K-Means Clustering

* Most popular
* Main use :- Customer segmentation
* Introduced by Stuart Lloyd in 1957 at Bell Labs

famous in 1967 by James MacQueen

Data

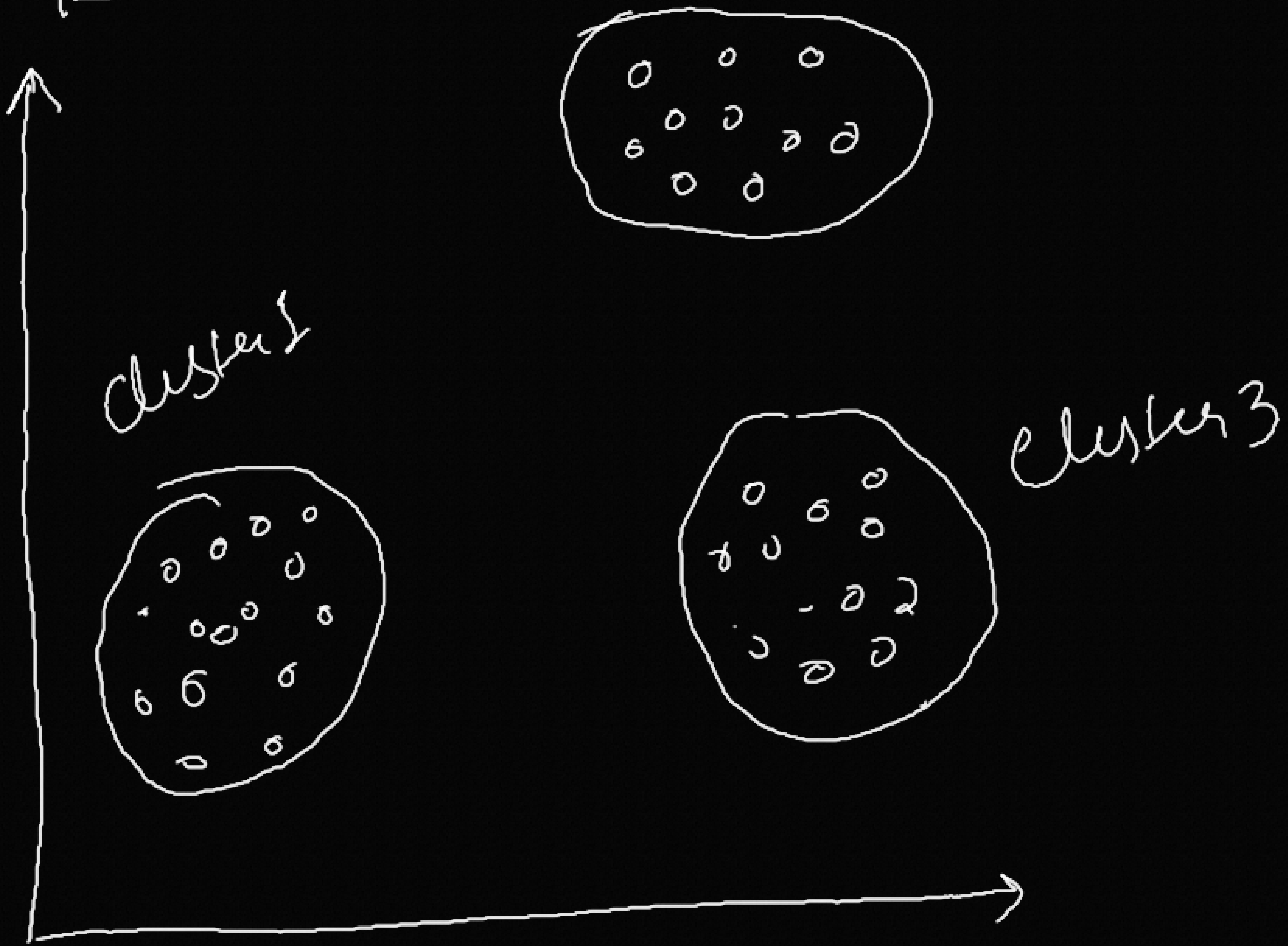
Graph plot

Assumption

① Nearly spherical

② Same size nearly

General



Mathematical

Data

X_1	X_2
1.0	1.5
1.5	1.8
5.0	8.0
6.0	8.5
1.2	1.0
6.0	6.0

1. Set the value of K (no of clusters)

   We set $K=2$.

