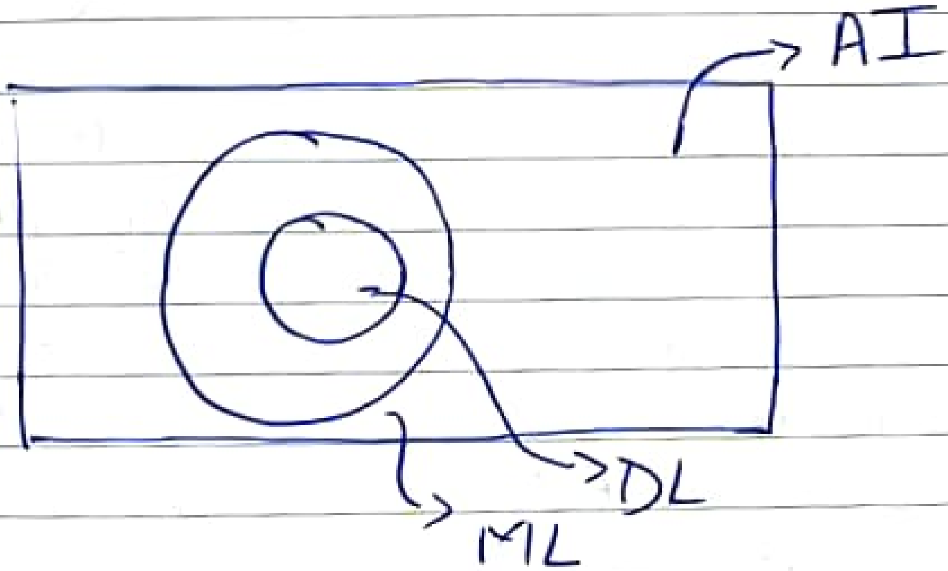
MACHINE LEARNING



-> AI application is able to do it own task w/o any human intervention.

UNSUPERVISED
↳ CLUSTERING
↳ In clustering we don't have target variable.
↳ We make cluster.

→ Clustering is a grouping Algo.

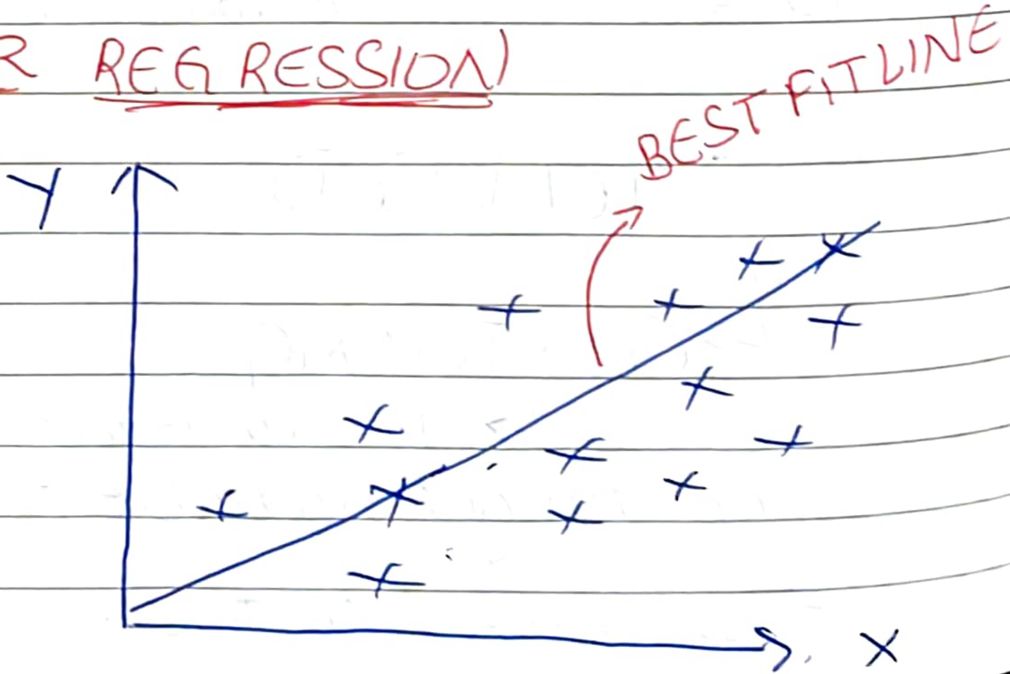
SUPERVISED LEARNING

(I) Linear Regression
(II) Ridge & Lasso
(III) Logistic Regression
(IV) Decision Tree
(V) ADA BOOST
(VI) Random Forest
(VII) Gradient Boosting
(VIII) XgBoost
(IX) Naives Bayes

UNSUPERVISED LEARNING

(I) K-Mean
(II) DBScan
(III) K-Nearest Neighbour Clustering
(IV) PCA
(V) LDA

(1) LINEAR REGRESSION



Best Fit line ⇒ $y = mx + c$

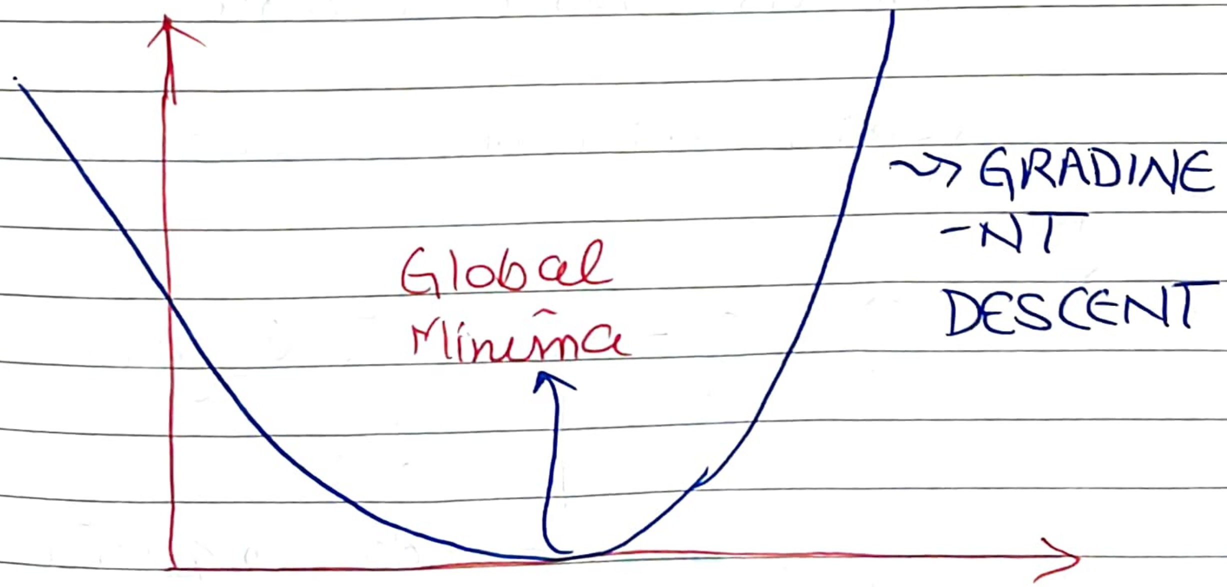
↳ LINEAR.

Cost Function / Square Error Function

↳ $\frac{1}{2m} \sum_{i=1}^{m} (\text{predicted value} - \text{actual value})^2$

$2m$ ↳ no. of datapoints

→ Aim → is to minimize cost function



Cost Function

Choose the Best Fit Line with Minimum Cost Function

CONVERGENCE ALGO / GRADIENT DESCENT ALGO

↳ Help you to come to global minim[a] in Gradient Descent.

Repeat until convergence

{

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

↳ Learning Rate.

