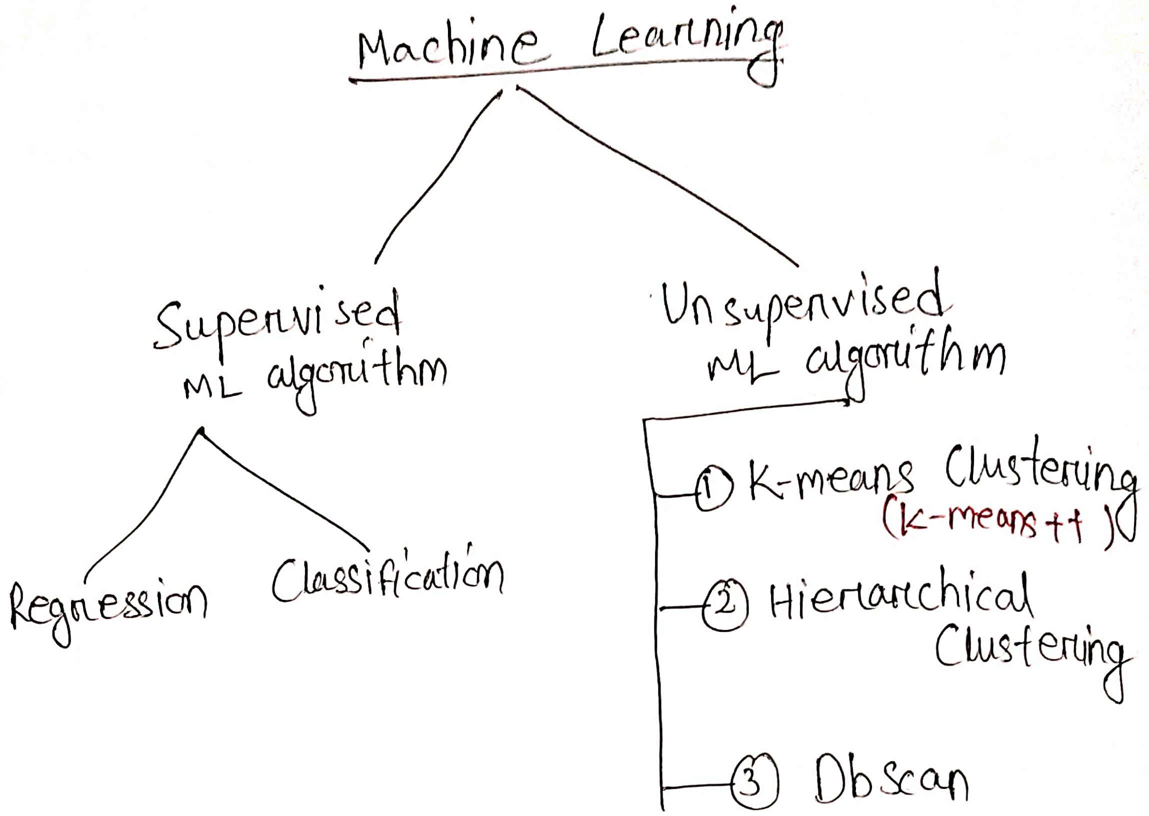
Clustering



In case of Supervised

	I/P	O/P
Height	weight	BMI
180	70	21
160	80	23
150	75	23
140	90	24
130	100	25
170	60	19

In case of Unsupervised

Height	weight	BMI	
180	70	~~a~~normal	①
160	45	under weight	②
150	70	normal	①
140	90	normal	①
130	100	over weight	③
170	60	normal.	①

→ we can • make cluster by using BMI values.

→ ③ groups

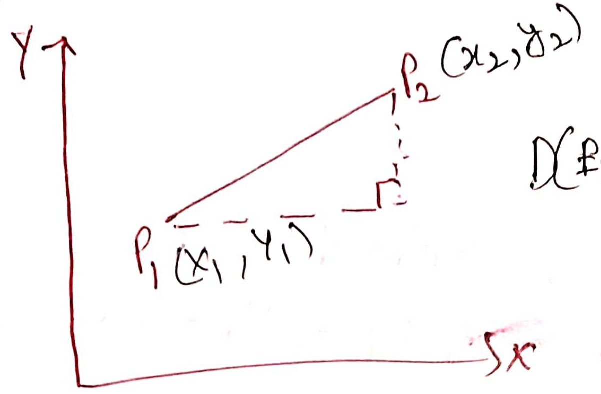
① K-means

→ Unsupervised algorithm works by calculating similarity score.

