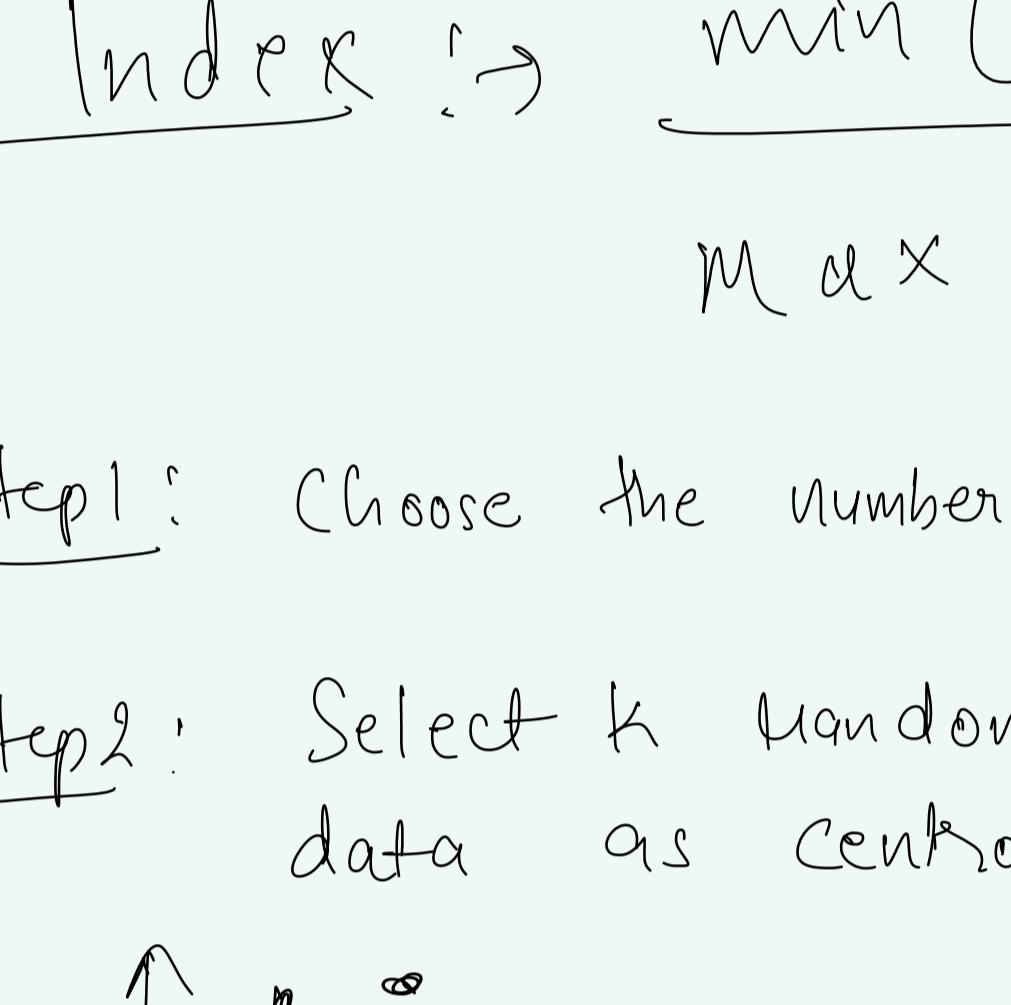
- Most common EDA techniques used to get an intuition about the structure of Data.
- It can be defined as the task of identifying Subgroups in the data such that points in a subgroup are homogenous in the nature.

K-Means Algorithm

K-means is an iterative algorithm that tries to partition the data into k -pre-defined subgroups where each point belongs to only 1 subgroup.

It assigns data points to a cluster such that the sum of the squared distances between the data points and the cluster's centroid is at the minimum.