}

→ Cost Function we use give global minima even if we have local minima

Predicted ⇒ $h_\theta(x) = \theta_0 + \theta_1 x.$

For $j=0$

$$\frac{\partial}{\partial \theta_0} J(\theta_0,\theta_1) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{i}) - y^{(i)})$$

For $j=1$

$$\frac{\partial}{\partial \theta_1} J(\theta_0,\theta_1) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{i}) - y^{(i)})x^{(i)}$$

CONVERGENCE ALGO (FINALLY)

Repeat until convergence

$$\{ \quad \theta_0 := \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})$$

$$\theta_1 = \theta_1 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x$$

$\}$

PERFORMANCE MATRIX

↳ We use this to check how good our model is.

R^2 & Adjusted R^2

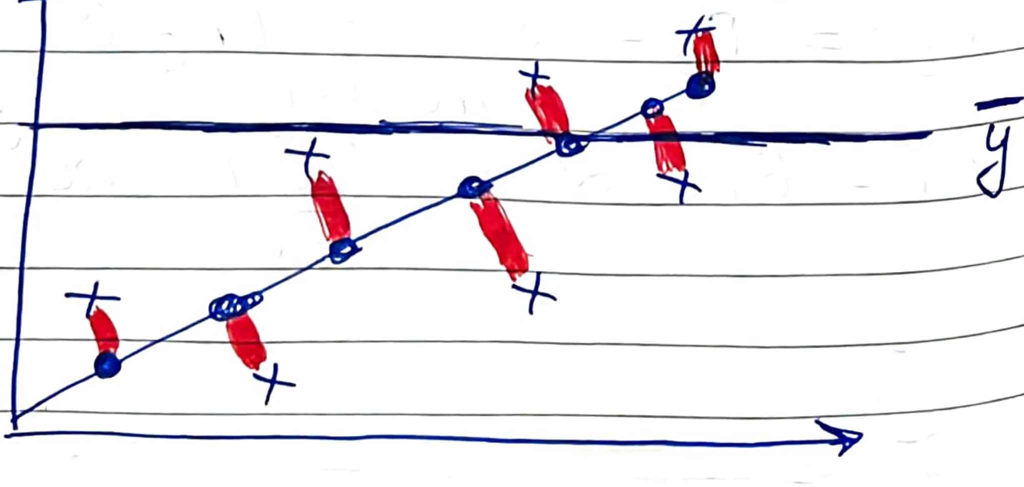
R^2

$h_\theta(x)$

$$\boxed{R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}}$$

↳ mean

$$R^2 = 1 - \frac{\text{Sum of Residual (1)}}{\text{Sum of Total}}$$



$R^2 \in (0,1)$

$R^2 \to 1 \to$ Indicates Better Fit.
$R^2 \to 0 \to$ Indicates Lack of Linear Relationship.

ADJUSTED R^2

$$\boxed{R^2_{adjusted} = \frac{1-(1-R^2)(N-1)}{N-p-1}}$$

$N \to$ No. of Samples
$p \to$ No. of Features

Q Between R^2 & Adjusted R^2 whose value will be Bigger?

$\to$ R^2 value will be bigger.

Date
Page

→ From R^2 & Adjusted R^2, Adjusted R^2 is preferred

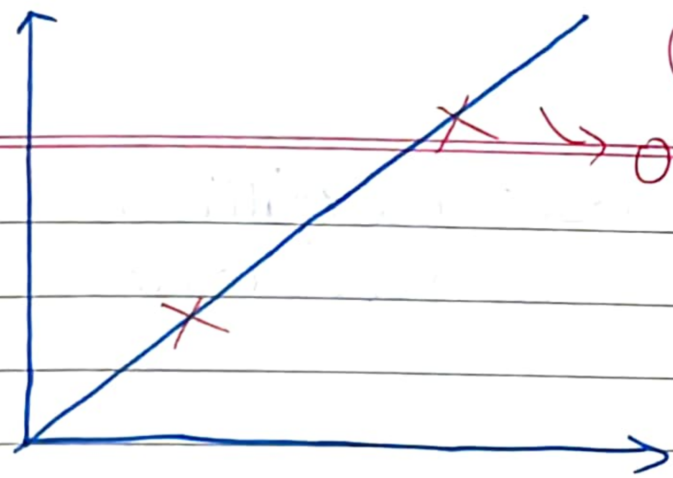
Adjusted R^2 prevent model from overfitting as if we increase no. of features it will automatically decrease Adj. R^2 value significantly if it is not important.

② RIDGE & LASSO REGRESSION

↳ TO PREVENT OVERFITTING.

Overfitting -> Model perform wells with Training Data (LOW Bias)
Fails to perform with Test Data. (High ~~performance~~ variance)

Underfitting -> Model perform bad with Training Data (HIGH BIAS)
Model perform bad also with Test Data
(HIGH VARIANCE)



↳ OVERFITTING

→ We need Generalise Model

↳ Low Bias

↳ Low Variance

RIDGE (L2 REGULARIZATION)

↳ $(\hat{y} - y^i)^2 + \lambda\,(slope)^2$

↓ $h_\theta(x)$

↳ Hyperparameter

LASSO (L1 REGULARIZATION)

↳ $(\hat{y} - y)^2 + \lambda\,|slope|$

↳ Feature selection

$|Slope| = \cancel{x_1 + x_2 +}\ |\theta_0 + \theta_1 + \theta_2 + \dots + \theta_n|$

Date
Page

LASSO

(I) Prevent overfitting
(II) Feature selection

$\lambda \rightarrow$ Cross-Validation

ASSUMPTION OF LINEAR REGRESSION

(I) Normal (Gaussian Distribution)

(II) Standardization $\rightarrow$ z-score

(III) Linearity.

(IV) Multi-collinearity.

(3) LOGISTIC REGRESSION

(CLASSIFICATION)

$\rightarrow$ Only Two Outcome

Logistic Regression → Sigmoid Function

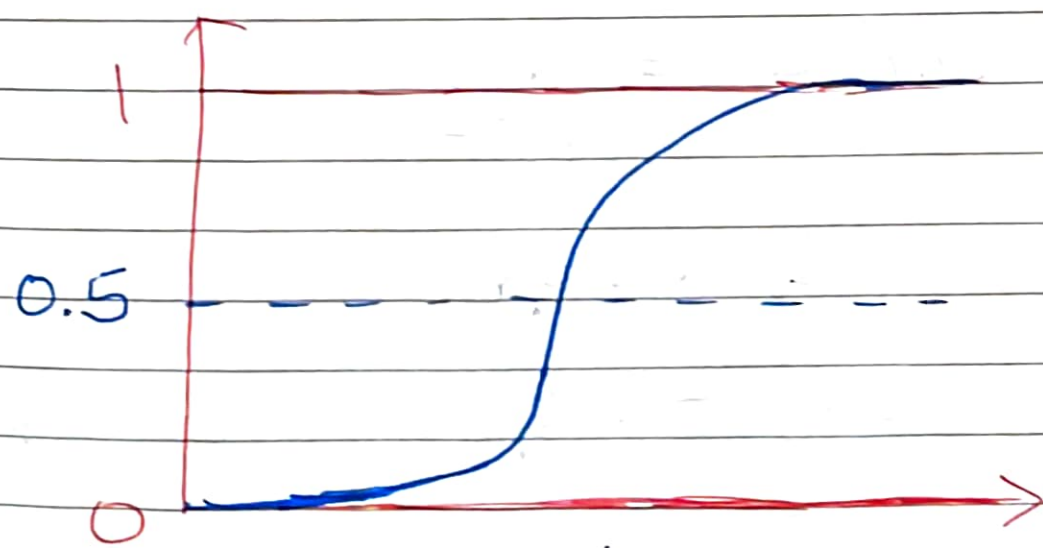
Sigmoid function

$$y = \frac{1}{1 + e^{-(a_0 + a_1 x)}}$$

$\downarrow$

$(0, 1)$

a_0 = Intercept.
a_1 = slope.



classmate
Date
Page

DECISION BOUNDARY LOGISTIC REGRESSION

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x_2 + \dots + \theta_n x_n$$

$$\boxed{h_\theta(x) = \theta^T x}$$

TRAINING SET

$$\{ (x^1, y^1), (x^2, y^2), \dots (x^n, y^n) \}$$

$$y \in \{0, 1\}$$

$$h_\theta(z) = \frac{1}{1 + e^{-z}}$$

$\hookrightarrow \theta_0 + \theta_1 x$

let our $z = \theta_1 x$ (intercept at 0)

