Machine Learning

{enable the m/c to learn pattern available in the data}

- Supervised learning (Target)
  - Regression (numerical)
  - Classification (Discrete)
- Unsupervised learning (No Target)
  - Clustering

→ Linear Regression :-

* Supervised algorithm
* Target column is numerical in nature.
* X (independent) & Y (dependent) have linear relationship.

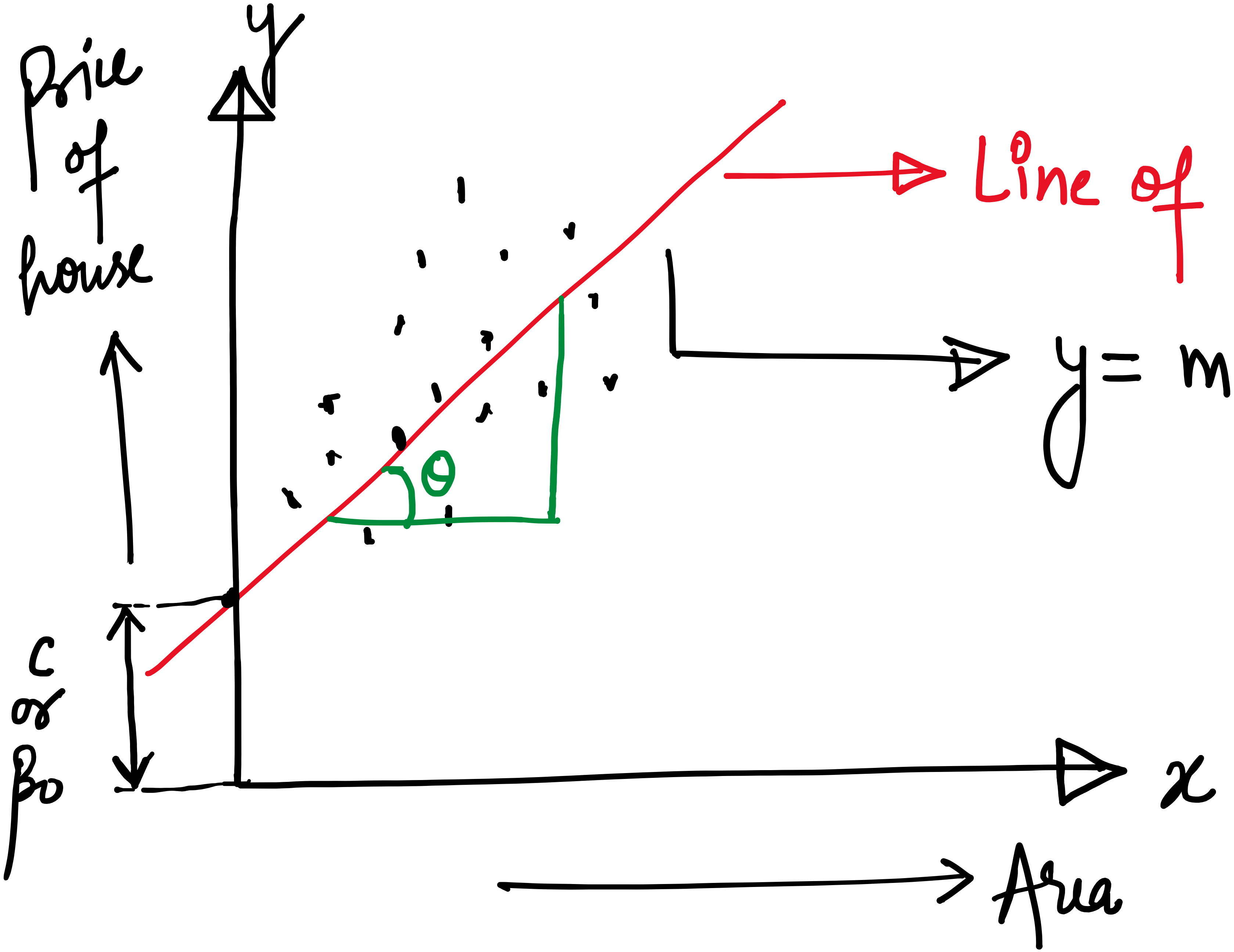
SLR (X=1)

$$y = mx + c$$

MLR (X>1)

n features

$$y = m_1x_1 + m_2x_2 + \dots + m_nx_n + c$$



Line of best fit

$y = mx + c$ or $\beta_0 + \beta_1 x$

m or β_1 → slope

c or β_0 → intercept

→ Evaluation Metrics:-

(i) R^2

$0 \le R^2 \le 1$

Higher the better

$$R^2 = 1 - \frac{RSS}{TSS}$$

RSS = Residual sum of squares (Error)

