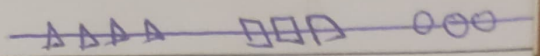
Kmeans clustering:

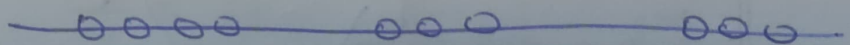
Original data:



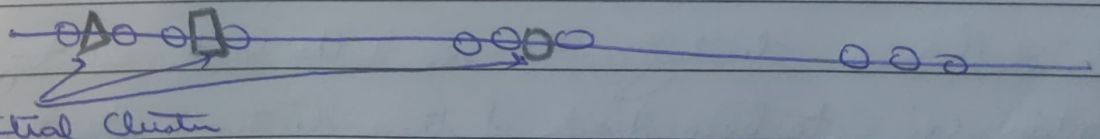
Let's see if computer can make these 3 clusters

Step 1 Select the no of cluster you want to identify in your data.
This is K in Kmean clustering

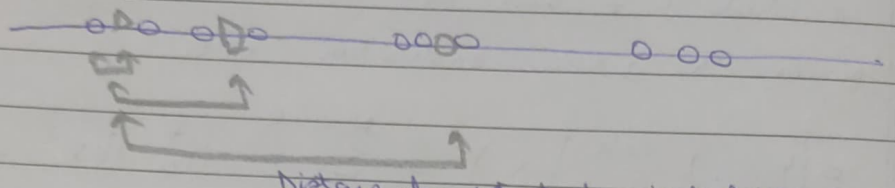
In this case K = 3



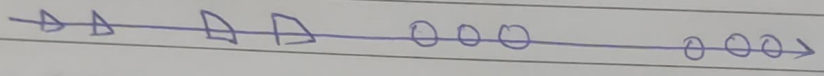
Step 2 Randomly select 3 pts (distinct)



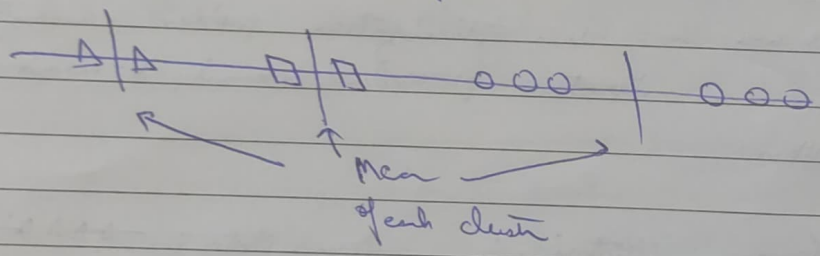
Step 3: Measure the distance b/w 1st pt of true initial cluster



Step 4: Assign 1st pt to nearest cluster.



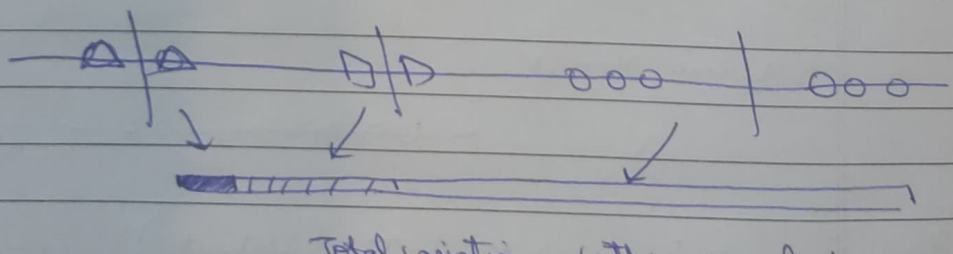
Step 5: Calculate mean of each cluster



Then we repeat what we just did (measure ad cluster) using the mean values.

I this case clustering won't change much

We can assess the quality of clustering by adding up the variation within each cluster

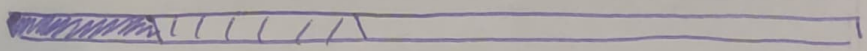


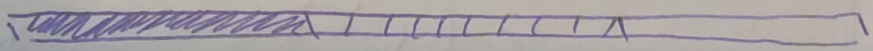
Since if mean clustering can't see the best clustering, its only option is to keep track of these clusters, ad this total variance, ad do the whole thing over again with diff starting pts.

PAGE EDGE
Date : / /

So we start again and the cluster all the remaining pts, calculate mean of each cluster, re-cluster based on new means. It repeats until clusters no longer change.

Now that the data are clustered, we sum the variation within each cluster.

1st attpt 

2nd attpt 

3rd attpt 

At this pt, K means clustering knows that 2nd attpt is best clustering so far. But it does not know if its best overall, so it will do a few more clusters (it does as many times as we tell) at one back at each one if it is still the best.

You can find by using elbow plot.

PAGE EDGE
Date : / /

Q) $X = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$
Perform K means clustering for $K = 2$

Ans) $K = 2$ ∴ K_1, K_2 - 2 clusters

Step 1 Randomly select (2 means) $m_1 = 4$, $m_2 = 12$

$\sqrt{(2-4)^2} = 2$
$\sqrt{(2-12)^2} = 10$ $K_1 = \{2, 3, 4\}$ $K_2 = \{10, 11, 12, 20, 25, 30\}$

Step 2 $m_1 = \frac{2+3+4}{3}$ $m_2 = \frac{10+11+12+20+25+30}{6}$

$= 3$ $= 18$

$K_1 = \{2, 3, 4, 10\}$ $K_2 = \{11, 12, 20, 25, 30\}$

Step 3 $m_1 = \frac{2+3+4+10}{4}$ $m_2 = \frac{11+12+20+25+30}{5}$

$= 4.75$ $m_2 = 19.6$

$K_1 = \{2, 3, 4, 10, 11, 12\}$ $K_2 = \{20, 25, 30\}$

Step 4 $m_1 = \frac{2+3+4+10+11+12}{7}$ $m_2 = \frac{20+25+30}{3} = 25$

$K_1 = \{2, 3, 4, 10, 11, 12\}$ $K_2 = \{20, 25, 30\}$

$m_1 = 7$ $m_2 = 25$

Since it gives same value for cluster m_1 at m_2, we have to stop.

$K_1 = \{2, 3, 4, 10, 11, 12\}$ $K_2 = \{20, 25, 30\}$

$m_1 = 7$ $m_2 = 25$