Cost Function

Linear Regression $J(\theta_1) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} (h_\theta(x^i) - y^i)^2$

Logistic Regression

$$h_\theta(x) = \frac{1}{1+e^{-(\theta, x)}}$$

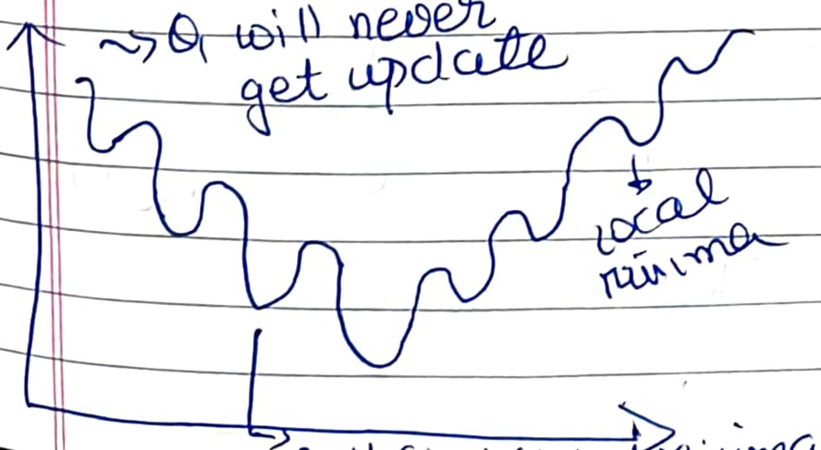
Cost Function

$$\hookrightarrow \cancel{\frac{1}{2}} \boxed{\frac{1}{2}\left(h_\theta(x^i) - y^{(i)}\right)^2}$$

$$h_\theta(x) = \frac{1}{1+e^{-(\theta, x)}}$$

We can't use this cost function $\because$ $h_\theta(x)$ is convex function
↳ not

NON-CONVEX FUNCTION



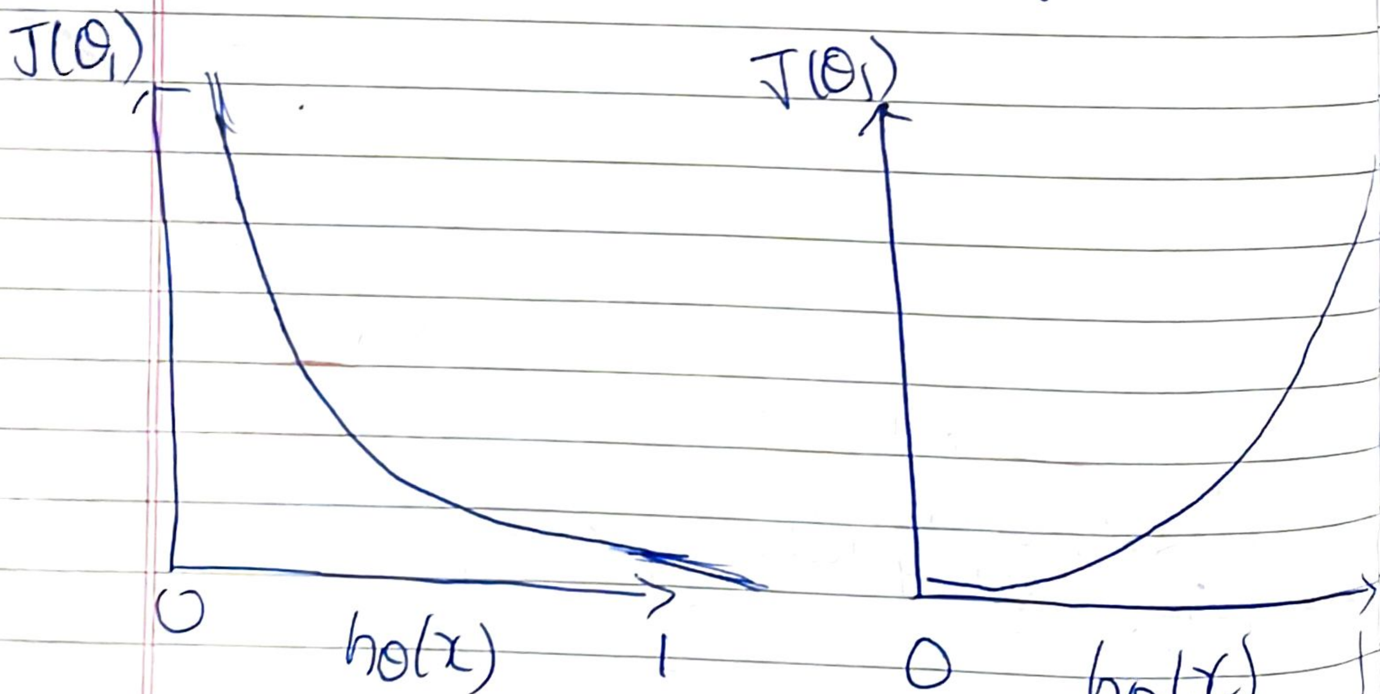
CONVEX FUNCTION

COST FUNCTION LOGISTIC REGRESSION

$$J(\theta_1) = \begin{cases} -\log(h_\theta(x)) \; ; & y=1 \\ -\log(1-h_\theta(x)) \; ; & y=0 \end{cases}$$

$$h_\theta(x) = \frac{1}{1+e^{-(\theta_1 x)}}$$

if $y=1$ if $y=0$



Classification

classmate

Date

Page

$$\boxed{\text{Cost}(h_\theta(x^i), y) = -y\log(h_\theta(x^i)) - (1-y)\log(1-h_\theta(x))}$$

if $y=1$

$$\text{Cost}(h_\theta(x^i), y) = -\log(h_\theta(x^i))$$

if $y=0$

$$\text{Cost}(h_\theta(x^i), y) = -\log(1-h_\theta(x^i))$$

FINAL COST FUNCTION

$$\boxed{J(\theta_1) = \frac{-1}{2m}\sum_{i=1}^{m}\left(y^i\log(h_\theta(x^i)) + (1-y^i)\log(1-h_\theta(x^i))\right)}$$

$$\boxed{h_\theta(x^i) = \frac{1}{1+e^{-\theta_1 x}}}$$

Repeat until convergence

$$\left\{ \theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} (J(\theta_i)) \right.$$

}

Q How it is different from Logisti Linear Regression?

PERFORMANCE MATRIX (classification Problem)

Features		y (Actual)	$\hat{y}$ (Predicted)
–	–	0	1
–	–	1	1
–	–	0	0
–	–	1	1
–	–	1	1
–	–	0	1
–	–	1	0

i) CONFUSION MATRIX

PREDICTED \ ACTUAL	1	0
1	3	2
0	1	1

PREDICTED \ ACTUAL		
	TP	FP
	FN	TN

$$\boxed{\underline{\text{ACCURACY}} = \frac{TP + TN}{TP + FP + FN + TN}}$$

→ Tells No. of Times you actually predicted correct

→ Biased Dataset can affect the accuracy.

(II) $\boxed{\text{Precision} = \frac{T.P}{T.P + F.P}}$

↳ Tells proportion of +ve prediction is actually correct.

(III) $\boxed{\text{Recall} = \frac{TP}{TP + FN}}$

↳ out of all the True Actual positive, How many have predicted correctly.

Predicted \ Actual	1	0
1	TP	FP
0	FN	TN

SPAM CLASSIFICATION ↳ PRECISION

CANCER → RECALL

STOCK MARKET PREDICTION → F-SCORE.