TSS = Total sum of squares

(ii) Adjusted R^2

$0 < Adj\ R^2 < 1$

Higher the better

$$Adj\ R^2 = 1 - \frac{(1-R^2)(N-1)}{(N-P-1)}$$

N = No. of data points

P = no. of features

→ Cost f^n :- Mathematical function needs to be minimised to find the line of best fit. or to find the optimal values of m & c.

$$C = \frac{RSS}{N}$$

*

$$\rightarrow \boxed{J = \frac{1}{N} \sum \left[y_a - (mx_i + c) \right]^2}$$

→ Gradient Descent Algorithm :- (Iterative Algorithm)

1. Initialize the value of m & c as zero.
2. Calulate the cost function.
3. Update the value of m & c

$$m' = m - \alpha \frac{\partial J}{\partial m}$$

$$c' = c - \alpha \frac{\partial J}{\partial m}$$

Error Gradients

4. Repeat step 2 & 3 untill we got the min error.

$\alpha \rightarrow$ learning rate

→ Modelling Steps:-

1. Data loading & data understanding.
2. Cleaning if required
3. EDA
4. Train test split
5. Missing value imputation

6. Outlier Treatment

7. Scaling

8. Feature Selection
   - RFE
   - p value (< 0.05)
   - VIF (≤ 5)

9. Training

10. Evaluation on test data

→ Logistic Regression :-

* Supervised algorithm (Classification)
* Target variable is discrete in nature.

