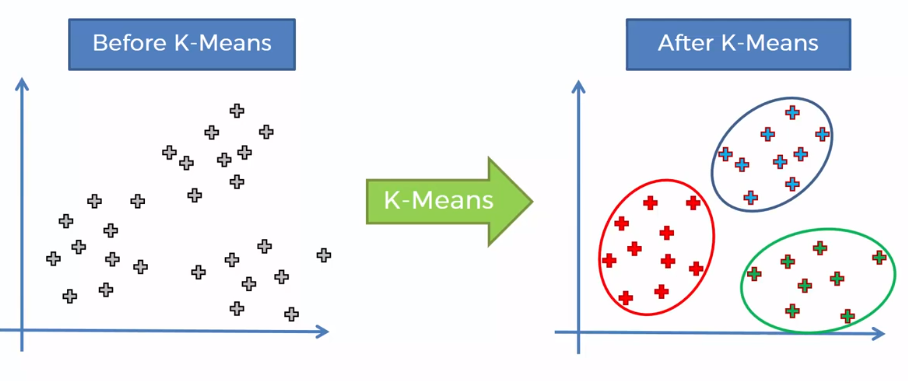
K-Means

Can we identify groups in our data set ?

K-means identify groups from clusters of datapoints in our dataset

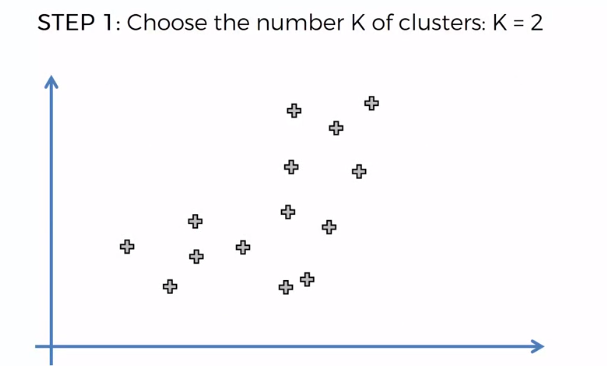


Step by step process:

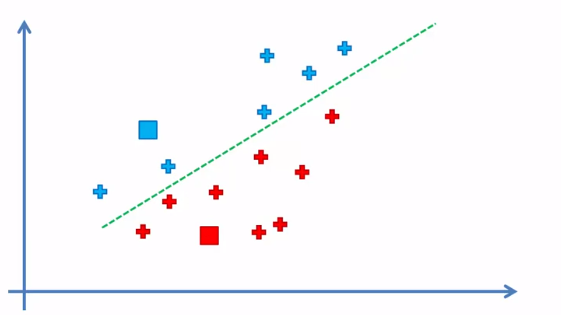
* Chose the number k of clusters
* Select at random k points, the centroids (not necessarily from your dataset )
* Assign each data point to the closest centroid, that forms k clusters
* Compute and place the new centroid of each cluster
* Reassign each data point to the new closest centroid
  + If any reassignment took place go to previous step otherwise FIN

Exmaple:

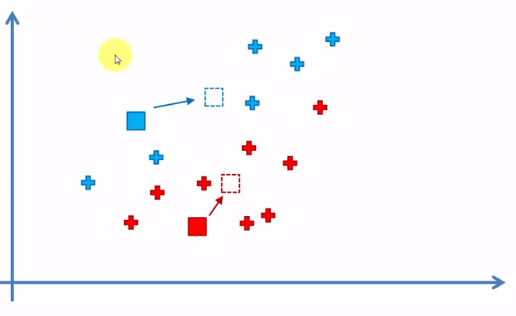
* Step 1: Chose the number k of clusters



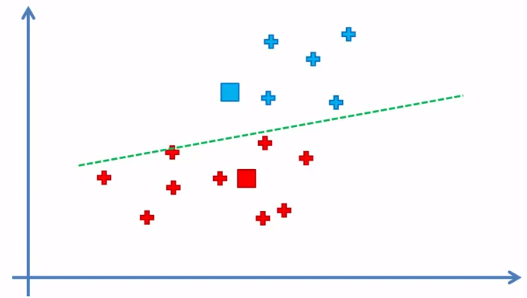
* Step 2: Select at random k points, the centroids (not necessarily from your dataset)



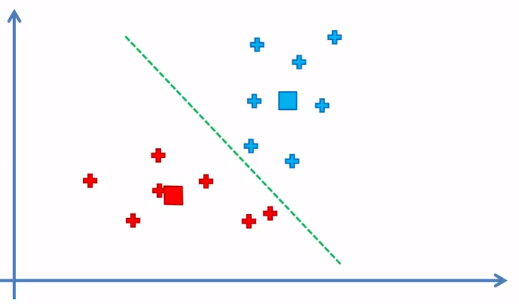
* Step 3: Compute and place the new centroid of each cluster



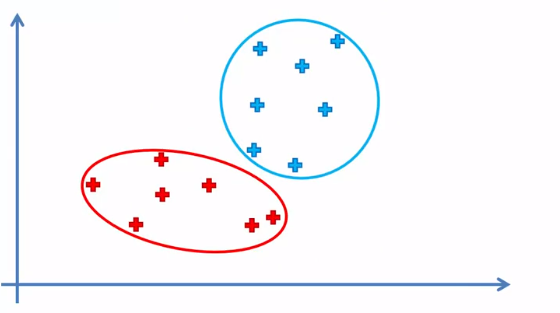
* Reassign each data point to the new closest centroid



* Step 5: Keep repeating step 3 until it is coverage



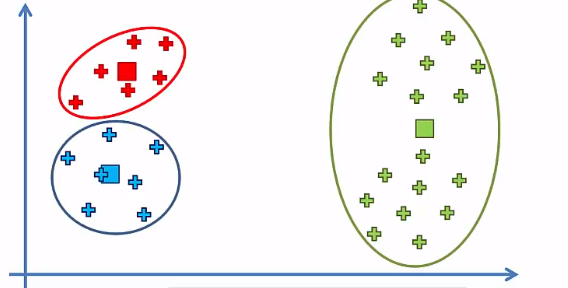
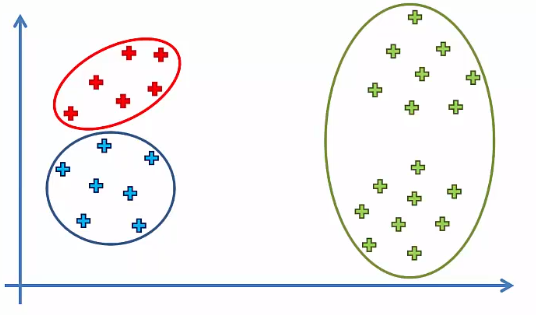
FIN model



K-means Random Initialization Trap:

What happen if we used bad random initialization?

Bad good

Counter this by using K-mean++

How to choose the right number of clusters?