(IV) F-SCORE

$$F\text{-}Beta = (1+\beta^2)\frac{Precision \times Recall}{Precision + Recall}$$

For $\beta = 1$

$$F\text{-}1 \Rightarrow \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Decreasing β means you are provid-ing more importance to false +ve than false -ve.

$$\boxed{\begin{array}{l} \beta\downarrow \Rightarrow FP >> FN \\ \beta\uparrow \Rightarrow FN >> FP \end{array}}$$

NAIVE BAYES INTUITION

{ classification }

↳ BAYE'S THEOREM

BAYE'S THEOREM

$$\boxed{P(B/A) = \frac{P(B) * P(A/B)}{P(A)}}$$

FEATURES				RESULT
x_1	x_2	x_3	x_4	YES
-	-	-	-	NO
—	—			YES
—	—			YES.

classmate

Date

Page

$$P(y=Yes|x_i) = \frac{P(Yes) * P(x_1|Yes) * (P(x_2|Yes) * \_\_ \_\_ * P(x_n|Yes)}{P(x_1) * P(x_2) * P(x_3) * \_\_ P(x_n)}$$

Ignore This Term

constant in Both (Fixed)

$$P(y=No|x_i) = \frac{P(No) * P(x_1|No) * \_\_ * P(x_n|No)}{P(x_1) * P(x_2) * P(x_3) * \_\_ P(x_n)}$$

fixed

constant

We need to do Normalization, cuz value of Probability will be v.small

$P(Yes|x_i) = 0.13$ | $P(No|x_i) = 0.05$

$> 0.5 \Rightarrow 1$

$< 0.5 \Rightarrow 0$

AFTER NORMALIZATION

72%

AFTER NORMALIZATION

28%

Q Question of NAIVE BAYES

DAY	OUTLOOK	TEMPERATURE	HUMIDITY
D1	SUNNY	HOT	HIGH
D2	SUNNY	HOT	HIGH
D3	OVERCAST	HOT	HIGH
D4	RAIN	MILD	HIGH
D5	RAIN	COOL	NORMAL
D6	RAIN	COOL	NORMAL
D7	OVERCAST	COOL	NORMAL
D8	SUNNY	MILD	HIGH
D9	SUNNY	COOL	NORMAL
D10	RAIN	MILD	NORMAL
D11	SUNNY	MILD	NORMAL
D12	OVERCAST	MILD	HIGH
D13	OVERCAST	HOT	NORMAL
D14	RAIN	MILD	HIGH

OUTLOOK

	YES	NO	P(YES)	P(NO)
SUNNY	2	3	2/9	3/5
OVERCAST	4	0	4/9	0/5
RAIN	3	2	3/9	2/5
Total →	9	5		

WIND	PLAY TENNIS
WEAK	NO
STRONG	NO
WEAK	YES
WEAK	YES
WEAK	YES
STRONG	NO
STRONG	YES
WEAK	NO
"	YES
"	YES
STRONG	YES
STRONG	YES
WEAK	YES
STRONG	NO

TEMP

	YES	NO	P(YES)	P(NO)
HOT	2	2	2/5	2/5
MILD	4	2	4/5	2/5
COLD	3	1	3/5	1/5
ToL	9	5		

classmate

Date

Page

$P(YES) = \frac{9}{14}$ $P(NO) = \frac{5}{14}$

→ P(YES/ (Sunny & HOT)

$$\Rightarrow \frac{P(YES) * P(Sunny/YES) * P(HOT/YES)}{\cancel{P(Sunny) * P(Hot)}}$$

$$= \frac{\cancel{9}}{14} \times \frac{2}{\cancel{9}} \times \frac{2}{9} = \frac{2}{63} = 0.031$$

→ $P(No/(Sunny \& Hot)) = \frac{5}{14} \times \frac{3}{5} \times \frac{2}{5} = 0.085$

NORMALIZE

$$P(NO/Sunny, Hot) = \frac{0.085}{0.085 + 0.031} = 0.73$$

$$P(Yes/Sunny, Hot) = 0.27$$

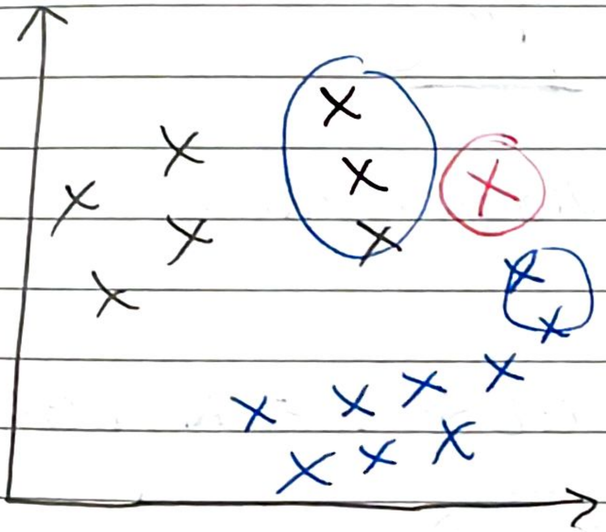
→ K-Nearest Neighbour

KNN ALGORITHM

(Classification & Regression)

↳ We use Eucledian / Manhattan Distan-ce.

(I)



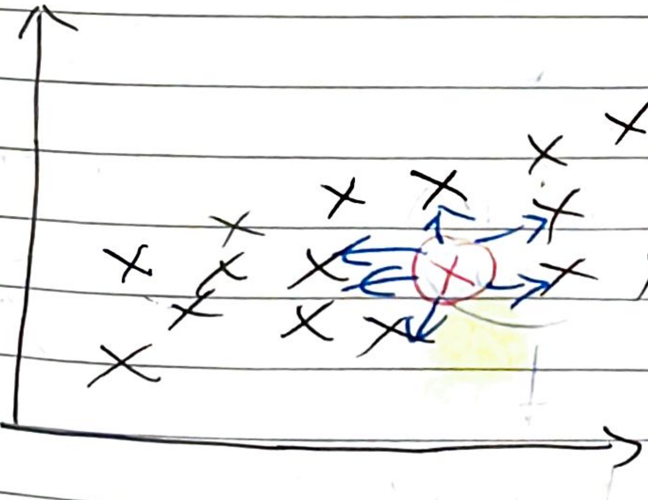
K=5

Max No.

Black => 3
Blue => 2

CLASSIFICATION

(II)