2. Initialize centroids

   Randomly initialize one centroid to each cluster

for eg.

C_1 (Cluster 1) $= (1.0, 1.5)$

C_2 (Cluster 2) $= (6.0, 8.0)$

③ Calculate distance (euclidean) (x_1, y_1) (x_2, y_2)

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

for (1.0, 1.5)

from C_1 (1.0, 1.5) $= \sqrt{(1.0 - 1.0)^2 + (1.5 - 1.5)^2}$

$= 0$

from C_2 (6.0, 8.0) $= \sqrt{(6.0 - 1.0)^2 + (8.0 - 1.5)^2}$

$= 8.2$

for (1.5, 1.8)

from C_1 = 0.583

from C_2 = 7.66

Calculate for all the points

Each point is assigned to nearest cluster (less distance)

Cluster 1	Cluster 2
(1.0, 1.5)	(5.0, 8.0)
(1.5, 1.8)	(6.0, 8.5)
(1.2, 1.0)	(6.0, 6.0)

④ Update Centroids (Mean of points of cluster)

$$C_{1\,new} = \left(\frac{1.0+1.5+1.2}{3}, \frac{1.5+1.8+1.0}{3}\right) = (1.23, 1.43)$$

$$C_{2\,new} = \left(\frac{5.0+6.0+6.0}{3}, \frac{8.0+8.5+6.0}{3}\right) = (5.67, 7.5)$$

We got the new centroids

⑤ Again repeat step 3-4 & again get new centroids

Until centroid value dont change significantly

When there is no significant change

$\Rightarrow$ Algorithm is "Converged"

Training complete.

Prediction

$C_1 = (1.23, 1.43)$

$C_2 = (5.67, 7.5)$

New point $= (2.0, 2.5)$

* Distance from $C_1 = 1.31$
* Distance from $C_2 = 6.2$

New point is closer to C_1

New point $\longrightarrow$ Cluster 1.

(output)

Elbow Method (find best K value)

5 points

A (1,2)   B (2,3)   C (3,4)   D (8,9)   E (9,10)

① Calculate SSD (sum of Squared distance)

Avg of all points in cluster

for K=1

$$\mu = \left(\frac{1+2+3+8+9}{5} , \frac{2+3+4+9+10}{5} \right) = (4.6, 5.6)$$

(Mean)

$$SSD = \Sigma \underbrace{\| x - \mu \|}^{2}$$

$\| \ \| \Rightarrow$ euclidean

$x \rightarrow$ point

$\mu \rightarrow$ Mean

$$\text{Distance}(A, \mu) = \| (1,2) - (4.6, 5.6) \|$$

$$(1,2)\ (4.6, 5.6) \Rightarrow (1-4.6)^2 + (2-5.6)^2$$

$$\Rightarrow \underline{\underline{25.92}}$$

$$\text{Distance}(C, \mu) = 5.12$$

$$\text{Distance}(B, \mu) = 13.52$$

$$\text{Distance}(E, \mu) = 38.72$$

$$\text{Distance}(D, \mu) = 23.12$$

$$SSD_1 = 25.92 + 13.52 + 5.12 + 23.12 + 38.72$$

$$= \underline{\underline{106.4}} \ //$$

for $K=2$

Cluster 1

A(1,2)
B(2,3)
C(3,4)

$$\mu_1 = (2,3)$$

SSD for cluster 1

$$(A, \mu_1) = 2$$
$$(B, \mu_1) = 0$$
$$(C, \mu_1) = 2$$

Cluster 2

D(8,9)
E(9,10)

$$\mu_2 = (8.5, 9.5)$$

SSD for cluster 2

$$(D, \mu_2) = 0.5$$
$$(E, \mu_2) = 0.5$$

$SSD_{\text{cluster 1}} = 2 + 0 + 2 = \underline{\underline{4}}$.

