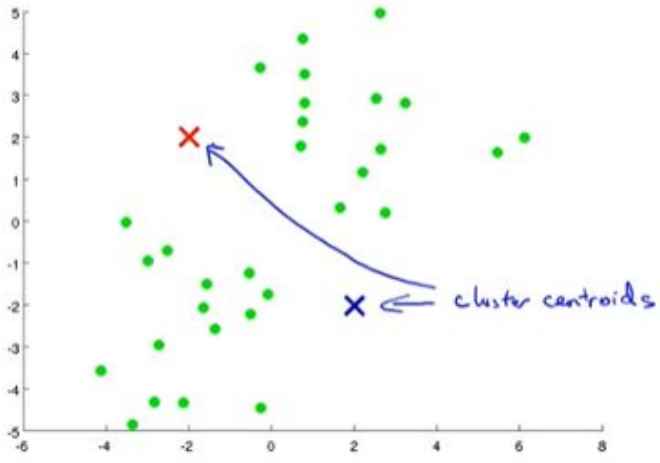
A Complete K Mean Clustering Algorithm From Scratch

K means clustering is the most popular and widely used unsupervised learning model. It is also called clustering because it works by clustering the data. Unlike supervised learning models, unsupervised models do not use labeled data.

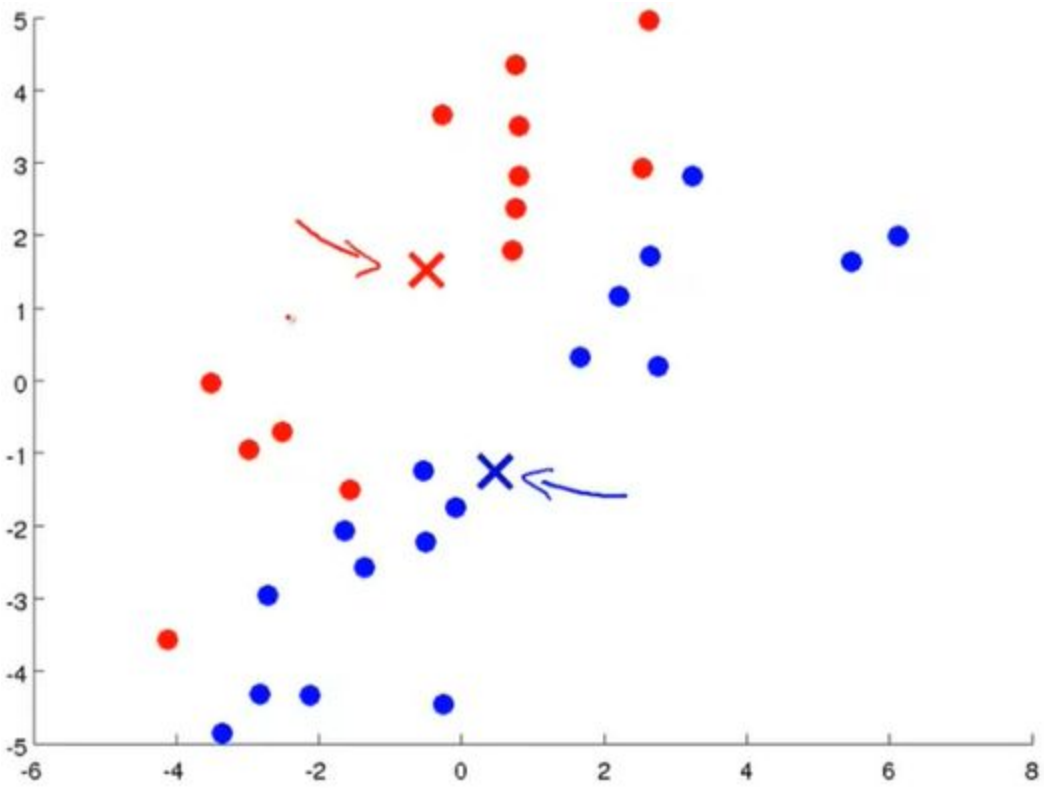
Here is how a k-mean clustering algorithm works:

1. The first step is to randomly initialize a few points. These points are called cluster centroids.



In the picture above, the red and blue points are cluster centroids. You can choose any number of cluster centroids. But the number of cluster centroids has to be less than the total number of data points.

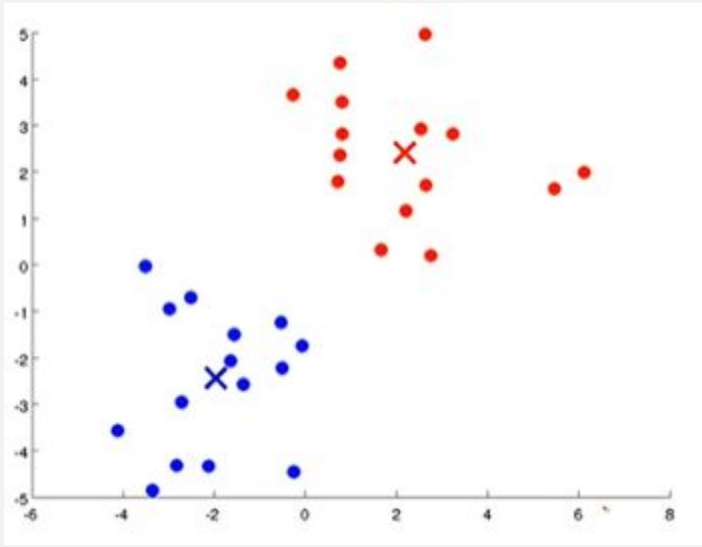
2.The second step is the cluster assignment step. In this step, we need to loop through each of the green dots. Depending on if the dot is closer to the red or blue point, we need to assign it to one of the points. In other words, color the green points either red or blue depending on if it is closer to blue cluster centroid or red cluster centroid.



3. The next step is to move the cluster centroids. Now, we have to take an average of all the red dots that are assigned to the red cluster centroid and move the red cluster centroid to that average. We need to do the same for the blue cluster centroid.

Now, we have new cluster centroids. We have to go back to number 2, the cluster assignment step. We need to rearrange the dots to the new cluster centroids. After that repeat number three.

Numbers 2 and 3 need to be repeated several times until both the cluster centroids are in suitable positions like this picture below.



Look, we just colored all the green dots as per the cluster centroids they are assigned to. The blue cluster centroid is in the center of the blue cluster and the red cluster centroid is in the center of the red cluster.

Procedure

1) Randomly generate **K** number of centroids (points)

2) Calculate the distance between each data point and each centroid

3) Assign each data point to the closest centroid

4) Calculate the mean of each clusters

5) Update the centroids to the calculated means of their respective clusters

6) Repeat step 2 to 5 until there are no changes to the centroids

Mathematical Formula

The distance calculation in step 2 is done using the Euclidean Distance

$d(i, j) = \| \mathbf{x}^i - \mu_j \|$

where $\mathbf{x}^i$ is *i-th* data point and μ_j is *j-th* centroid.

The mean calculation of a cluster in step 4 is done with

$\mu(\omega) = \frac{1}{|\omega|} \sum_{x \in \omega} x$

where ω is one of the specific cluster and x is a data point.