  ↓

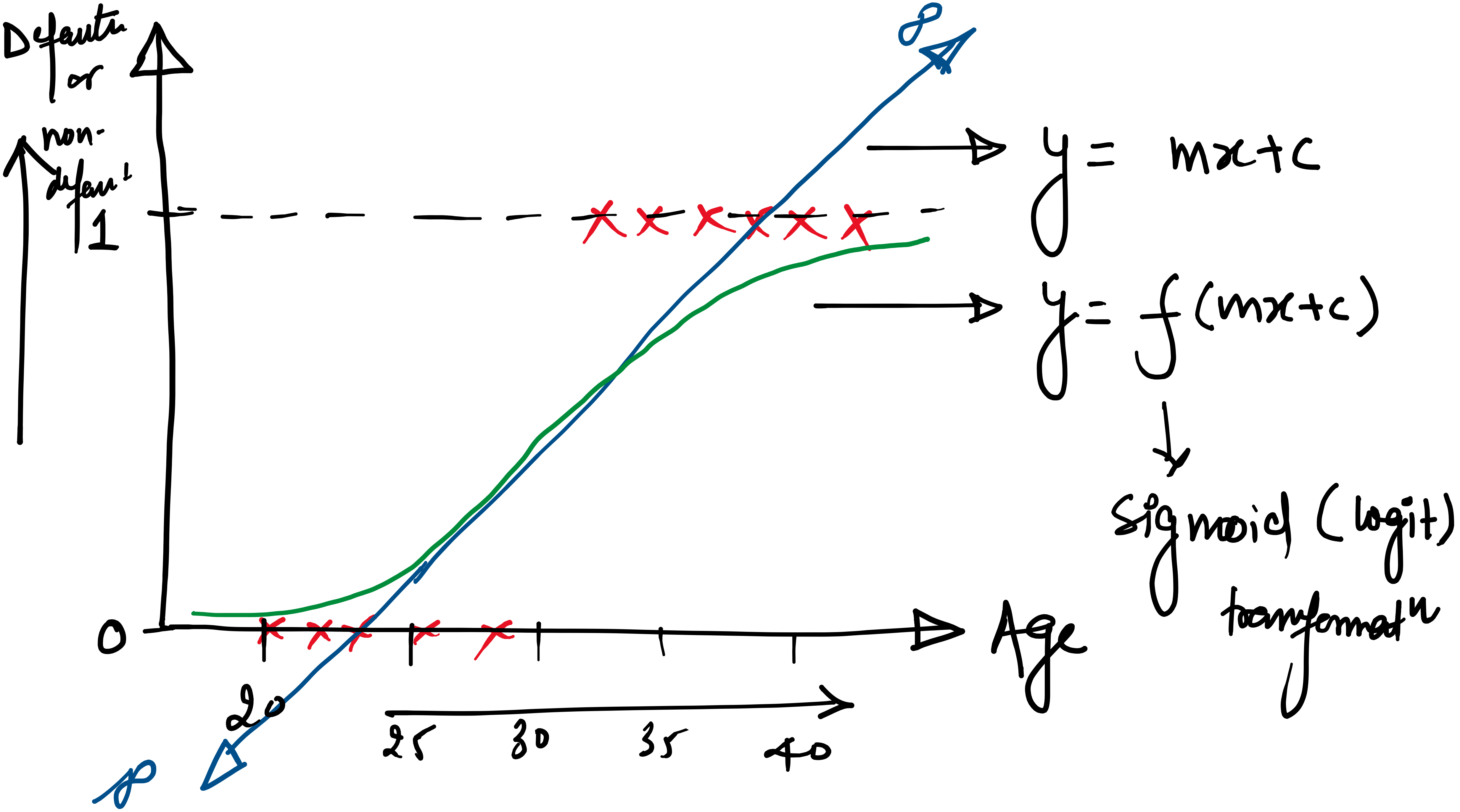
  Categorical

→ Binary Classifier

2 categories (Target)

→ Multi class classification

>2 categories (Target)



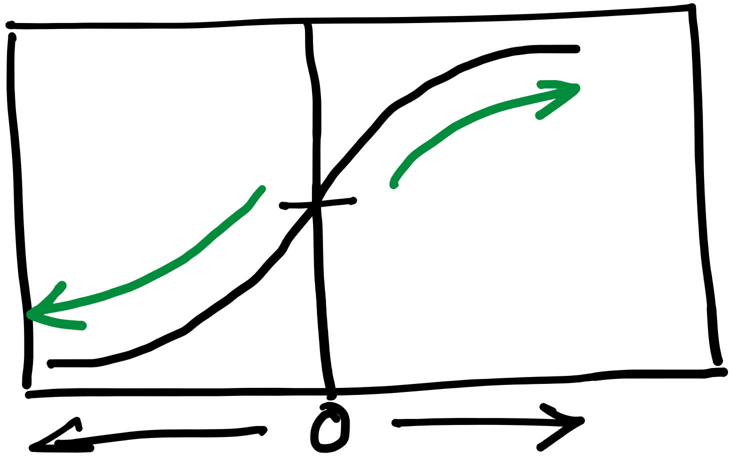
Default or non-default
1
0
Age
20
25
30
35
40
∞
∞
y = mx+c
y = f(mx+c)
sigmoid (logit) transformatn

Logistic Regression = Logit f^n + Regression
(Sigmoid)

$$f(x) = \frac{1}{1+e^{-x}}$$

* $0 \leq f(x) \leq 1$

* S-Shaped Curve



→ Evaluation Metrices :-

1 → +ve
0 → -ve

* Confusion Metrics :-

		Predicted +	Predicted -
Actual	+	TP	FN
	-	FP	TN

TP → True Positive ↑

TN → True Negative ↑

FP → False positive ↓

FN → False negative ↓

① Accuracy:

$$A = \frac{TP+TN}{TP+TN+FP+FN}$$

② Recall: $R = \frac{TP}{TP+FN}$

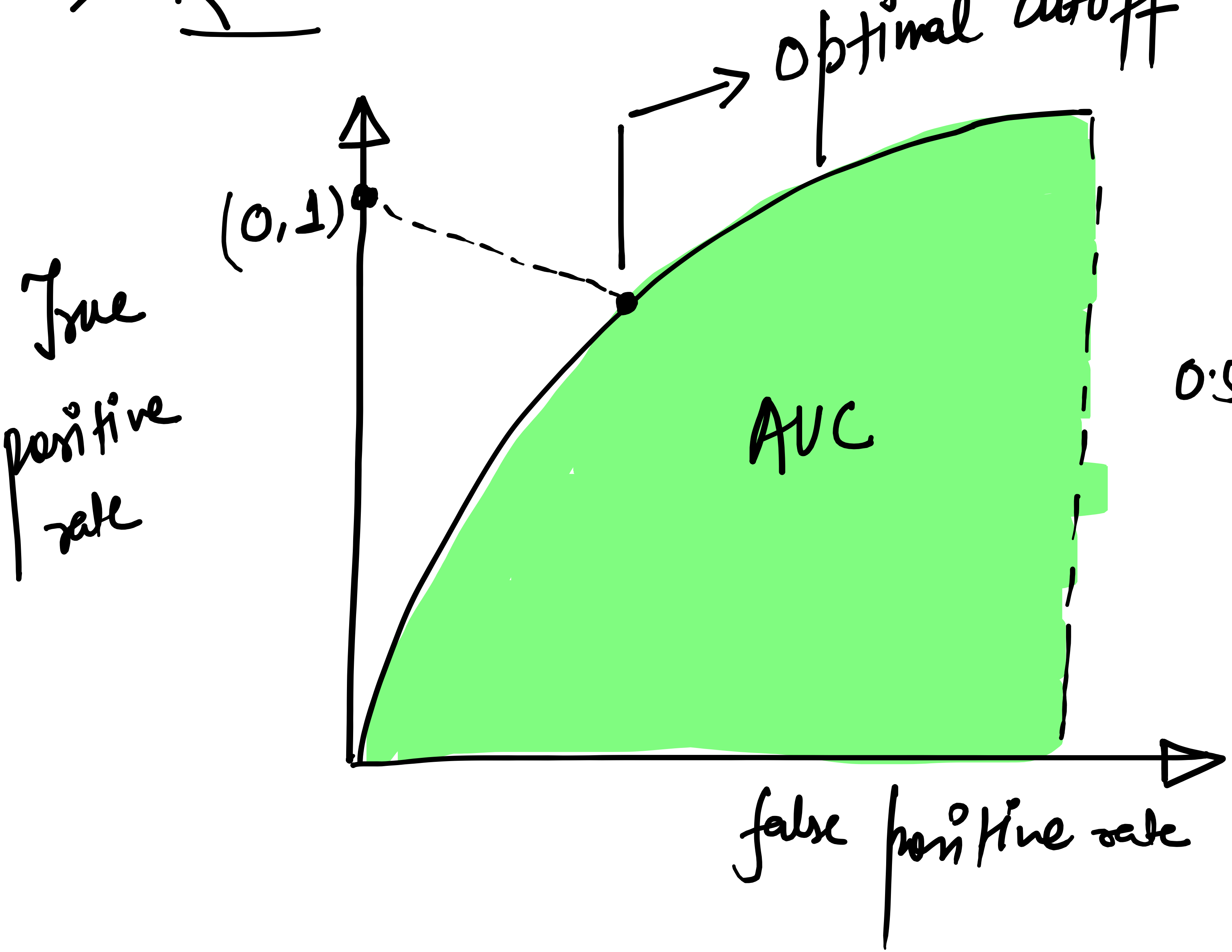
③ Precision:

$$P = \frac{TP}{TP+FP}$$

④ F1-Score: (Harmonic mean of Precision & Recall)

$$F_1 = \frac{2 \times P \times R}{P+R}$$

→ ROC :-



→ optimal cutoff

→ AUC (Area Under Curve

$0.5 \leq AUC \leq 1$

* various cutoffs

→ Cost function :-

* Maximum likelihood Estimation (MLE)

$$\left\{ MLE = Max\left[\underbrace{(1-p_0)(1-p_1)(1-p_2)}_{y=0} \cdots \underbrace{p_{n-1}\, p_n}_{y=1} \right] \right.$$

Clustering

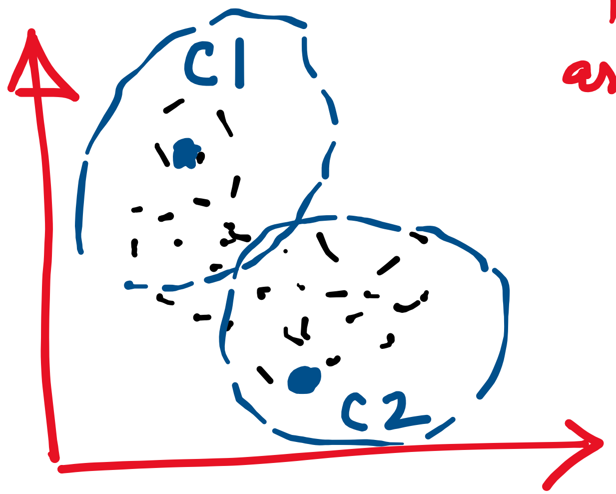
- K-Means Clustering
- Hierarical Clustering

* Unsupervised algorithm

* In K-means, no. of clusters to be pre known but thats not the scenario for hierarical clustering