→ ① Eucledean distance. ③ cosine similarity.

② Manhatten distance.

① Euclidean distance



$$D(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

② Manhatten distance

$$D(P_1 - P_2) = |x_1 - x_2| + |y_1 - y_2|$$

③ cosine similarity

$$\cos(\theta) = \frac{A \cdot B}{|A| |B|}$$

Height weight → Clustering. using K-means.

	Height	weight	
	185	72	C_1
	170	56	C_2
3	168	60	
4	179	68	
5	182	72	
6	188	77	
7	180	71	
8	183	84	
9	180	88	
10	180	67	
11	167	76	

Step ① Intialization of centroid

In case of k-means centroid initialization is random. Let

Let $K = 2$ (K = no. of centroid)

Lets assume 1st and 2nd row as C_1 and C_2

Step ② Euclidean distance.

So to make cluster we have to find euclidean distance to other point from C_1 and C_2 (two cluster)

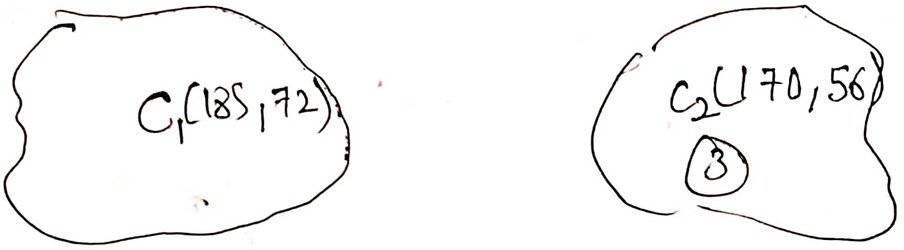
$$D(C_1, 3) = \sqrt{(185-168)^2 + (72-60)^2}$$

$$= \sqrt{(17)^2 + (12)^2} = 20.8$$

$D(C_2, 3) = \sqrt{(168-170)^2 + (60-56)^2}$

$= \sqrt{2^2 + 4^2} = 4.4$

so $D(C_2, 3) < D(C_1, 3)$

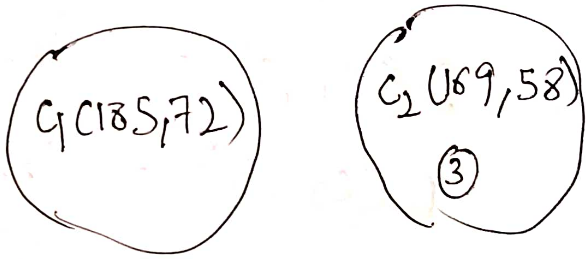


As ③ belongs to C_2 (cluster-2) C_2 need to ~~be~~ be updated.

step-3

new $C_2 = \left(\frac{170+168}{2}, \frac{56+60}{2}\right)$

$= (169, 58)$



Again step ②

$D(C_1, 4) = \sqrt{(185-179)^2 + (72-68)^2}$

$= \sqrt{6^2 + 4^2} = 7.07$

$D(C_2, 4) = \sqrt{(169-179)^2 + (58-68)^2}$

$= \sqrt{10^2 + 10^2} = 14.14$

As $D(C_1, 4) < D(C_2, 4)$

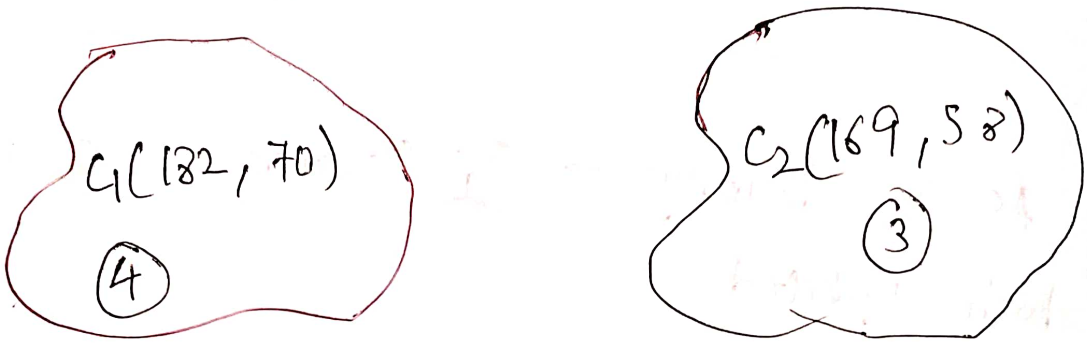
S ④ belongs to C_1 (centroid-1)

