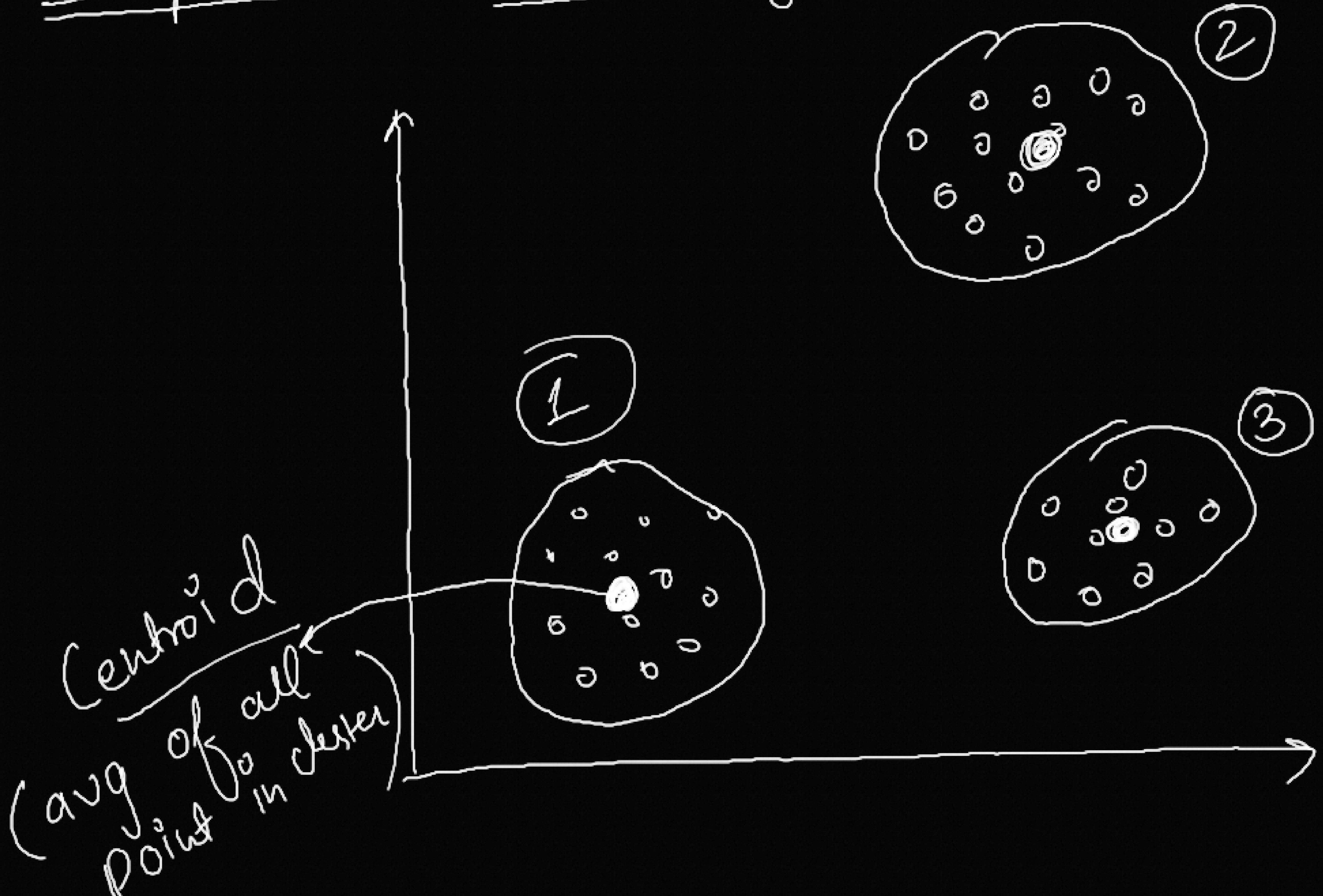
Unsupervised Learning Metrics

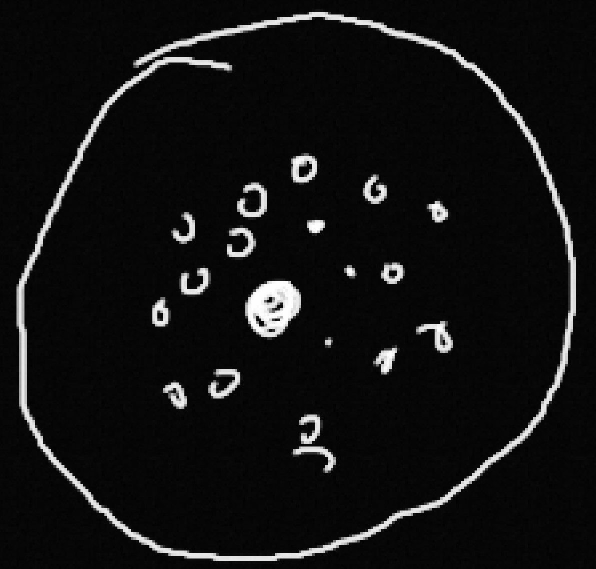
1. Inertia
2. Silhouette Score



① <u>Inertia</u>

How well points are clustered around their centroid. (check quality of cluster)

<u>Low inertia</u> ⇒ Better is quality of cluster
Better is the model
Closer points are to centroid



Mathematical

Sum of squared distance b/w **each** data point & centroid.