$SSD_{\text{cluster 2}} = 0.5 + 0.5 = \underline{\underline{1}}$.

$SSD_2 = 4 + 1$

$= \underline{\underline{5}}$.

for $K = 3$.

Cluster 1	Cluster 2	Cluster 3
A	C	D
B		E

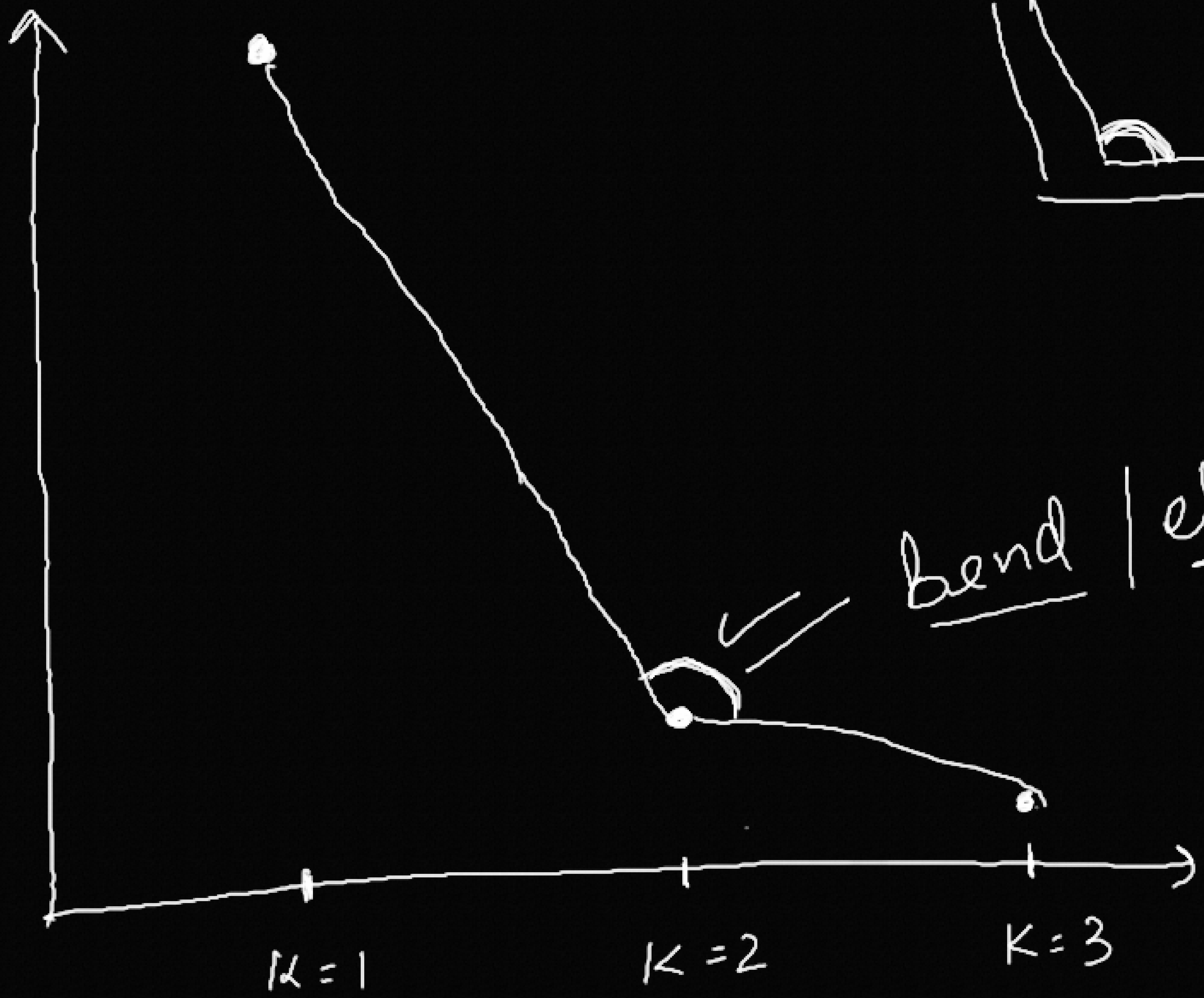
$SSD_3 = 2$.

