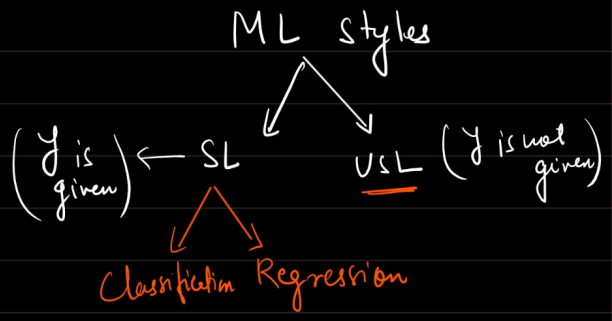
Introduction to Unsupervised learning

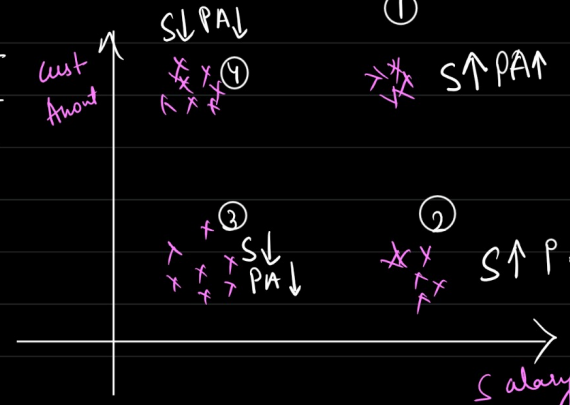
* Target variable (Y) is Not given

* USL groups / segments your data (Finds pattern in the data)



Examples Zara Store

Customer purchase amount	Customer Salary in Lakhs
2500	10
—	—
—	—
—	—
—	—



Clothes mix up

Shop Keeper is blind → will try to touch the cloth and understand the pattern, different rocks

To increase revenue → Group 2 customers should be focussed more.

→ Group 1 → loyalty card & Extra discount.

* depends on you how you want interpret the groups and bring business.

Motivation

→ Need to launch a campaign based on income / sales.
→ Different ways of fraud / money laundering.
→ Grouping of images / patterns of specific image
→ Document / Article Analysis.
→ Cohort Analysis.

* In higher dimensions, you cannot identify groups / patterns manually. Therefor you need USL.

USL

① K-means clustering
② Hierarchical clustering
③ DBSCAN

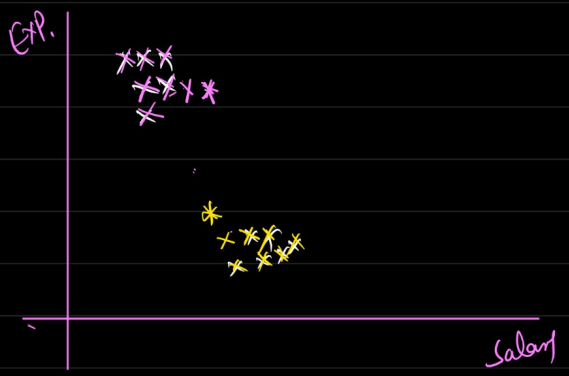
* K means clustering

Steps of K means clustering

Step-1 intialize Centroid (K)
- → can be random dp
- → Centroid of all the dp
- → A rrandom dp

Step-2 → Points nearer to the centroid will be labelled as that group

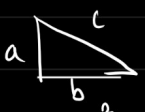
Step-3 Re-Calculate the centroid ↳ Avg.
Again repeat Skp 2 untill Centroid doesnt change.



Exp	Salary
2	5
3	6
5	7

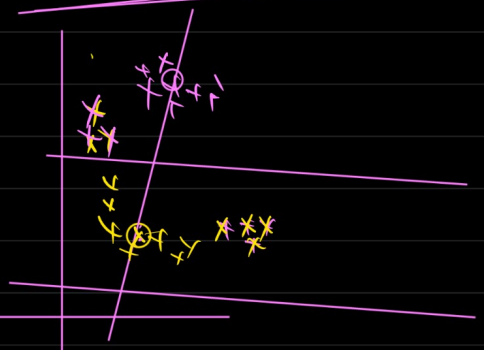
$\left(\frac{2+3+5}{3} \quad \frac{7+6+5}{3}\right)$

( )



