Clustering

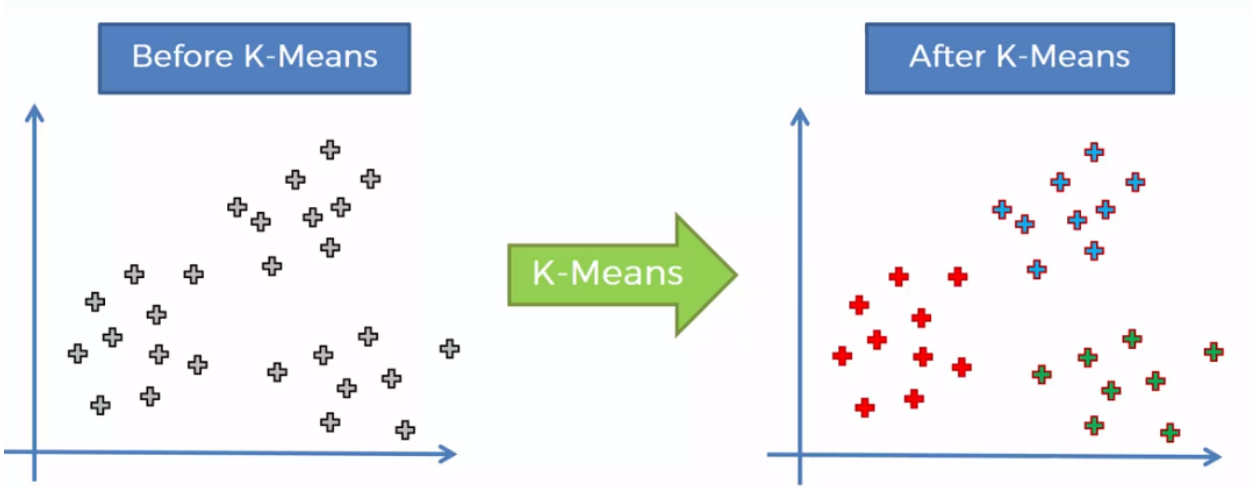
When you look for things such as music, you might want to look for meaningful groups. It could be from a particular artist, a particular genre, a particular language or a particular decade. How you group items gives you more insights about it.

You might find you have a strong linking for vintage music, or something upbeat.

In machine learning, we can often group examples to understand more about the data. Grouping unlabelled examples is called clustering.

For example, let's say that you have a dataset of flowers, with different petal and sepal size but you want to identify what kind of a flower they are. For this, you can group the blobs in the scatter plot and then based on the attributes of the cluster, you can identify what flower it is.

One of the most widely used algorithms for clustering is the K-means algorithm.



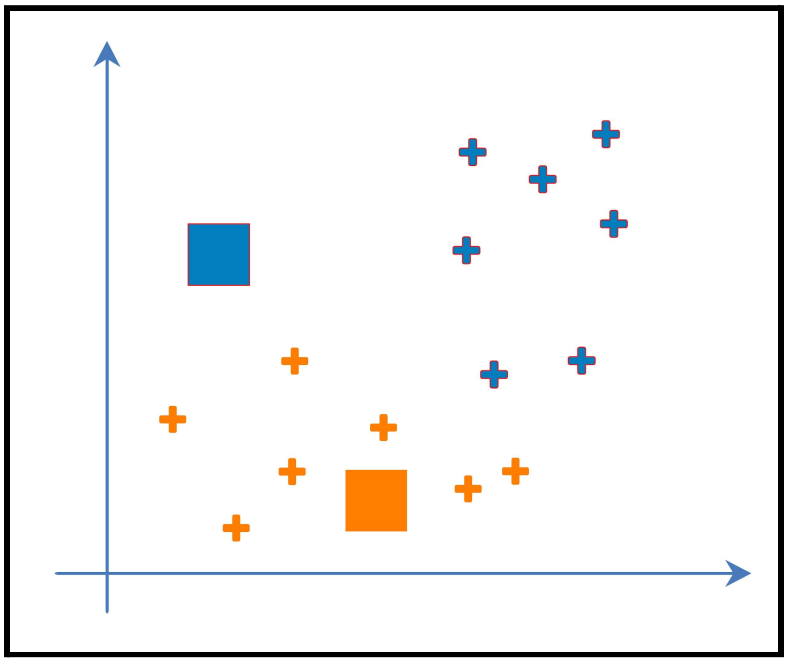
Steps to perform the K-means Algorithm -