$K=1 \longrightarrow 106.4$

$K=2 \longrightarrow 5$

$K=3 \longrightarrow 2$

Do elbow plotting



Bend | elbow

K = 2 best K-value

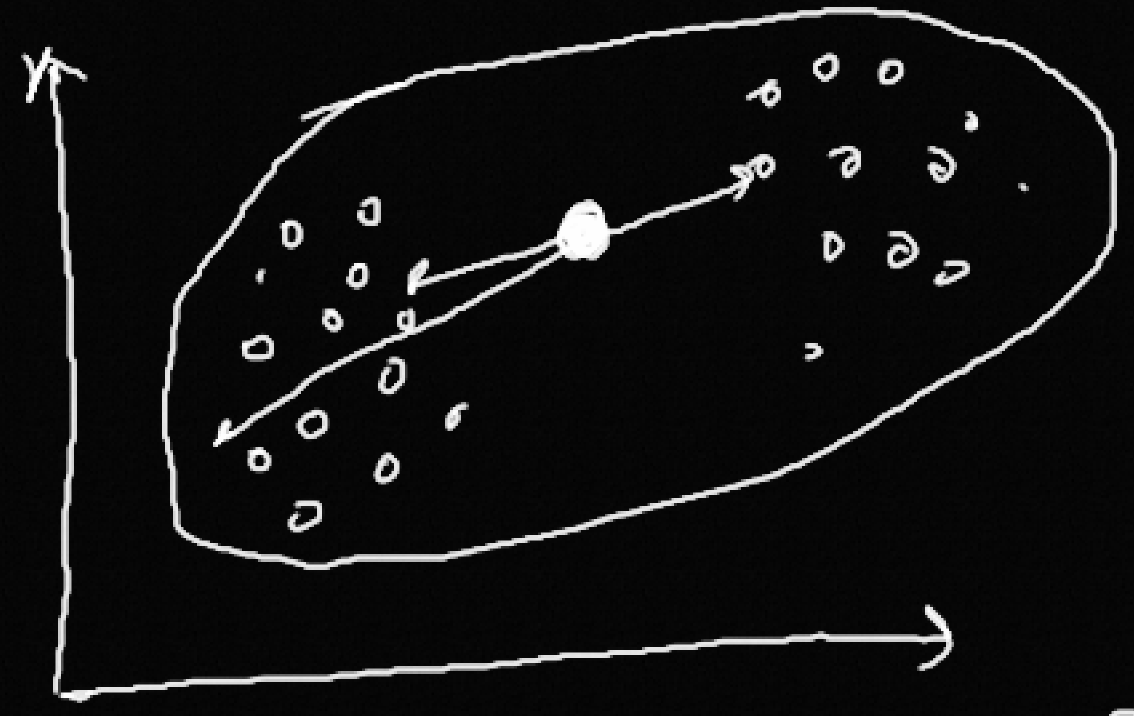
Value

K↑ SSD↓

for $K=1$

$SSD = 106.4$ ⇒ very high

(how far are the points)



Centroid ⟷ points

very high

$K=2$ ⇒ drop greatly.

$SSD = 5.$ good decision

$K=3$ ⇒ drop but very less

$SSD = 2$ not good decision

(Bend formed is the best value)

When to Use

* Large datasets
* Clusters are spherical or of same size
* Continous numerical features (best)
* Fast

When Not to Use

* Irregular clusters.
* Outlier in data

Assumption is biggest disadvantage.

→ fight these problem we have DBSCAN