→ K-Means Clustering :- K=2

C1 = Mean of data assigned to it



① Random Initialization ✓

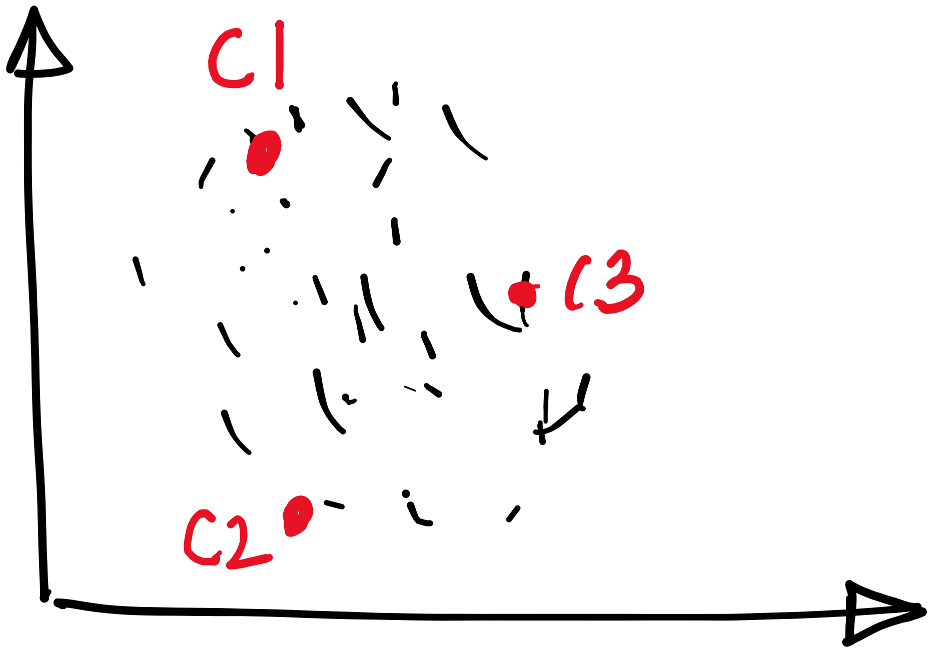
② Assignment

③ Updation