Step 1

Choose the number K of clusters

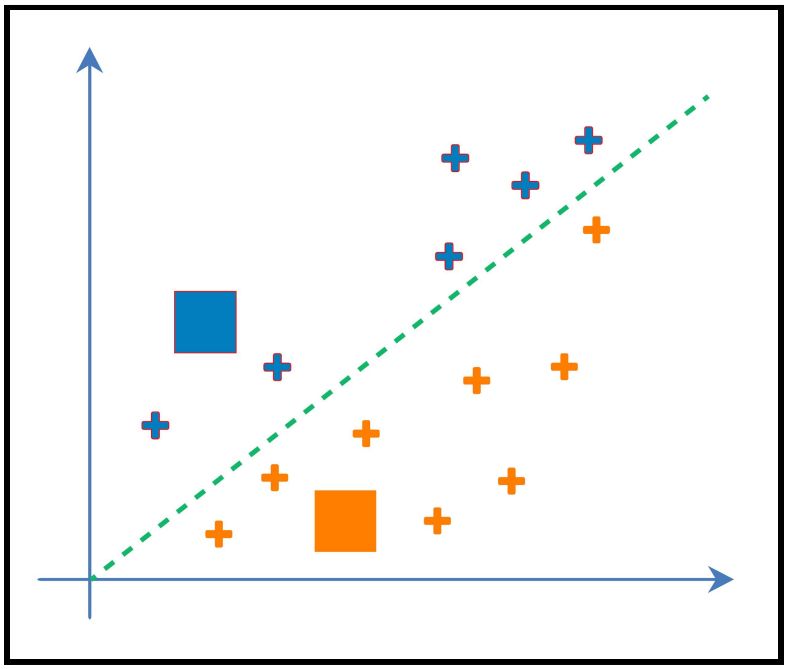
Step 2

Select randomly the center points (centroids) for the K clusters (2 in this case)



