K-Means Clustering Notes

Introduction

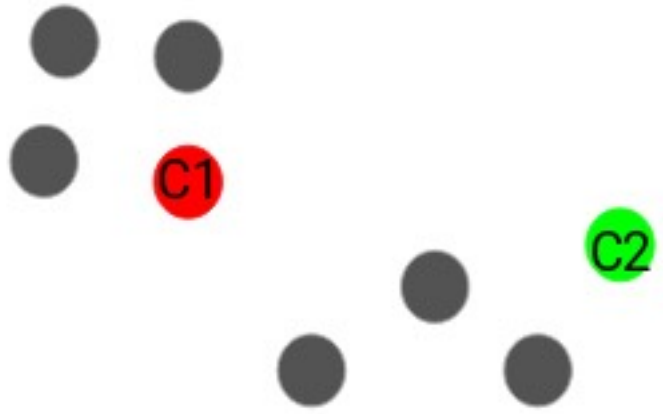
- **K-Means Clustering** is an unsupervised learning algorithm used for partitioning `n` observations into `k` clusters where each observation belongs to the cluster with the nearest mean.

Key Concepts

1. **Cluster**: A group of data points aggregated together due to certain similarities.
2. **Centroid**: The mean position of all the points in a cluster, which represents the center of the cluster.

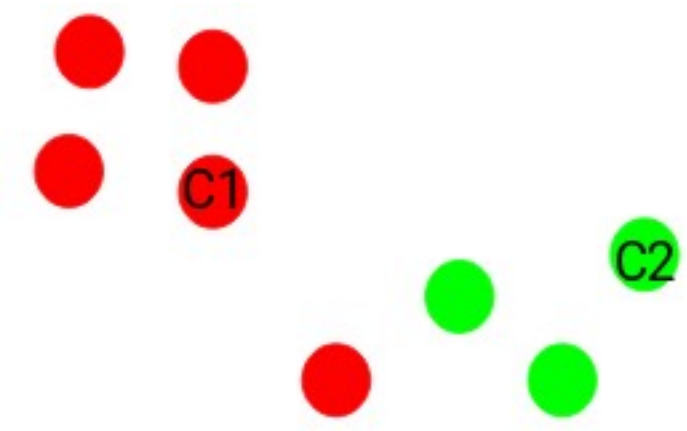
We have these 8 points, and we want to apply k-means to create clusters for these points. Here's how we can do it.

1. **Choose the number of clusters *k:*** The first step in k-means is to pick the number of clusters, k.
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:



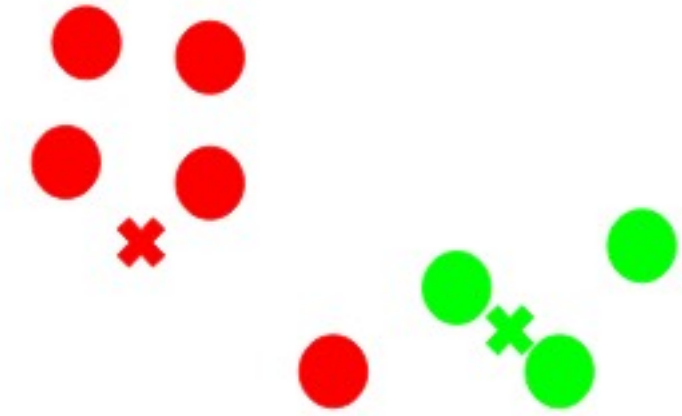
   Here, the red and green circles represent the centroid for these clusters.

3. **Assign all the points to the closest cluster centroid:** Once we have initialized the centroids, we assign each point to the closest cluster centroid:



Here you can see that the points closer to the red point are assigned to the red cluster, whereas the points closer to the green point are assigned to the green cluster.

4. **Recompute the centroids of newly formed clusters:** Now, once we have assigned all of the points to either cluster, the next step is to compute the centroids of newly formed clusters:



Here, the red and green crosses are the new centroids.

5. **Repeat steps 3 and 4:** We then repeat steps 3 and 4:

The *step of computing the centroid and assigning all the points to the cluster*

based on their distance from the centroid is a single iteration. But wait – when should we stop this process? It can't run till eternity, right?

Stopping Criteria for K-Means Clustering

There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:

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

We can stop the algorithm if the centroids of newly formed clusters are not changing. Even after multiple iterations, if we are getting the same centroids for all the clusters, we can say that the algorithm is not learning any new pattern, and it is a sign to stop the training.

Another clear sign that we should stop the training process is if the points remain in the same cluster even after training the algorithm for multiple iterations.

Finally, we can stop the training if the maximum number of iterations is reached. Suppose we have set the number of iterations as 100. The process will repeat for 100 iterations before stopping.