This approach is k's Expectation Maximization



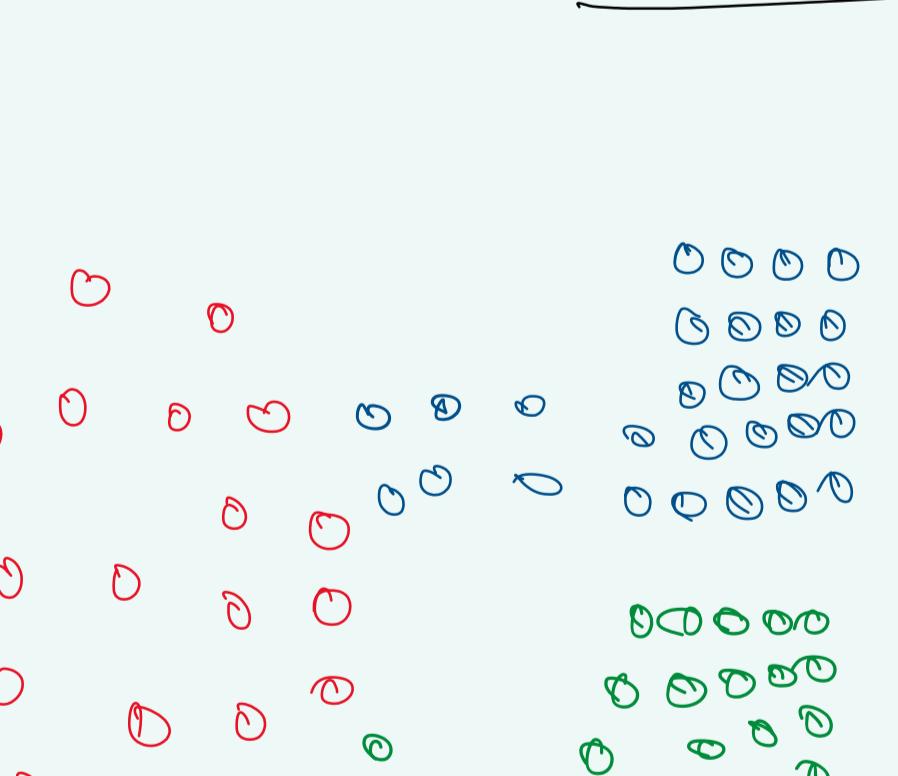
for good cluster, inertia should be as low as possible.

The sum of distances of all ten points within a cluster from the centroid of that cluster is called inertia.

$$\text{Dunn Index} \rightarrow \frac{\min(\text{Inter cluster distance})}{\max(\text{Intra cluster distance})}$$

Step 1: Choose the number of clusters k

Step 2: Select k random points from the data as centroid.

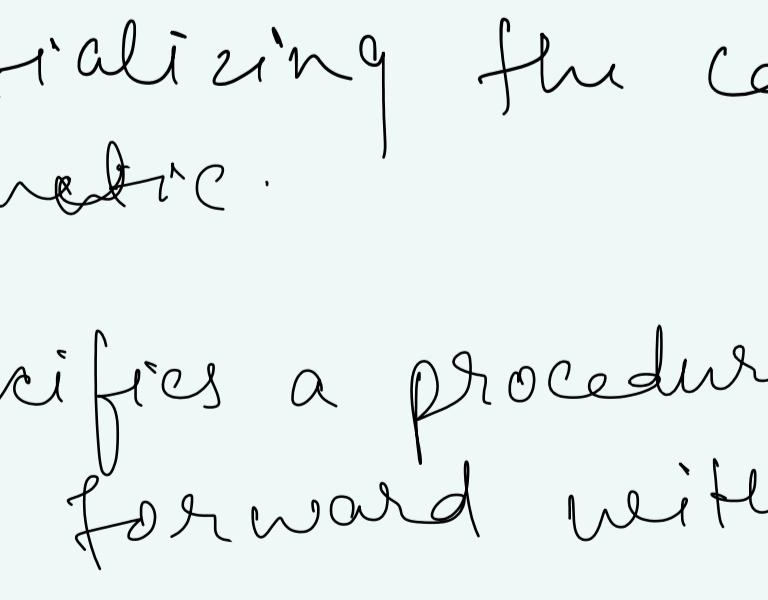


$C=2$

Step 3: Assign all the points to the closest centroid



Step 4: Recompute the centroid of newly formed cluster.