K=6

Regression

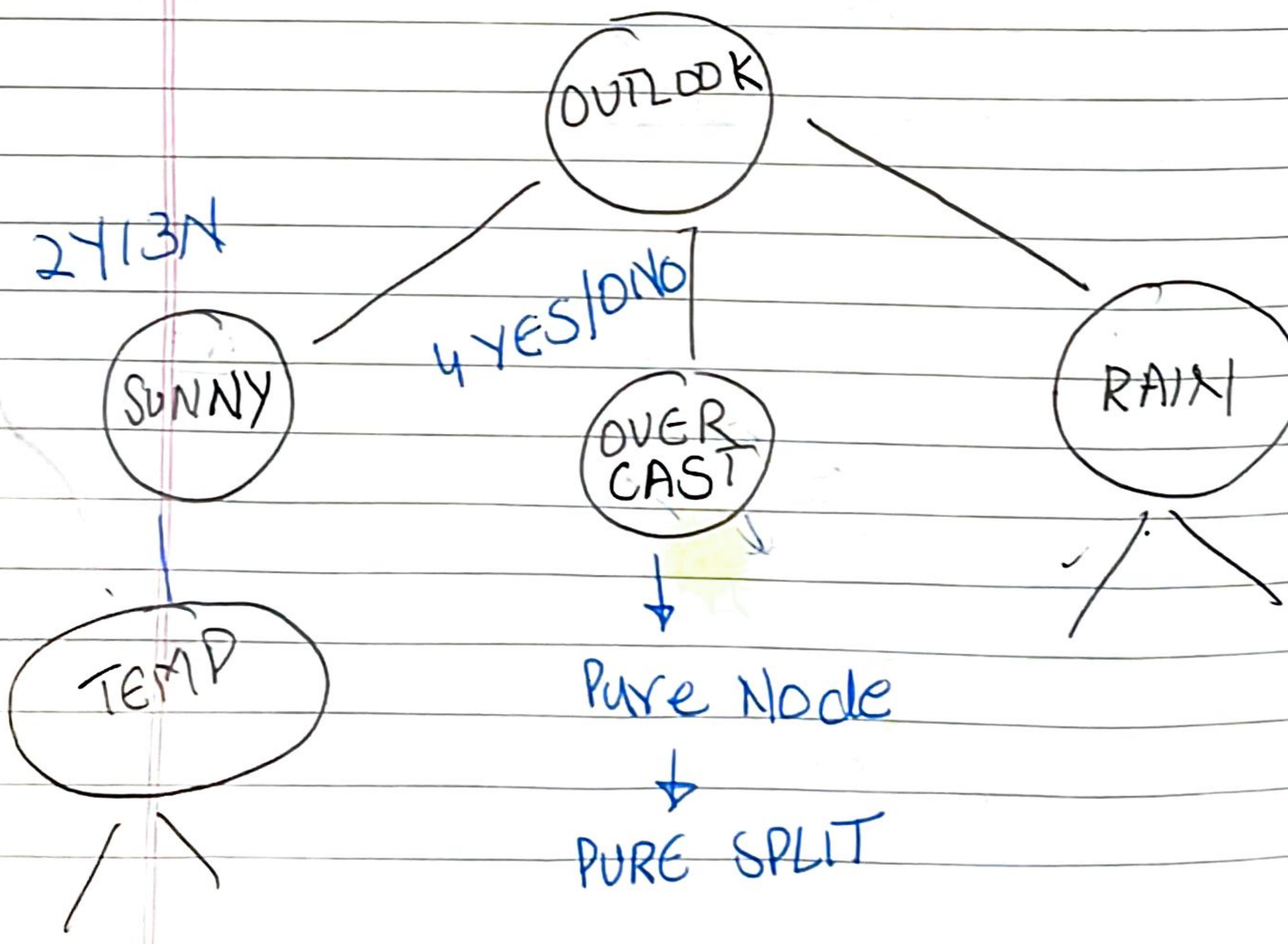
KNN WORKS V. BAD UNDER 2 CONDITIONS

(I) Outlier
(II) Imbalanced Dataset.

DECISION TREE

(CLASSIFICATION & REGRESSION)

↳ NESTED IF ELSE. ⇒ NODES.



→ PURE SPLIT
→ IMPURE SPLIT.

Q HOW TO KNOW THE SPLIT IS PURE SPLIT OR NOT W/O ACTUALLY LOOKING D.T?

Sol ⇒ Entropy
Gini Coefficient Impurity.

Q How the features are selected?

Sol ⇒ Information Gain

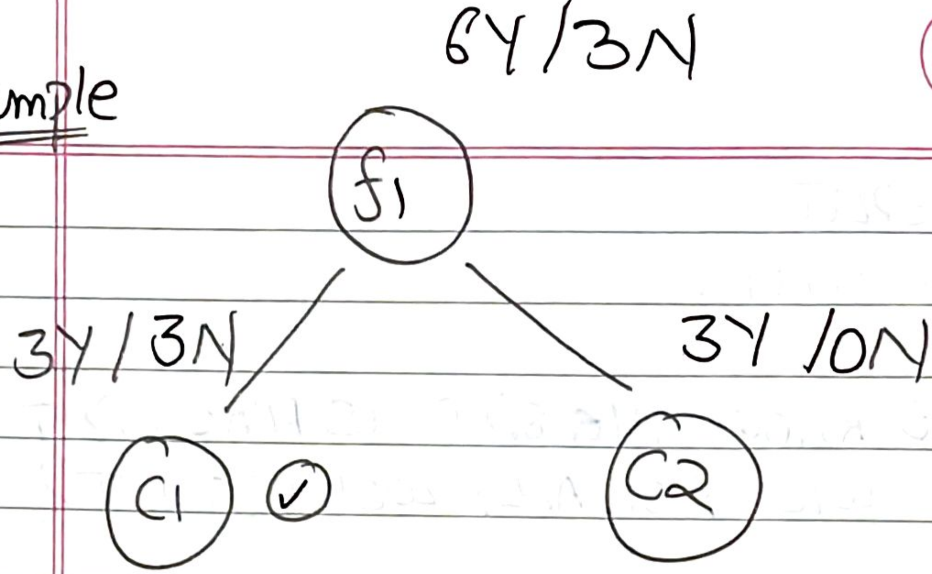
① ENTROPY

$\cancel{H(S) = -P_+ \log_2 P_+ - P_- \log}$

$$\boxed{H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-}$$

P_+ ⇒ Probability of Yes

Example



ENTROPY => $-\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}$

ENTROPY $= -1\log_2 1 = 0$

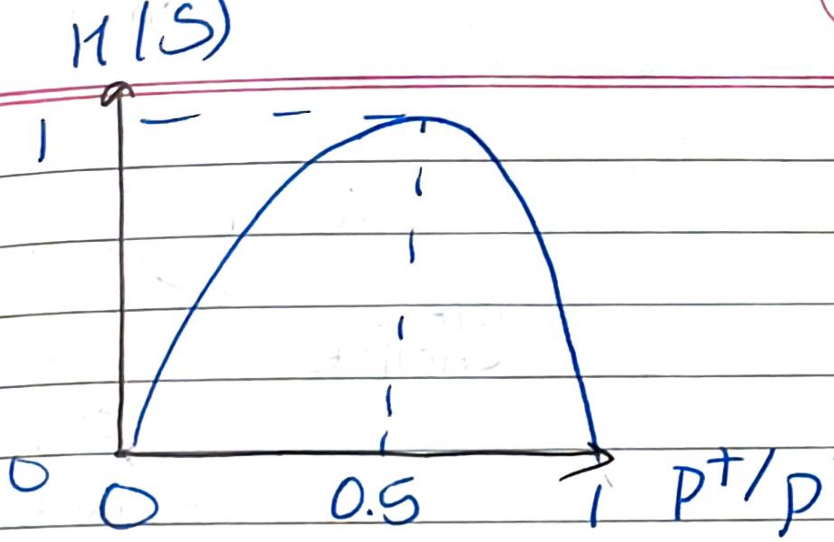
↳ of PURE SPLIT

☆ ENTROPY OF PURE SPLIT IS ALWAYS 0.

ENTROPY $= -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$.

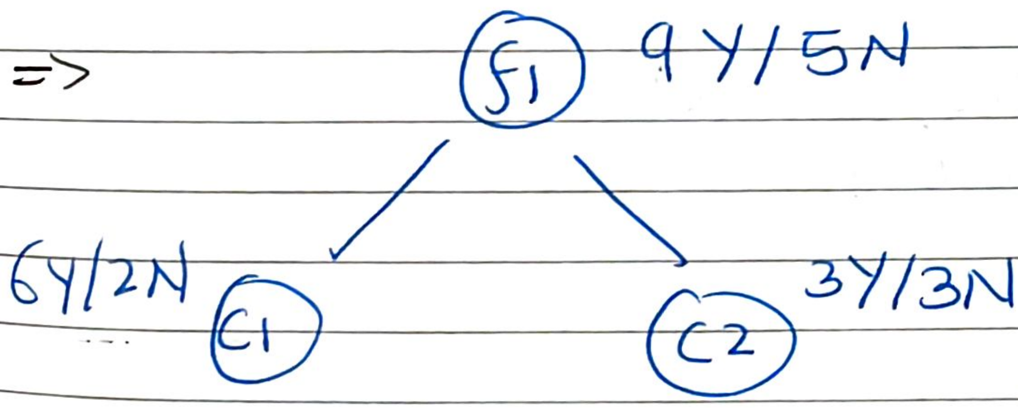
ENTROPY = 1 $(p^+ = 0.5,\ p^- = 0.5)$

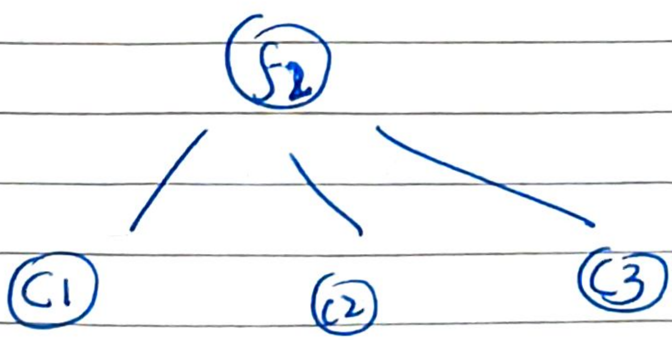
☆ ENTROPY is always $\in (0 \rightarrow 1)$



② Which Feature to take to split??