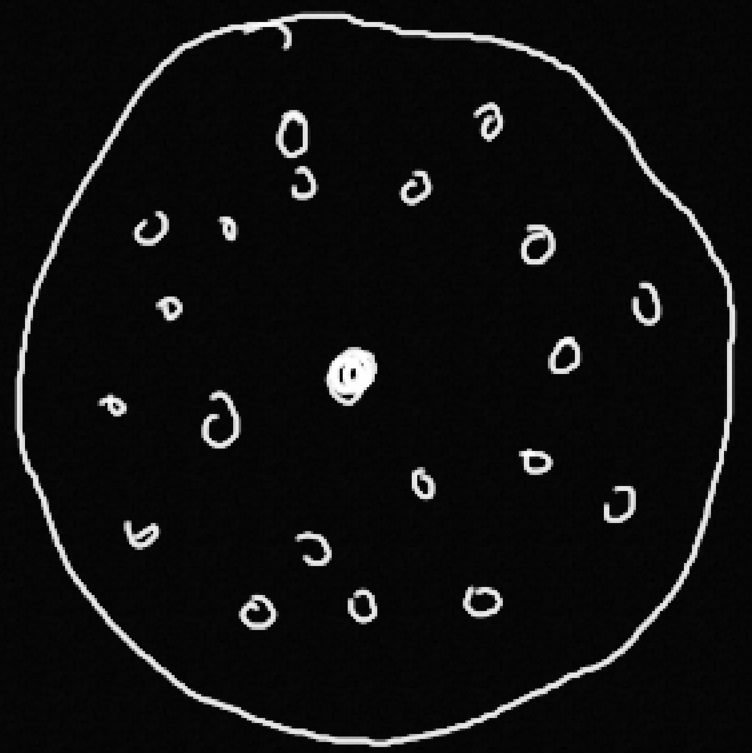
formula

$$\text{Inertia} = \sum \| x_i - \mu_c \|^2$$

x_i ⇒ data point

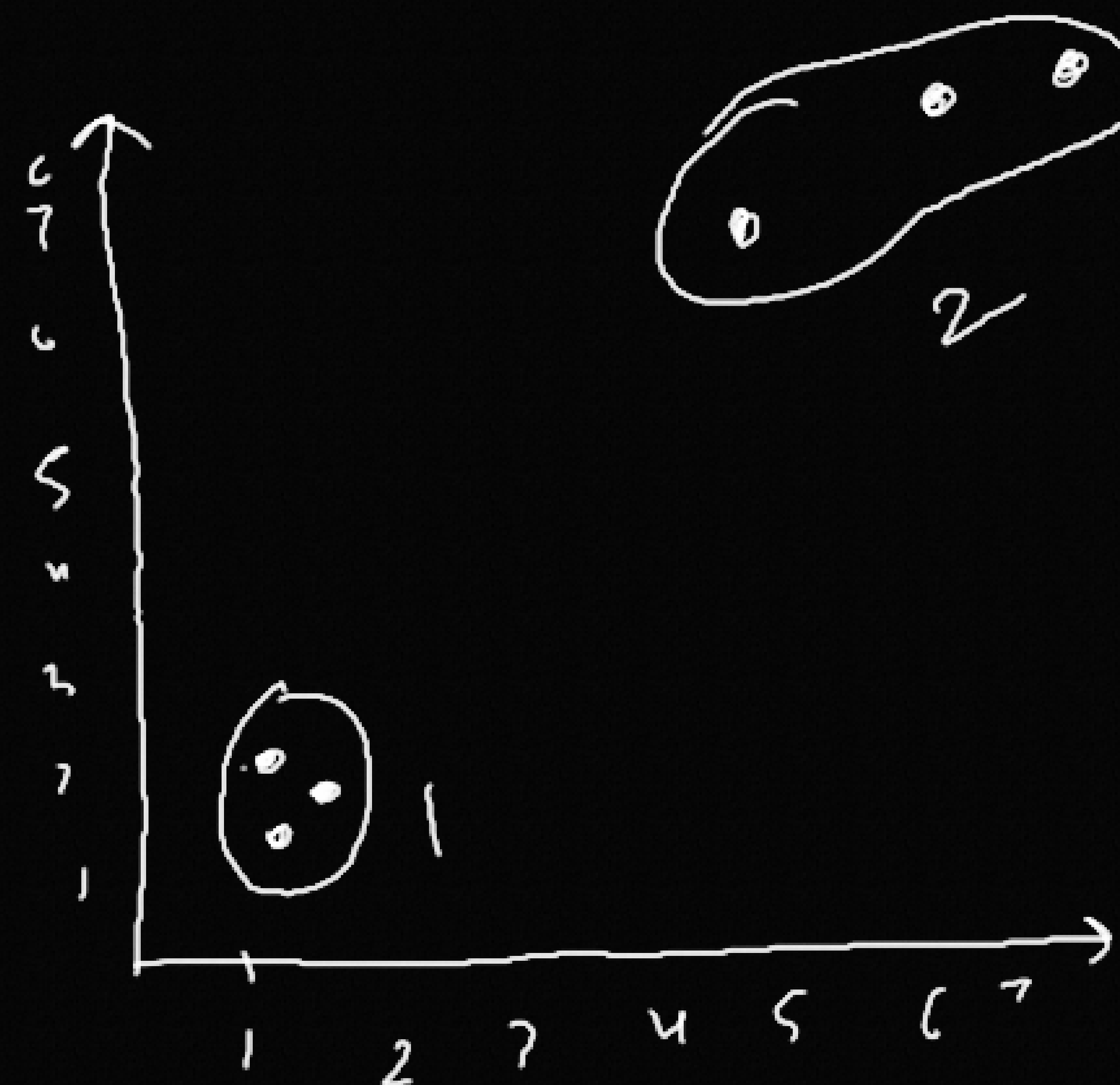
μ_c ⇒ centroid

$\| \ \|$ ⇒ euclidean distance



Example

Data Point	X-co-ordinate	Y-cordinate
P1	1	2
P2	1.5	1.8
P3	5	8
P4	8	8
P5	1.	0.6
P6	9	11



Cluster 1 | **Cluster 2**

P1 $(1,2)$ P2 $(1.5, 1.8)$ P5 $(1, 0.6)$

$$\text{Centroid}_1\ (\mu_{c_1}) = \left(\frac{1+1.5+1}{3}, \frac{2+1.8+0.6}{3}\right)$$