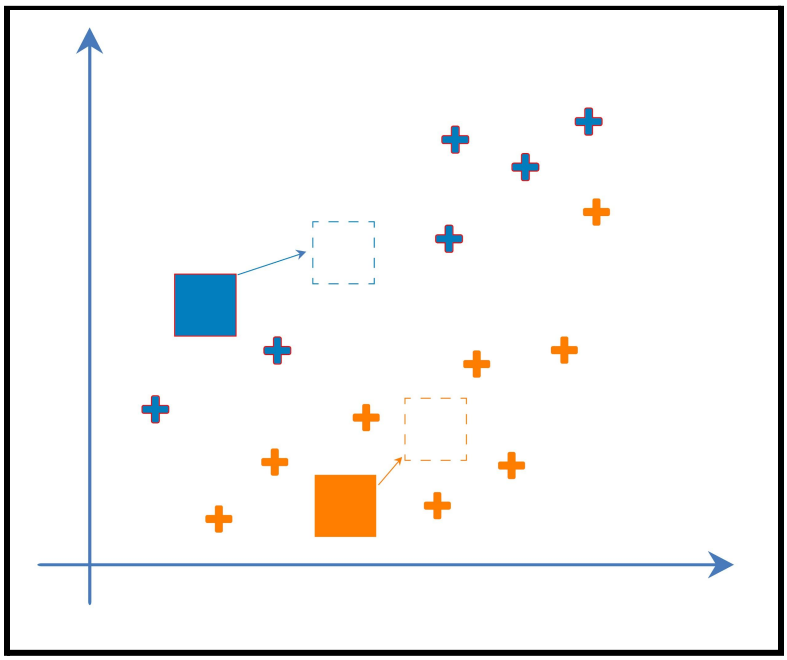
Step 3

Assign each data point to the closest centroid



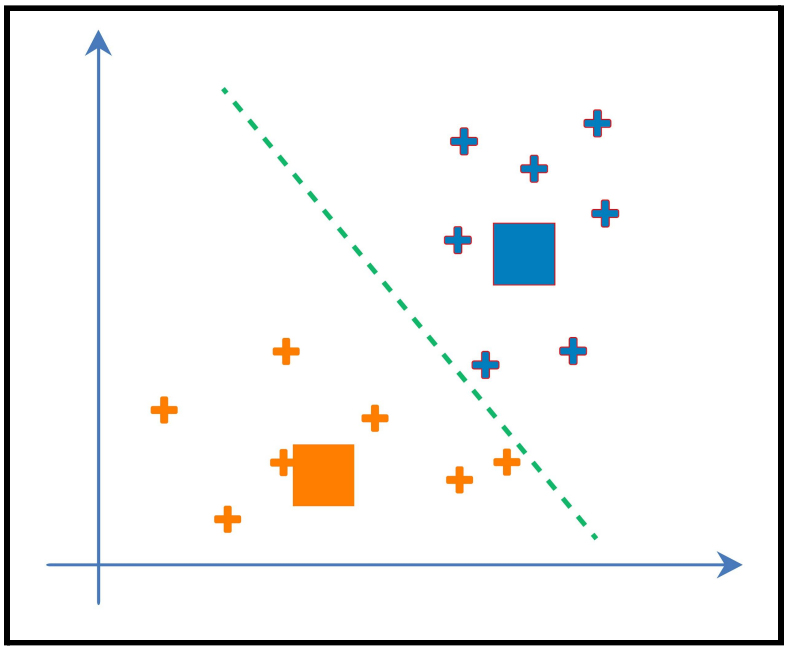
Step 4

Shift the centroids a little for all the clusters

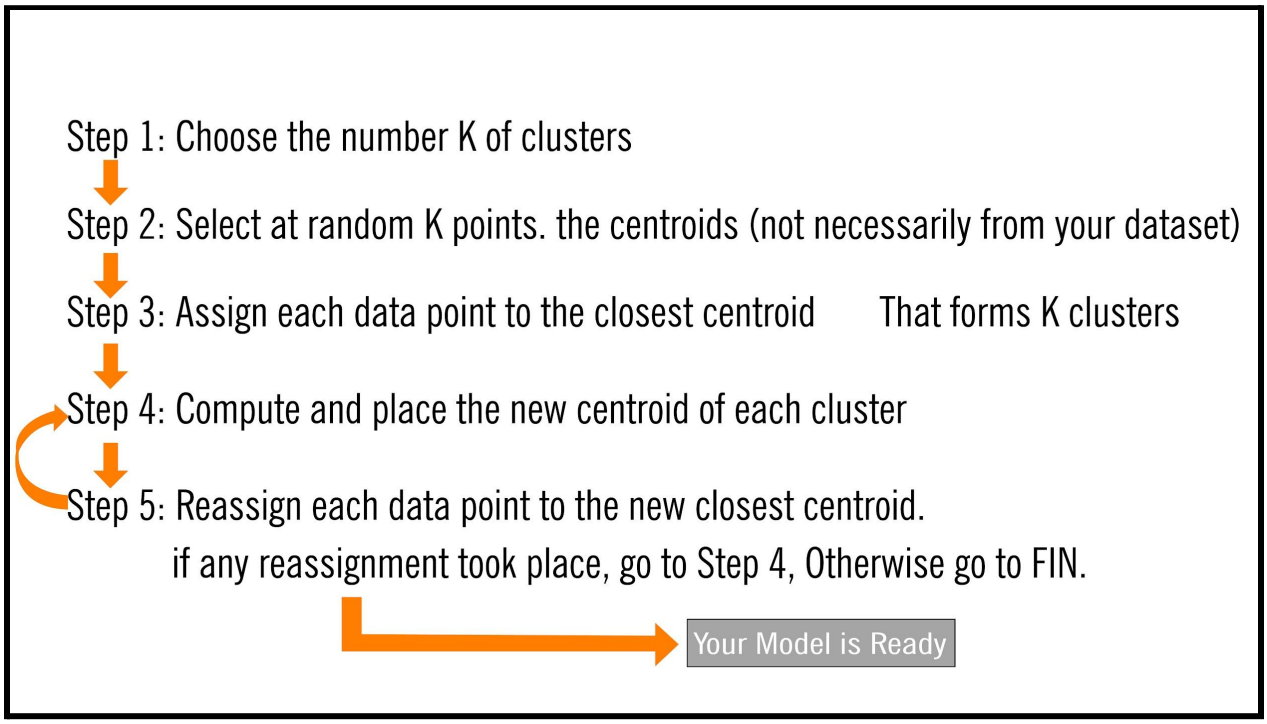


Step 5

Re-assign each data point to the new closest centroid. If any points get reassigned, repeat Step 4 again otherwise the model is ready.



Summary



How to choose the right K?

We use the WCSS parameter to evaluate the choice of K. WCSS stands for Within Cluster Sum of Squares. What this means is that we are going to choose a center point for a cluster, from where all the points falling inside that cluster will be closest.
Then, we will calculate the distance of all the points from the center, add up all the distances and then note the value.
We will then take 2 centre points and do the same. We will choose the value of K to be the one which has the minimum sum of all the distances.

The Elbow method can be used to choose the best value for K.