Q =>





INFORMATION GAIN

Entropy of Root Node

SAMPLE FOR CATEGORIES

$$I.G \Rightarrow Gain(S, f_1) = H(S) - \sum_{v \in val} \frac{|S_v|}{|S|} H(S_v')$$

TOTAL SAMPLE

ENTROPY OF CATAGERIOS C1, C2

So, We Need

$H(S)$ ✓
$H(C_1)$ ✓
$H(C_2)$ ✓
⋮

$$Gain(S, f_1) = 0.94 - \left[\frac{8}{14} \times 0.81 + \frac{6}{14} \times 1\right]$$

$$\boxed{Gain(S, f_1) = 0.049}$$

$$\boxed{Gain(S, f_2) = 0.051}$$

→ Select the feature with More Info. Gain

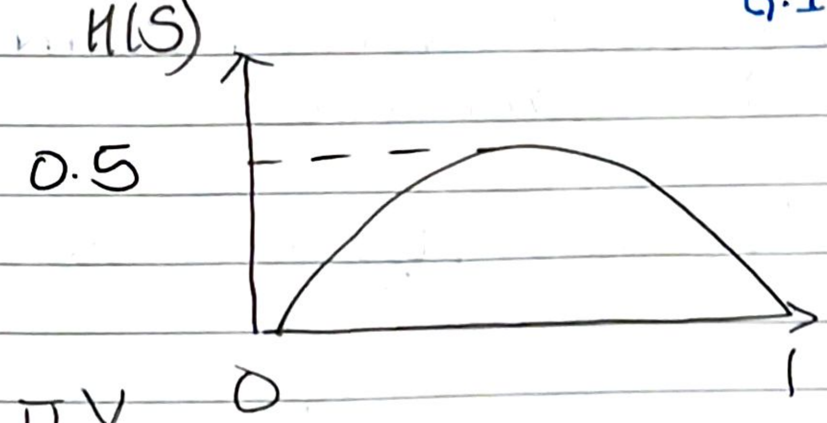
GINI IMPURITY

↳ $1 - \sum_{i=1}^{n} (p)^2$

For 2 o/p.

G. Impurity $= 1 - [(p_+)^2 + (p_-)^2]$.

G. Impurity $\in (0, 0.5)$



COMPLEXITY

ENTROPY ⟶ Log

GINI IMPURITY → Simple Maths.

FOR Decision Tree, Try to apply GINI IMPURITY

↳ AS IT WILL BE EASY.

FAST

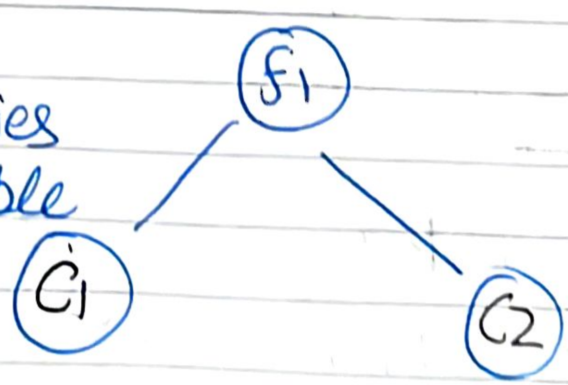
GINNI > ENTROPY

FEATURE SELECTION IN DECISION TREE REGRESSO

OP is an contineous feature

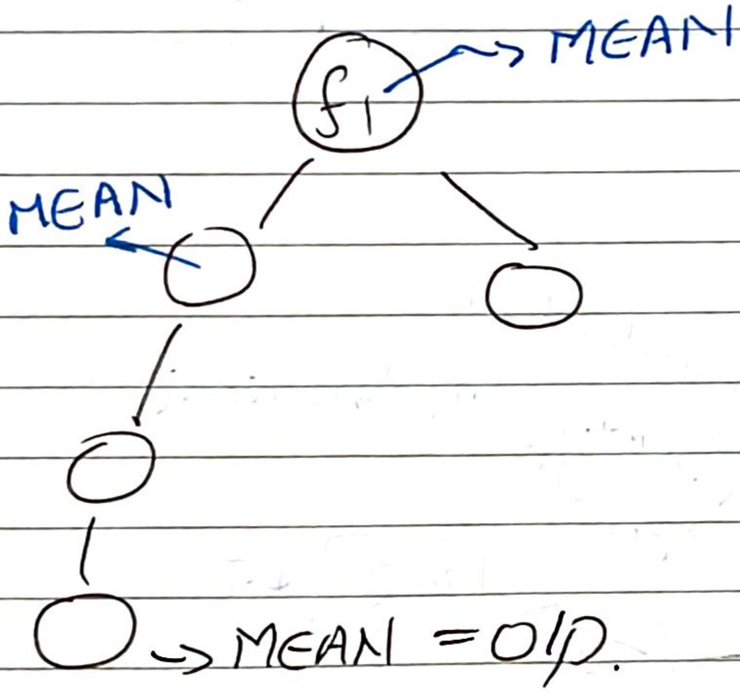
- MEAN of o/p will be calculated.
- Mean square error or MEAN absolute error is calculated.
- Splitting will happen based on that

Based on categories of contineous variable



- Mean of categories will be calculated then MSE " " "

- As MSE will get reduced, we are reaching towards leaf node
- Finally, when you follow this path what will be the mean value will be the output.



HYPERPARAMETER

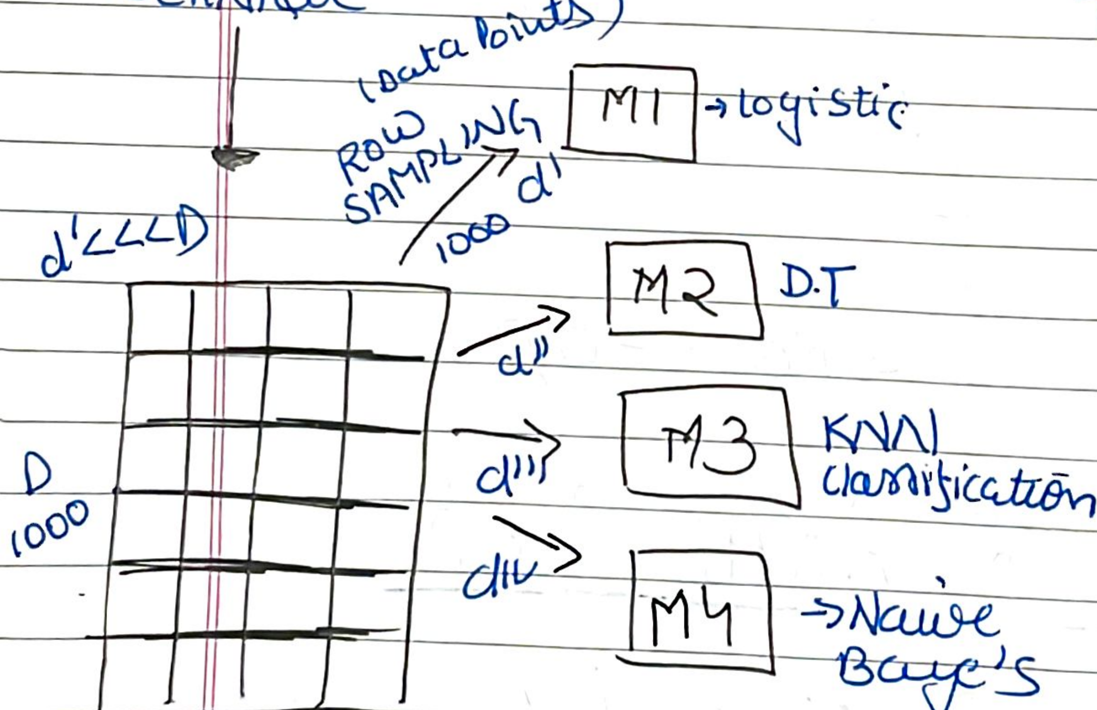
↳ Do Pre pruning
↳ max-depth, max_leaf.