→ improvised version → K-means++

④ Repeat ② & ③ untill centroids become stable

K=3

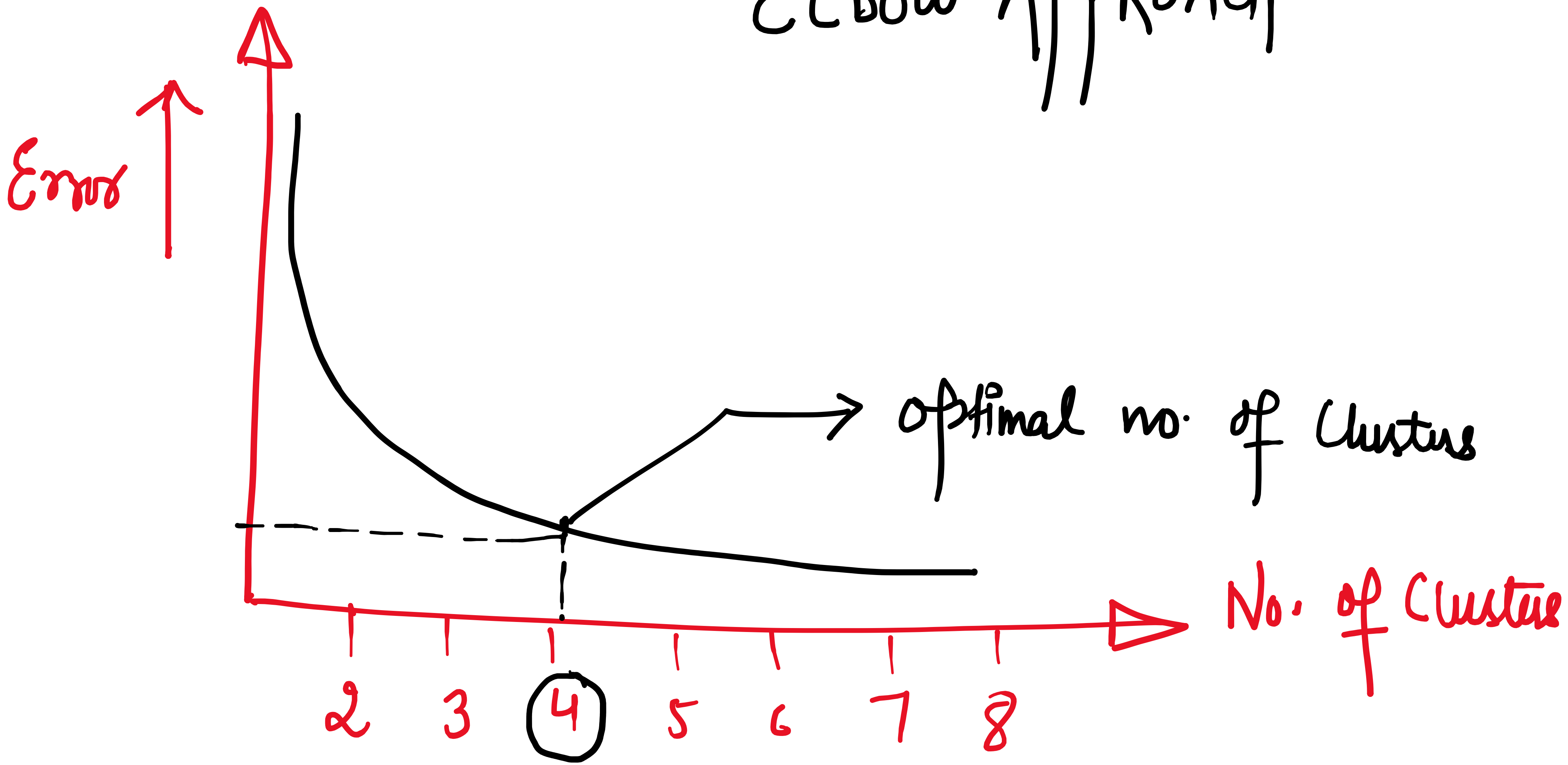


C3

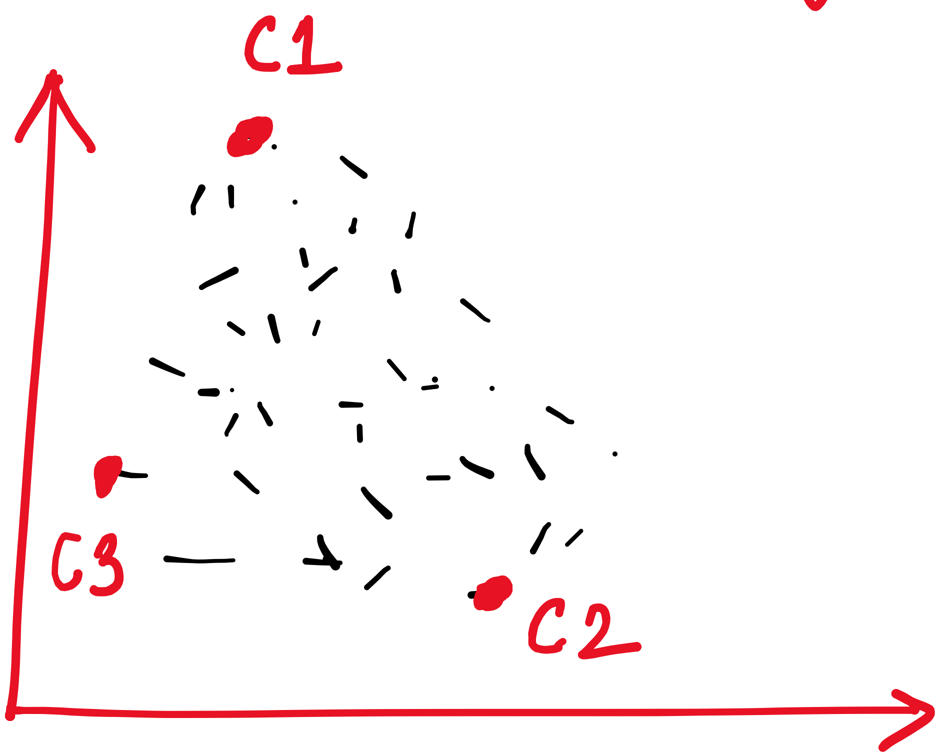
K=3