Step-3

As 4 belongs to C_1, C_1 centroid needs to update

$$\text{new } C_1 = \frac{185+179}{2}, \frac{72+68}{2}$$

$$= (182, 70)$$



Again step 2

$$D(C_1, 5) = \sqrt{(182-182)^2 + (72-70)^2} = 2$$

$$D(C_2, 5) = \sqrt{(182-169)^2 + (72-58)^2}$$

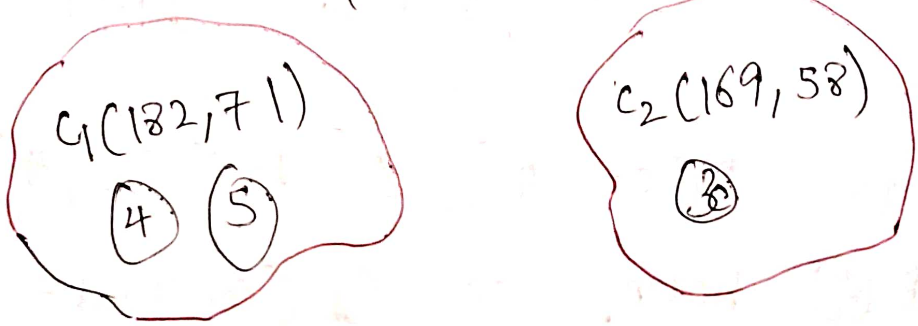
$$= \sqrt{(13)^2 + (14)^2} = 19.1$$

As $D(C_1, 5) < D(C_2, 5)$

↓ step 3 Repeted.

$$\text{new } C_1 = \left(\frac{182+182}{2}, \frac{70+72}{2}\right)$$

$$= (182, 71)$$



steps

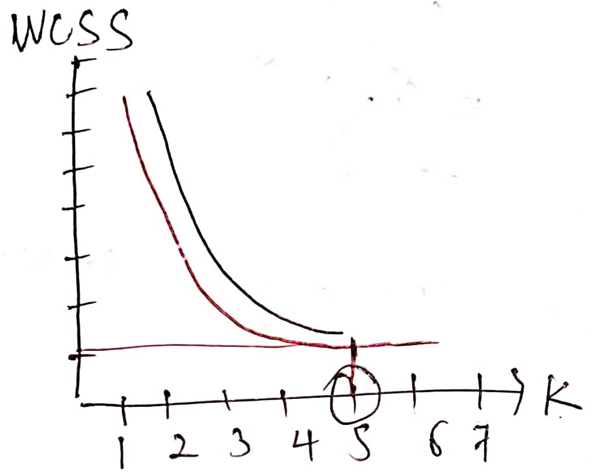
① centroid

② Distance (compare minimum distance)

③ include point in cluster & update centroid.

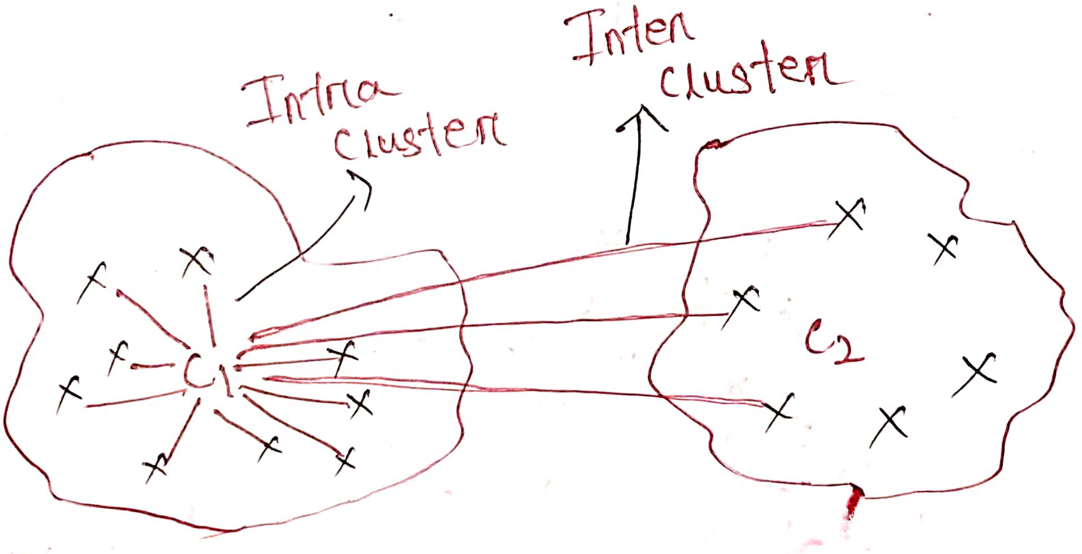
Q. How to decide k value ?