$$= (1.16, 1.46)$$

P3 $(5,8)$ P4 $(8,8)$ P6 $(9,11)$

$$\text{Centroid}_2 = \left(\frac{5+8+9}{3}, \frac{8+8+11}{3}\right)$$

$$= (7.33, 9)$$

Calculate Inertia

① **Cluster 1** :-

P1 & Centroid $= \sqrt{(1-1.16)^2 + (2-1.46)^2}$

(1, 2) (1.16, 1.46) $= \underline{\underline{0.54}}$

P2 & Centroid $= \underline{\underline{0.48}}$

P5 & Centroid $= \underline{\underline{0.87}}$

Sum of squared $= (0.54)^2 + (0.48)^2 + (0.87)^2$

$= \underline{\underline{1.27}}$ ✓ Inertia 1

② Cluster 2 $\Rightarrow$ P3 & Centroid $\Rightarrow$ 2.43

P4 & " $\Rightarrow$ 1.25

P6 & " $\Rightarrow$ 2.69

Sum of Squared distance $= (2.43)^2 + (1.25)^2 + (2.69)^2$

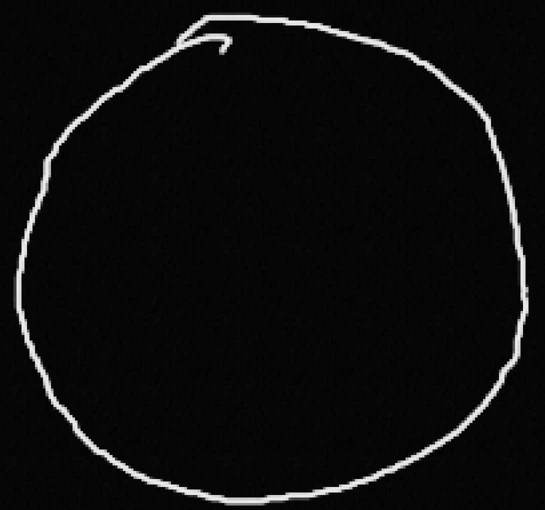
$= 14.16$ Inertia 2

Total Inertia $\Rightarrow 1.27 + 14.16$

$= 15.43$ $\longrightarrow$ (As low as possible)

When not to use Inertia ?

1. Biased towards Spherical clusters
   When you have irregular shape
   inertia may not be best choice.



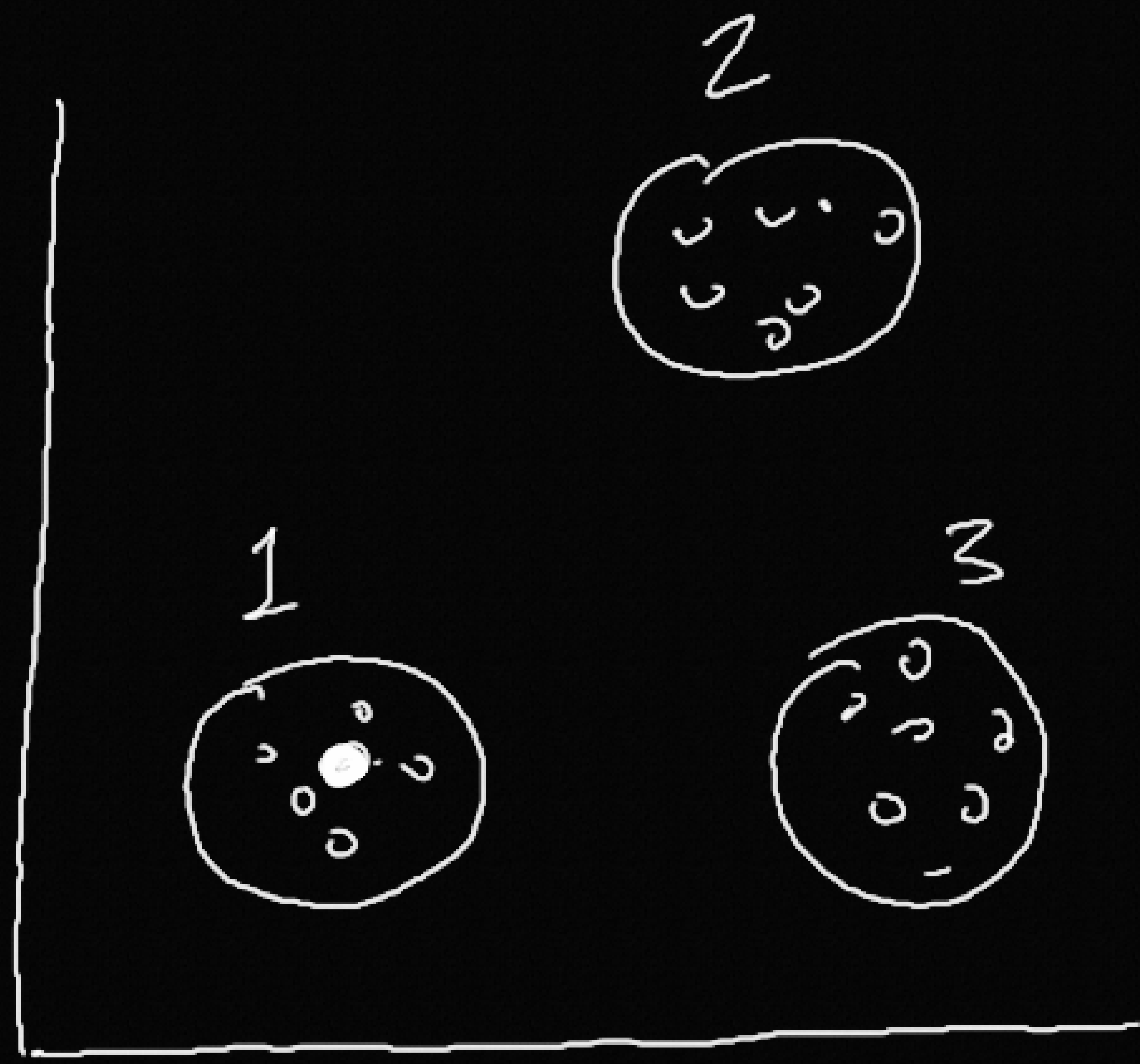
2. Not Scale friendly
   - Weight (Kg) ⇒ 30 - 120
   - Salary (₹) ⇒ 10,000 - 100,000

dominate.