\rightarrow Intra cluster distance

\rightarrow Inter cluster distance

\rightarrow Step 5: Repeat step 3 and step 4

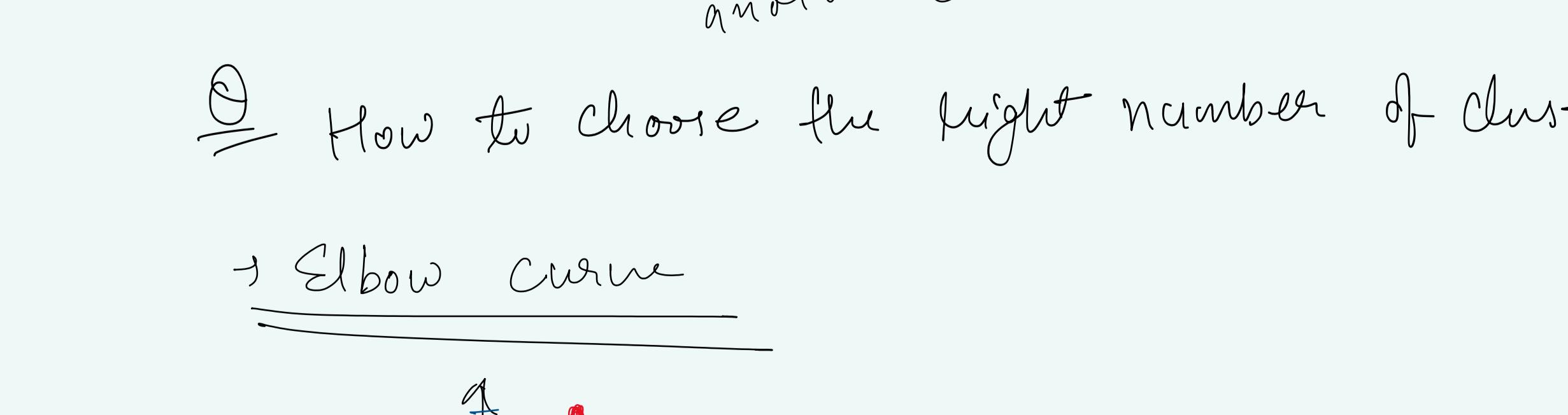
Stopping Condition of k-means :-

① Centroid of newly formed clusters do not change.

② Points remain in the same cluster.

③ Maximum number of iterations are reached.

Challenges with K-means



\rightarrow prob: Due to Density

Sol:- Use a higher number of cluster.

\rightarrow K-Means ++

\rightarrow So initializing the centroids in k-means is kinda problematic.

\rightarrow It specifies a procedure to initialize the cluster centers before moving forward with the standard k-means clustering algorithm.

\rightarrow Step 1: \rightarrow Randomly choose k points as centroids.

\rightarrow Step 2: Assign each data point to the nearest centroid.

\rightarrow Step 3: Compute the mean of all points assigned to each centroid.

\rightarrow Step 4: Repeat steps 2 and 3 until the centroids no longer change.

\rightarrow Step 5: Use the final centroids to cluster the data.

\rightarrow Step 6: Assign each data point to its nearest centroid.

\rightarrow Step 7: Compute the mean of all points assigned to each centroid.

\rightarrow Step 8: Repeat steps 6 and 7 until the centroids no longer change.

\rightarrow Step 9: Use the final centroids to cluster the data.

\rightarrow Step 10: Assign each data point to its nearest centroid.

\rightarrow Step 11: Compute the mean of all points assigned to each centroid.

\rightarrow Step 12: Repeat steps 10 and 11 until the centroids no longer change.

\rightarrow Step 13: Use the final centroids to cluster the data.

\rightarrow Step 14: Assign each data point to its nearest centroid.

\rightarrow Step 15: Compute the mean of all points assigned to each centroid.

\rightarrow Step 16: Repeat steps 14 and 15 until the centroids no longer change.

\rightarrow Step 17: Use the final centroids to cluster the data.

\rightarrow Step 18: Assign each data point to its nearest centroid.

\rightarrow Step 19: Compute the mean of all points assigned to each centroid.

\rightarrow Step 20: Repeat steps 18 and 19 until the centroids no longer change.

\rightarrow Step 21: Use the final centroids to cluster the data.

\rightarrow Step 22: Assign each data point to its nearest centroid.

\rightarrow Step 23: Compute the mean of all points assigned to each centroid.

\rightarrow Step 24: Repeat steps 22 and 23 until the centroids no longer change.

\rightarrow Step 25: Use the final centroids to cluster the data.

\rightarrow Step 26: Assign each data point to its nearest centroid.

\rightarrow Step 27: Compute the mean of all points assigned to each centroid.

\rightarrow Step 28: Repeat steps 26 and 27 until the centroids no longer change.

\rightarrow Step 29: Use the final centroids to cluster the data.

\rightarrow Step 30: Assign each data point to its nearest centroid.

\rightarrow Step 31: Compute the mean of all points assigned to each centroid.

\rightarrow Step 32: Repeat steps 30 and 31 until the centroids no longer change.

\rightarrow Step 33: Use the final centroids to cluster the data.

\rightarrow Step 34: Assign each data point to its nearest centroid.

\rightarrow Step 35: Compute the mean of all points assigned to each centroid.

\rightarrow Step 36: Repeat steps 34 and 35 until the centroids no longer change.

\rightarrow Step 37: Use the final centroids to cluster the data.

\rightarrow Step 38: Assign each data point to its nearest centroid.

\rightarrow Step 39: Compute the mean of all points assigned to each centroid.

\rightarrow Step 40: Repeat steps 38 and 39 until the centroids no longer change.

\rightarrow Step 41: Use the final centroids to cluster the data.

\rightarrow Step 42: Assign each data point to its nearest centroid.