Ans. ELBOW method.

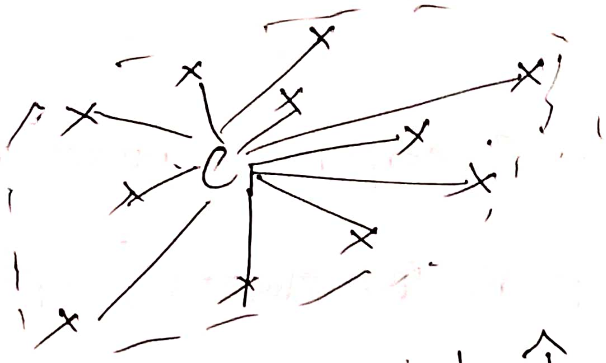


K=5

WCSS → within - cluster sum of square.

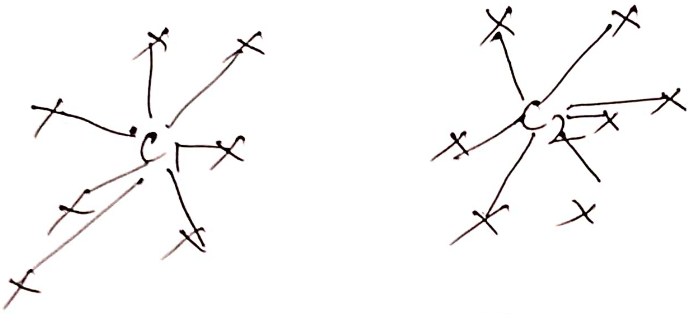


let $\boxed{k = 1}$



$WCSS_1 =$ high ↑

$\boxed{k = 2}$

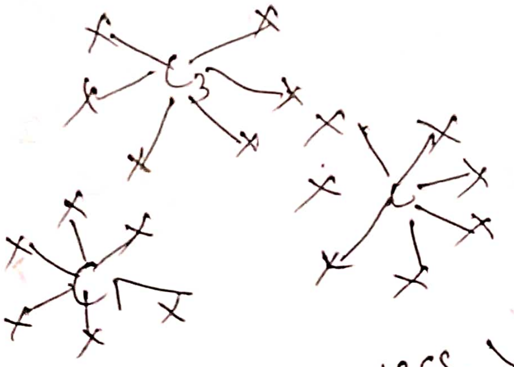


$WCSS_2 =$ Low (As compare to $k=1$)

$$\boxed{WCSS = \sum_{i=1}^{n} d(C, x_i)^2}$$

$WCSS_1 > WCSS_2$

$\boxed{k = 3}$



$WCSS_1 > WCSS_2 > WCSS_3$

wcss is minimum at specific value of k
assume (k = 5)

☆ How to validate clusten value (k = 5]

① Dunn index.

② Silhouette score

① Dunn index $= \frac{\max dist(x_i, x_j)}{\max dist(y_i, y_j)}$

② Silhouette score $= \frac{b_i - a_i}{\max(b_i - a_i)}$ → [-1 to 1]

In silhouette score

a_i = intra cluster distance

b_i = inter cluster distance.

If $a_i > b_i$

silhouett score -ve (wrost value)

→ In case of k-means ++, random initialization of k-means problem is solved.

→ Kmeans is a centroid based algorithm.