→ <u>optimum no. of</u> <u>Clusters</u> :-

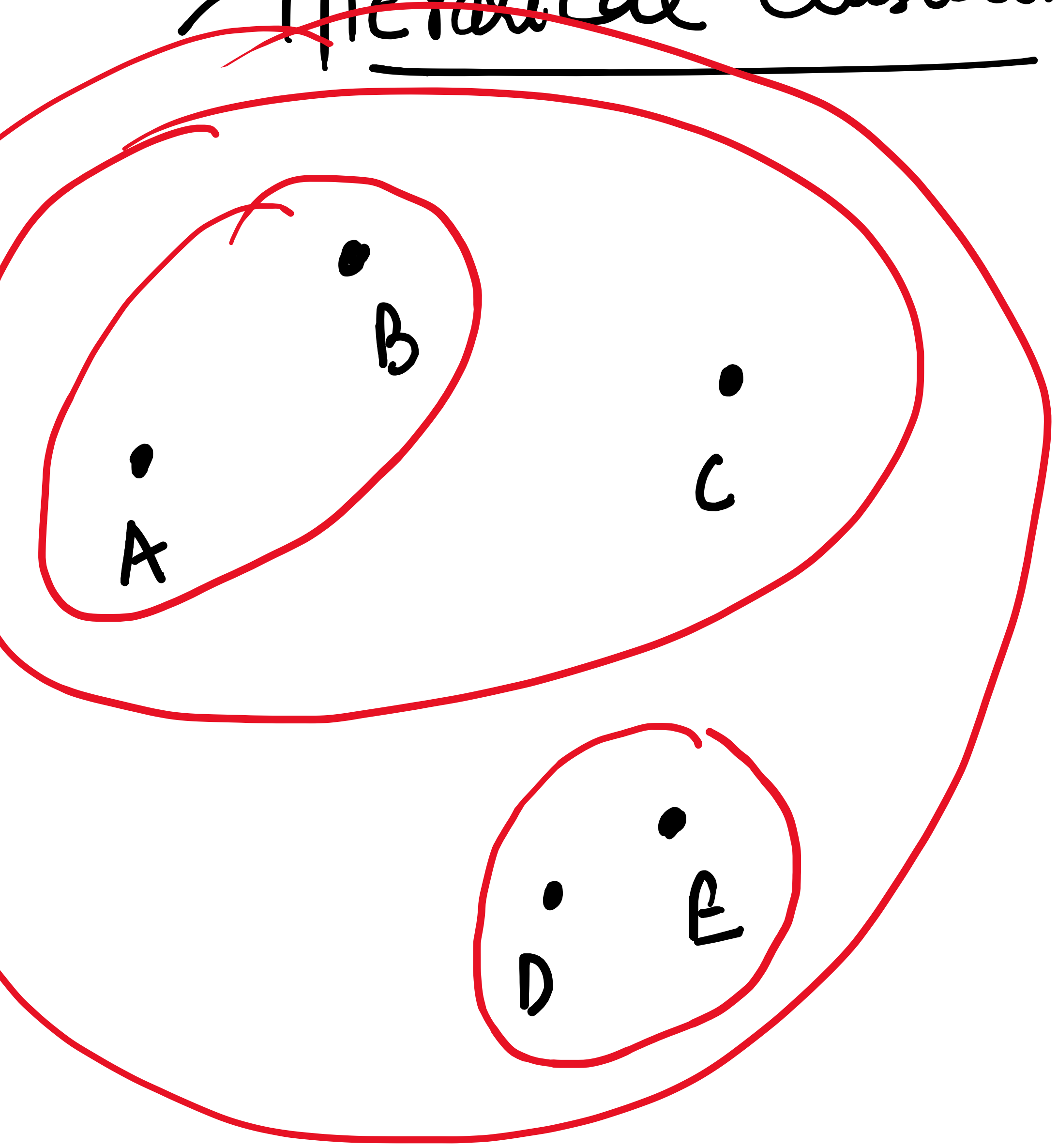
ELBOW APPROACH



→ In K-means ++ centroids are randomly initialised from the dataset only.