→ Overfitting can happen if don't do hyperparameter tuning.

ENSEMBLE TECHNIQUES

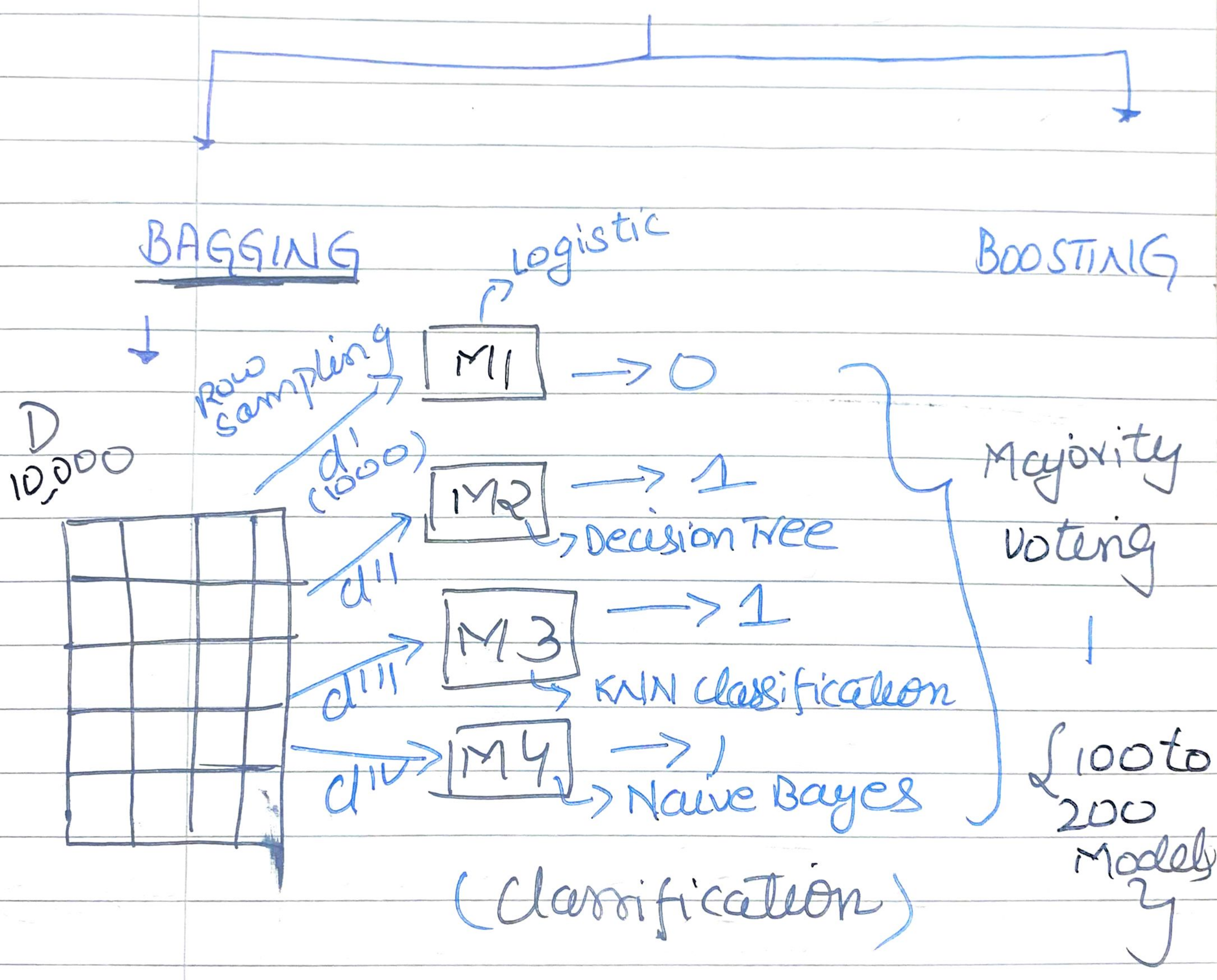
BAGGING TECHNIQUE

BOOSTING TECHNIQUE



→ Overfitting can happen if you don't do hyperparameter tuning.

ENSEMBLE TECHNIQUE



{Bootstrap Aggregating}

↓

Also known as Bagging.

__/__/____

ENSEMBLE TECHNIQUE → Ensemble technique in ML involve combin-ining predictions from multiple model to improve overall performa-nce.

→ In Bagging, in case of Baggin Regression entire mean will be taken

BOOSTING

↳ Sequential combination of model

whereas Boosti Bagging is parallel & independent multiple model

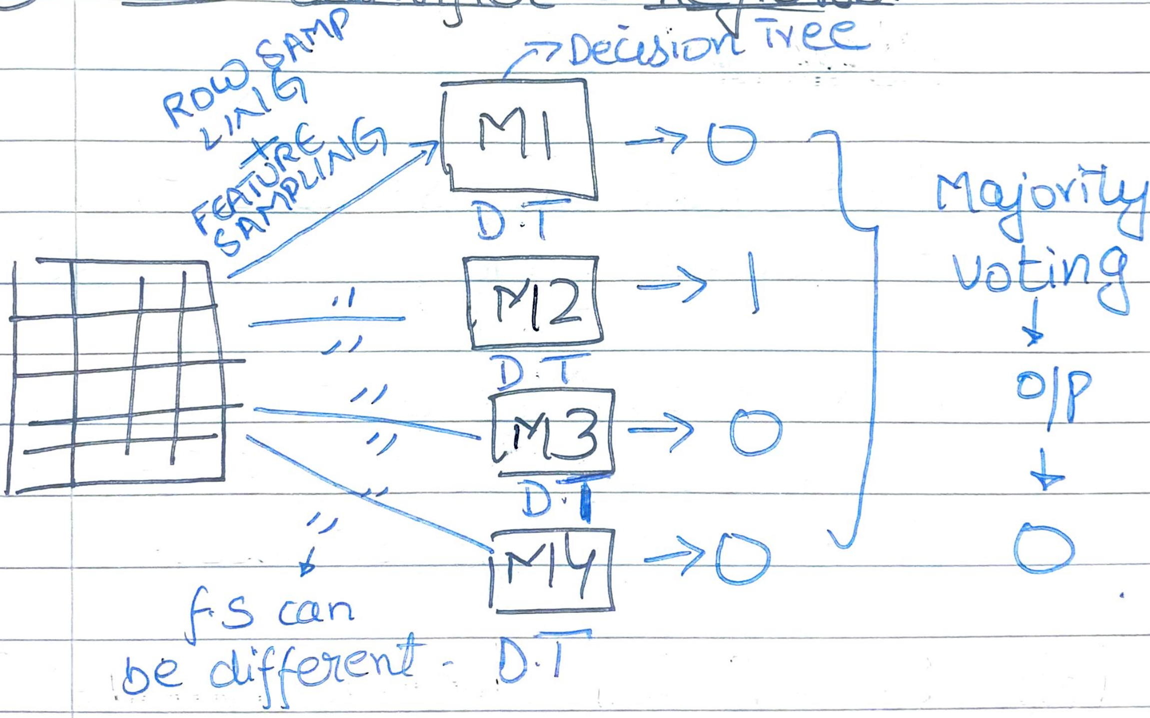
→ M1 → M2 → M3 → M4 → o/p

weak Learner, weak learner, ", "

STRONG LEARNER

widely used	BAGGING	BOOSTING
(I)	R.F (Classifier)	(I) AdaBoost
(II)	R.F (Regression)	(II) Gradient Boost.
		(III) XgBoost.