$c^2 = a^2 + b^2$

⟹ after recalculating Centroid



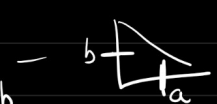
*Distance

① Euclidean distance

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

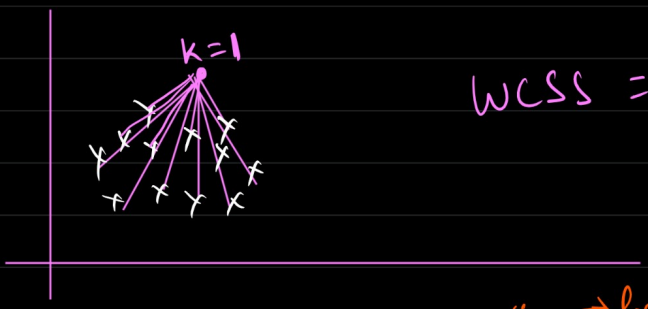
② Manhattan distance

$$\text{Manhattan distance} = |x_2-x_1| + |y_2-y_1|$$



How to Select the k value

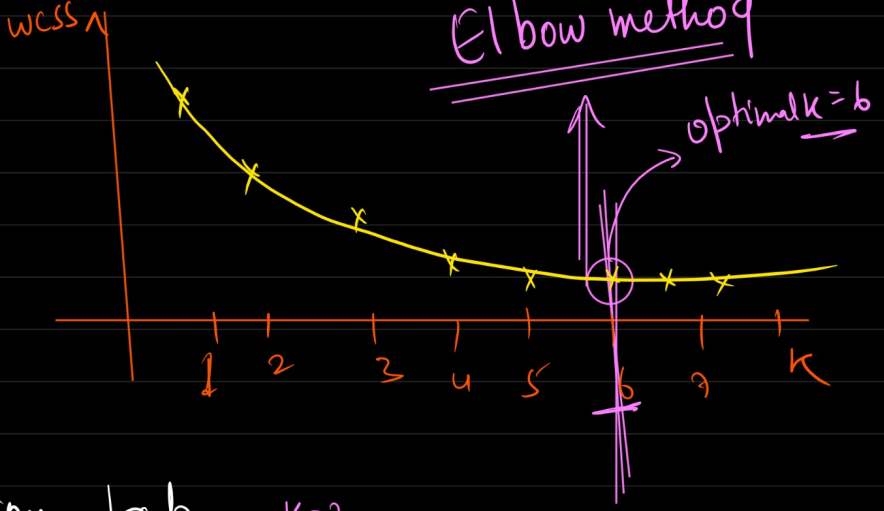
WCSS → Within Cluster Sum of squares distance



$$WCSS = \sum_{i=1}^{n} \left(\text{distance b/w point to nearest Centroid} \right)^2$$

⇒ for k=2 → As you increase k, wcss decreases
→ And at optimal k, wcss stops changes.

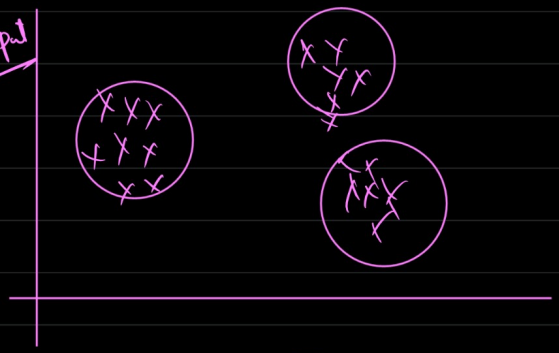
Elbow method



* Random Initialization trap

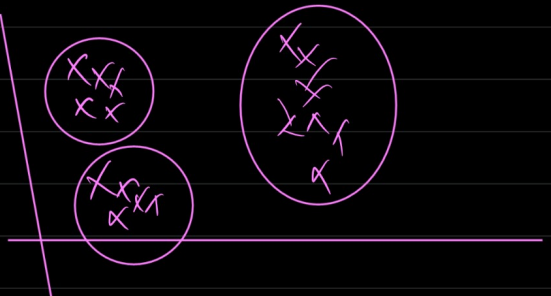
K=3

Kmeans →



Another possibility

Kmeans →



* The final cluster depends on the initialization of k.

⇓

Kmeans + +

⇓

initialize the k's as much far as possible