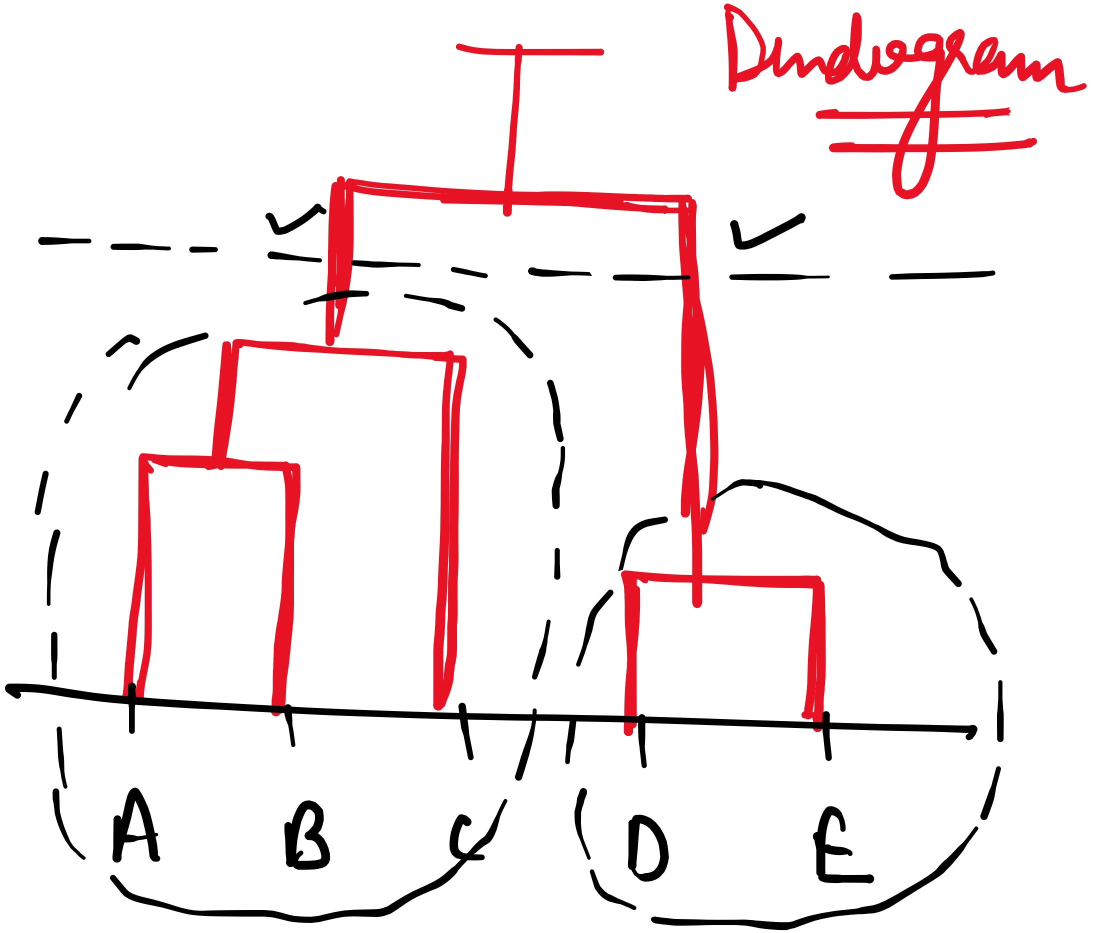


→ Hierarical Clustering :-

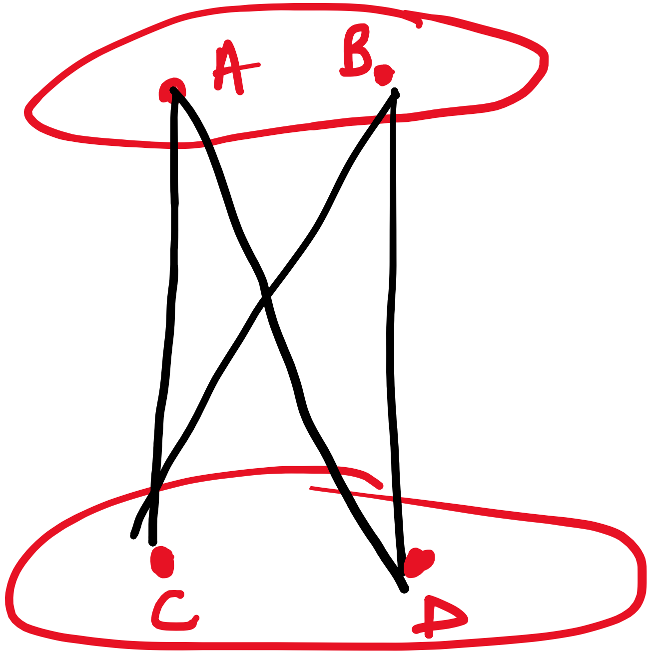


$K = 2$

	AB	C	DE
AB		−	−
C			−



→ Distance b/w two Clusters :-



3 linkages :-

* Average : Average
* Complete : Max
* Single : Min

Av(AC, AD, BC, BD)

Max(AC, AD, BC, BD)

Min (AC, AD, BC, BD)