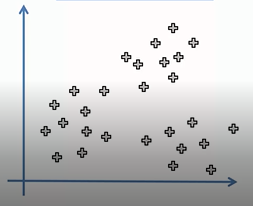
**K – MEANS**

K-MEANS CLUSTERING ALGORITHM

K-Means is an algorithm that allows you to cluster your data.

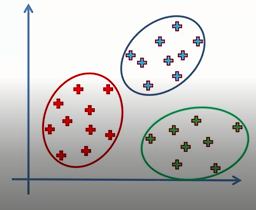
We have got a scattered plot.



We have to variables.

Can we identify certain groups among all variables? And how can we do this?

-> K – means allows you to easily identify clusters.



This is a simplified example. K – means can also work with multidimensional objects.

STEPS:

step1: choose the number ‘k’ of clusters.

step2: select at random k points, the centroids(not necessarily from your dataset).

step3: assign each data point to the closest centroid of each clusters.

step4: compute and place the new centroid of each clusters.

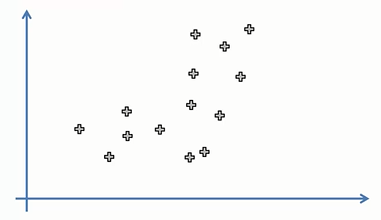
step5: reassign each data point to the newclosest centroid. If any reassignment took place, go to

step4, otherwise go to FIN.

Your model is ready.

Can you visually just identify the final clusters in the following graph?

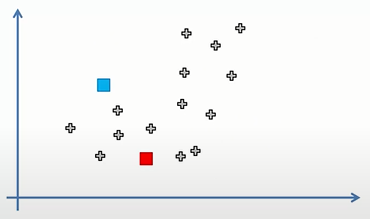
No, it’s difficult. And it becomes more difficult when there are more than two variables.



Here, K – means clustering algorithm is used.

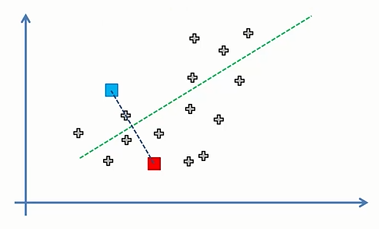
Step 1 says choose the number ‘k’ of clusters. So , suppose we somehow decided 2 clusters.

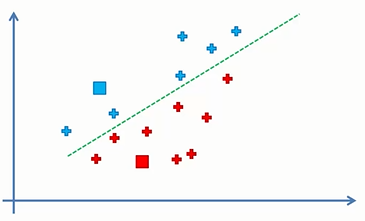
Step 2 says select at random k points, the centroids(not necessarily from your dataset), so here is the graph.



Step 3 says assign each data point to the closest centroid of each clusters

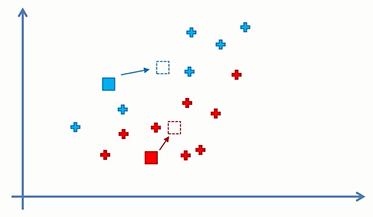
trick is join both, draw perpendicular from center then any point on this line is equidistant from blue and red.





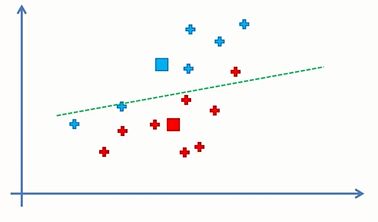
Now any point above this perpendicular line is closer to blue and any point below this perpendicular line is closer to red.

Step 4 says compute and place the new centroid of each clusters.

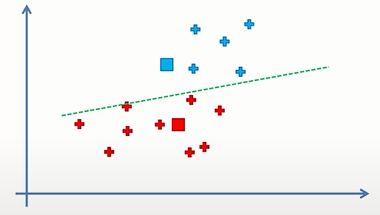


Step 5 says reassign each data point to the new closest centroid. If any reassignment took place, go to step4, otherwise go to FIN

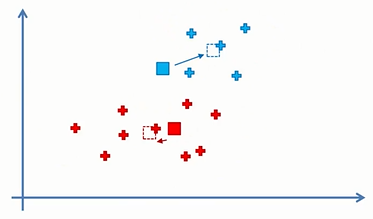
so again put a line



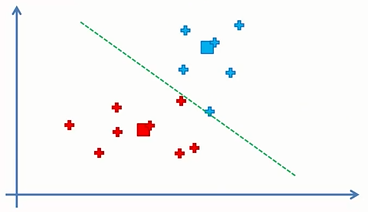
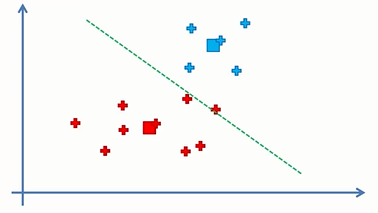
some reassignment did take place so colour them

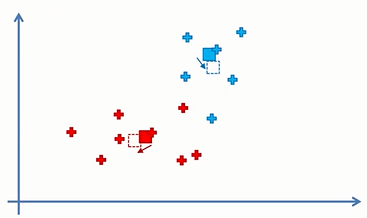


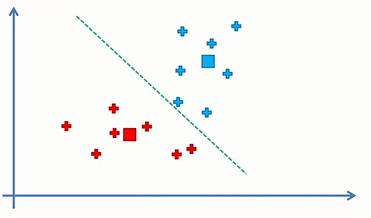
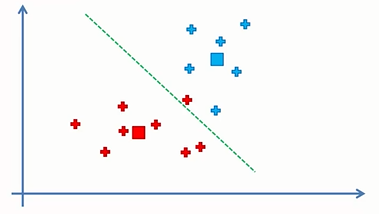
new clustering, therefore step 4 again, i.e compute and place the new centroid of each clusters

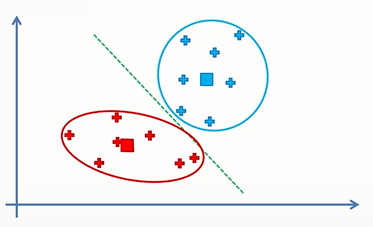
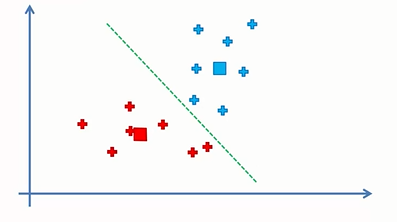
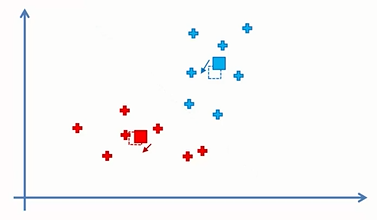


Then step 5 then again steps 4 and 5 repeatedly till there are no reassignments in step 5 and we get the final graph.









Thus the model is now ready and this is the final graph.