Scaling Techniques

Sihouette Score

* Measures how a data point is fitted in assigned cluster as compared to other clusters.



Range :- -1 to 1

$1 \Rightarrow$ Well clustered (far from other clusters)

$-1 \Rightarrow$ Point is in wrong cluster

$0 \Rightarrow$ B/w two clusters

Two factors

Intra cluster distance (a)

Avg distance b/w one point & all other point in same cluster.

Inter cluster distance (b)

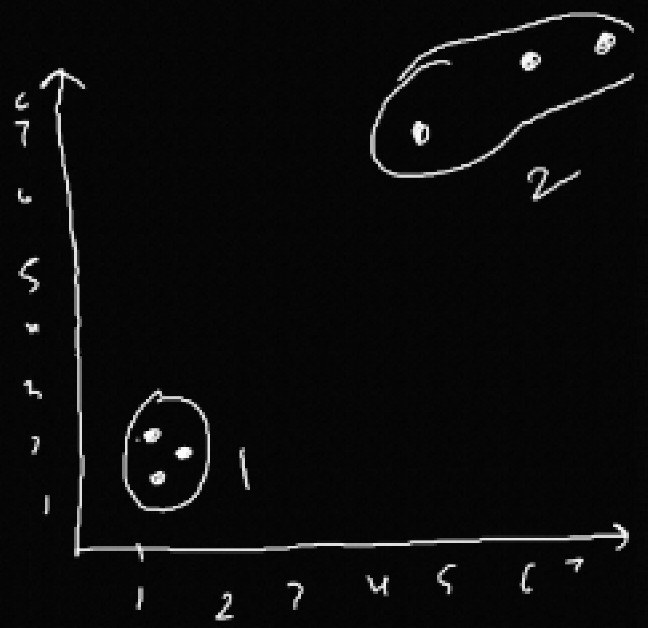
Avg distance b/w point & all other point in nearest cluster that point does not belong

Silhouette Score

$$\text{SH score} = \frac{b - a}{\max(a, b)}$$

Example

Data Point	X-co-ordinate	Y-cordinate
P1	1	2
P2	1.5	1.8
P3	5	8
P4	8	8
P5	1	0.6
P6	9	11



Cluster 1	Cluster 2
P1 P2	P3 P4
P5	P6

Taking only 2 points.

① Calculate a

for P1 (cluster 1) :- P1 & P2 ⇒ 0.54
P1 & P5 ⇒ 1.4

$$a(P1) = \frac{0.54 + 1.4}{2} = 0.97$$

for P3 (cluster 2) :- P3 & P4 = 3 P3 & P6 = 5

$$a(P3) = \frac{3+5}{2} = 4$$

② Calculate b

for P1 (cluster 1) :- P1 & P3 ⇒ 7.21
P1 & P4 ⇒ 9.22
P1 & P6 ⇒ 12.04

$$b(P1) = \frac{7.21 + 9.22 + 12.04}{3} = 9.49$$

for P3 (cluster 2) :- P3 & P1 ⇒ 7.21
P3 & P2 ⇒ 7.01
P3 & P5 ⇒ 8.27

$$b(P3) = 7.5$$

③ Calculate Silhouete score

$$\text{formula} = \frac{b-a}{\max(a,b)}$$

for P1

$$\text{SH score} = \frac{9.49 - 0.97}{\max(0.97, 9.49)} = \frac{8.52}{9.49} = \underline{\underline{0.9}}$$

for P3.

$$\text{SH score} = \frac{7.5-4}{\max(4,7.5)} = \frac{3.5}{7.5} = \underline{\underline{0.47}}.$$

(4) Once you get all SH scores,

Then

Do average of all the SH scores

You will get final **Silhouette Score**

Limitations

1. Not good with irregular shapes
   But better than inertia.

2. Computational Cost. very high.