\rightarrow Step 43: Compute the mean of all points assigned to each centroid.

\rightarrow Step 44: Repeat steps 42 and 43 until the centroids no longer change.

\rightarrow Step 45: Use the final centroids to cluster the data.

\rightarrow Step 46: Assign each data point to its nearest centroid.

\rightarrow Step 47: Compute the mean of all points assigned to each centroid.

\rightarrow Step 48: Repeat steps 46 and 47 until the centroids no longer change.

\rightarrow Step 49: Use the final centroids to cluster the data.

\rightarrow Step 50: Assign each data point to its nearest centroid.

\rightarrow Step 51: Compute the mean of all points assigned to each centroid.

\rightarrow Step 52: Repeat steps 50 and 51 until the centroids no longer change.

\rightarrow Step 53: Use the final centroids to cluster the data.

\rightarrow Step 54: Assign each data point to its nearest centroid.

\rightarrow Step 55: Compute the mean of all points assigned to each centroid.

\rightarrow Step 56: Repeat steps 54 and 55 until the centroids no longer change.

\rightarrow Step 57: Use the final centroids to cluster the data.

\rightarrow Step 58: Assign each data point to its nearest centroid.

\rightarrow Step 59: Compute the mean of all points assigned to each centroid.

\rightarrow Step 60: Repeat steps 58 and 59 until the centroids no longer change.

\rightarrow Step 61: Use the final centroids to cluster the data.

\rightarrow Step 62: Assign each data point to its nearest centroid.

\rightarrow Step 63: Compute the mean of all points assigned to each centroid.

\rightarrow Step 64: Repeat steps 62 and 63 until the centroids no longer change.

\rightarrow Step 65: Use the final centroids to cluster the data.

\rightarrow Step 66: Assign each data point to its nearest centroid.

\rightarrow Step 67: Compute the mean of all points assigned to each centroid.

\rightarrow Step 68: Repeat steps 66 and 67 until the centroids no longer change.

\rightarrow Step 69: Use the final centroids to cluster the data.

\rightarrow Step 70: Assign each data point to its nearest centroid.

\rightarrow Step 71: Compute the mean of all points assigned to each centroid.

\rightarrow Step 72: Repeat steps 70 and 71 until the centroids no longer change.

\rightarrow Step 73: Use the final centroids to cluster the data.

\rightarrow Step 74: Assign each data point to its nearest centroid.

\rightarrow Step 75: Compute the mean of all points assigned to each centroid.

\rightarrow Step 76: Repeat steps 74 and 75 until the centroids no longer change.

\rightarrow Step 77: Use the final centroids to cluster the data.

\rightarrow Step 78: Assign each data point to its nearest centroid.

\rightarrow Step 79: Compute the mean of all points assigned to each centroid.

\rightarrow Step 80: Repeat steps 78 and 79 until the centroids no longer change.

\rightarrow Step 81: Use the final centroids to cluster the data.

\rightarrow Step 82: Assign each data point to its nearest centroid.

\rightarrow Step 83: Compute the mean of all points assigned to each centroid.

\rightarrow Step 84: Repeat steps 83 and 84 until the centroids no longer change.

\rightarrow Step 85: Use the final centroids to cluster the data.

\rightarrow Step 86: Assign each data point to its nearest centroid.

\rightarrow Step 87: Compute the mean of all points assigned to each centroid.

\rightarrow Step 88: Repeat steps 87 and 88 until the centroids no longer change.

\rightarrow Step 89: Use the final centroids to cluster the data.

\rightarrow Step 90: Assign each data point to its nearest centroid.

\rightarrow Step 91: Compute the mean of all points assigned to each centroid.

\rightarrow Step 92: Repeat steps 91 and 92 until the centroids no longer change.

\rightarrow Step 93: Use the final centroids to cluster the data.

\rightarrow Step 94: Assign each data point to its nearest centroid.

\rightarrow Step 95: Compute the mean of all points assigned to each centroid.

\rightarrow Step 96: Repeat steps 95 and 96 until the centroids no longer change.

\rightarrow Step 97: Use the final centroids to cluster the data.

\rightarrow Step 98: Assign each data point to its nearest centroid.

\rightarrow Step 99: Compute the mean of all points assigned to each centroid.

\rightarrow Step 100: Repeat steps 99 and 100 until the centroids no longer change.

\rightarrow Step 101: Use the final centroids to cluster the data.

\rightarrow Step 102: Assign each data point to its nearest centroid.

\rightarrow Step 103: Compute the mean of all points assigned to each centroid.

\rightarrow Step 104: Repeat steps 103 and 104 until the centroids no longer change.

\rightarrow Step 105: Use the final centroids to cluster the data.