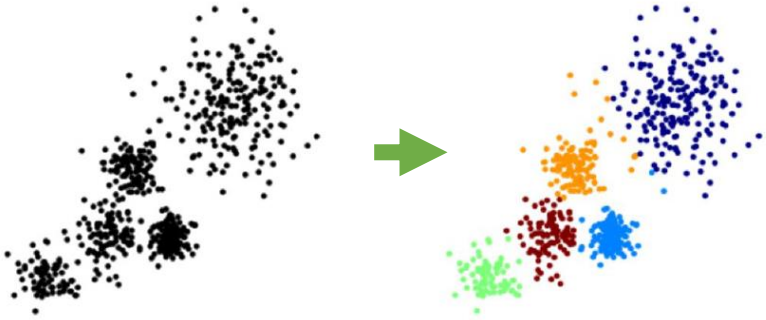
K Means Clustering

K Means Clustering → An unsupervised learning algorithm that will attempt to group similar clusters together in your data.



Used when…

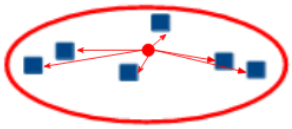
- We have an idea of how many groups we're expecting to find a priori also known as the value of K.

Properties of Clusters

1. All the data points in a cluster should be similar to each other
2. The data points from different clusters should be as different as possible

Evaluation Metrics

1. **Inertia** calculates the sum of distances of all the points within a cluster from the centroid of that cluster



Intra cluster distance

➢ The lesser the inertia value, the better our clusters are.

2. **Dunn Index** ensures that different clusters should be as different from each other as possible.

Intra cluster distance Inter cluster distance

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

Numerator should be higher

Denominator should be lower

- The greater the Dunn Index, the better our clusters are.

How it works

The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.

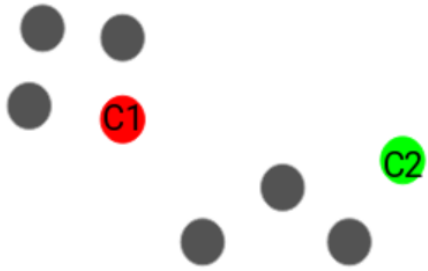
We have these 8 points and we want to apply k-means to create clusters for these points. Here's how we can do it.

Step 1: Choose the number of clusters *k*

The first step in k-means is to pick the number of clusters, k. k value is \ selected as 2 for this example.

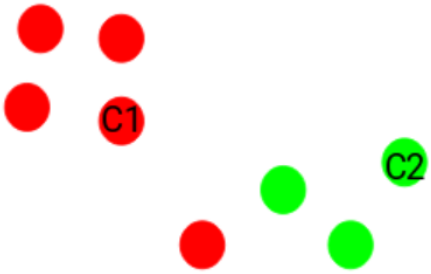
Step 2: Select k random points from the data as centroids

Next, we randomly select the centroid for each cluster. Let's say we want to have 2 clusters, so k is equal to 2 here. We then randomly select the centroid:



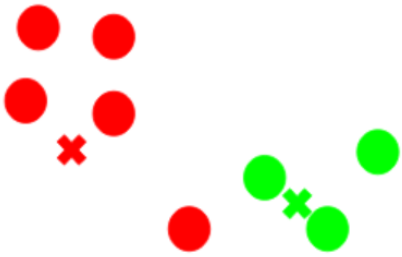
Hence C1 and C2 are randomly selected as centroid for the cluster.

Step 3: Assign all the points to the closest cluster centroid



Points which are closer to the red point are assigned to the red cluster whereas the points which are closer to the green point are assigned to the green cluster.

Step 4: Recompute the centroids of newly formed clusters



Here, the red and green crosses are the new centroids

Step 5: Repeat steps 3 and 4

- Steps 3 and 4 is repeated for multiple iterations until it met 3 conditions:

    1. Centroids of newly formed clusters do not change
    2. Points remain in the same cluster
    3. Maximum number of iterations are reached