(I) <u>R.F Classifier & Regressor</u>



☆ Every D.T will create Low Bias & High Variance
↳ But if we combine using Bagging it will create Low Bias, low

variance which is required

Q1) Is Normalization required in R.F? / D.T

Q2) In KNN is Normalization or Standardization Required?

Ans 1) No, D.T basically do the splits, if you minimize the data that split will not be that beneficial.

Ans 2) Yes, Here we use two things Eucledian, Manhatten Distance.

KNN calculates distance which can be sensitive to differences in scale among features.

Q3) Is R.F impacted by outlier?

Ans 3) No, They have some resilence towards outliers.

Q Is KNN impacted by Outlier?

→ Big Yes.

★ You can also make custom Bagging put which ever model you want

BOOSTING

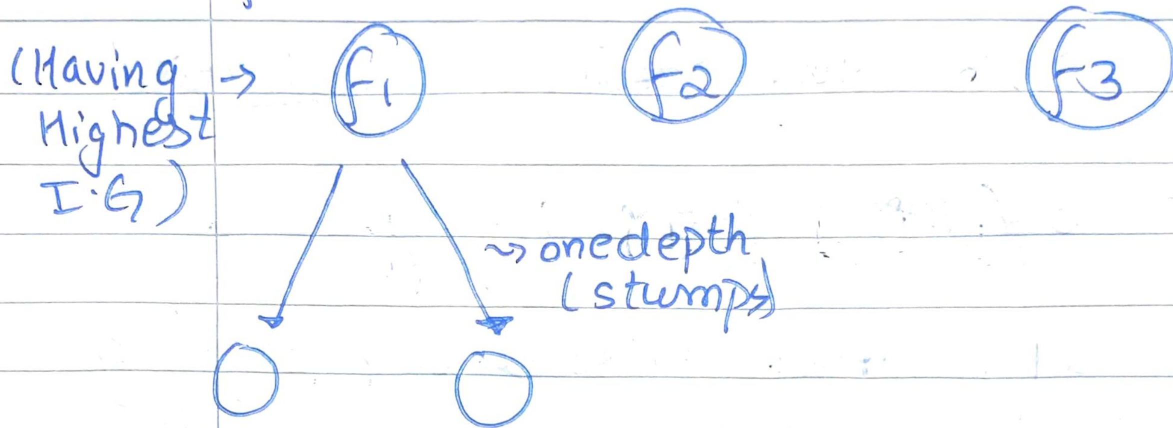
1) ADABOOST

over wt = 1

f1	f2	f3	f4	o/p	weight	New weight
–	–	–	–	Yes	1/7	0.05
–	–	–	–	No	1/7	0.05
–	–	–	–	Yes	1/7	0.05
–	–	–	–	"	1/7	0.349
–	–	–	–	"	1/7	0.05
–	–	–	–	No	1/7	0.05
–	–	–	–	Yes	1/7	0.05

Error

Feature selection on I. Gain



{weak learner}

→ Now, we will train the model & will calculate Total error.

- Let it gives only one error in o/p.

$$\text{Total error} = \frac{1}{7}$$

- Performance of Stumps

$$\text{Performance} \Rightarrow \frac{1}{2}\log_e\left(\frac{1-TE}{TE}\right)$$

$$\Rightarrow \frac{1}{2}\log_e\left(\frac{1-1/7}{1/7}\right)$$

$$\Rightarrow 0.895.$$

- Next Step, Update weight

<u>Correct Records</u>

$$\boxed{\text{New Sample weight} = \text{Weight} \times e^{-(\text{Performance of stumps})}}$$

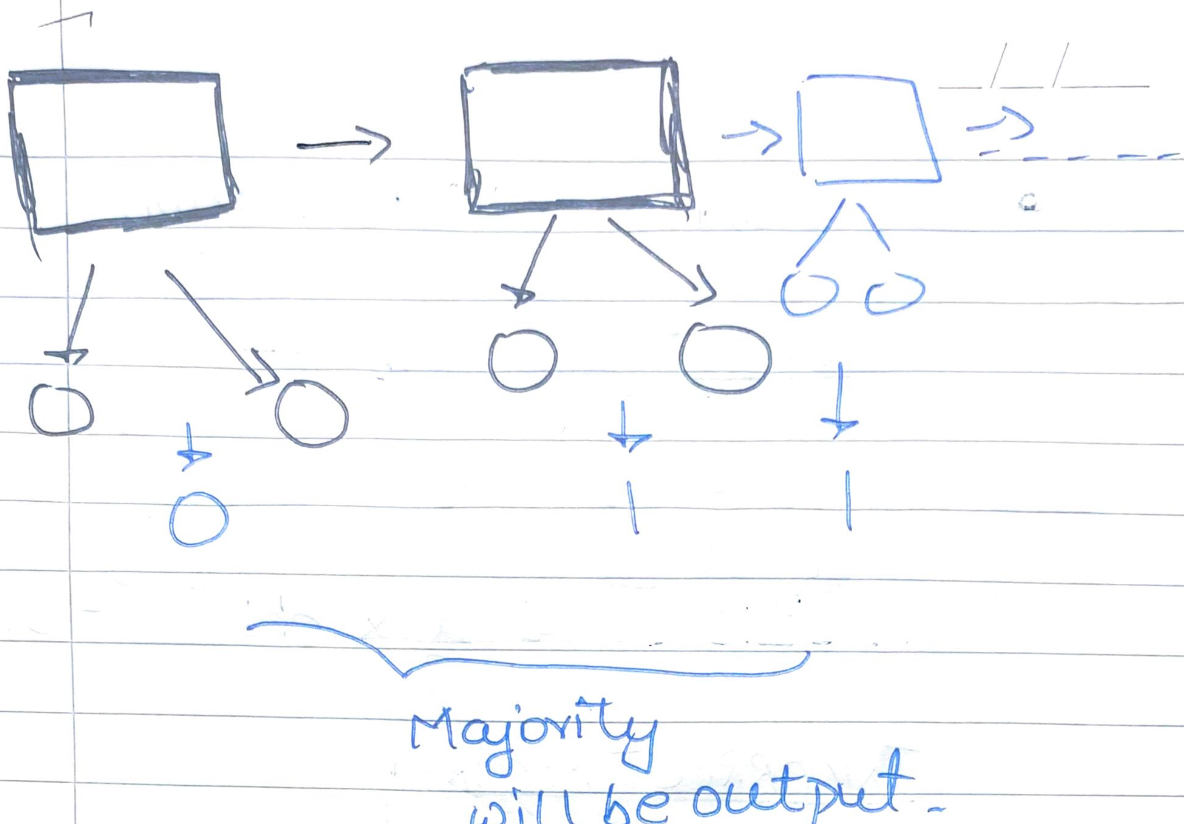
$$= \frac{1}{7} \times e^{-0.895} = 0.05.$$

<u>INCORRECT RECORD</u>

$$\boxed{\text{New Sample weight} = \text{Weight} \times e^{+P_S}}$$

$$\Rightarrow \frac{1}{7} \times e^{0.895} = 0.349.$$

New weight		Normalized weight	BUCKETS
0.05	÷ 0.649	0.07	[0-0.07)
0.05	''	0.07	[0.07-0.14]
0.05	''	0.07	[0.14-0.21]
0.349	''	0.537	[0.21 - ~~0.28~~] 0.747
0.05	''	0.07	[0.[illegible] 0.[illegible]] 0.751
0.05	''	0.07	[- -]
0.05	''	0.07	[- -].
0.649		⇒ 1	



→ Most of wrong records will be selected & given to other stumps

BLACK BOX MODEL

↳ Internal mechanism and decision making processes of black-box model are not easily discernible.

WHITE BOX MODEL

↳ Internal working of white box models are clear and easy to interpret.

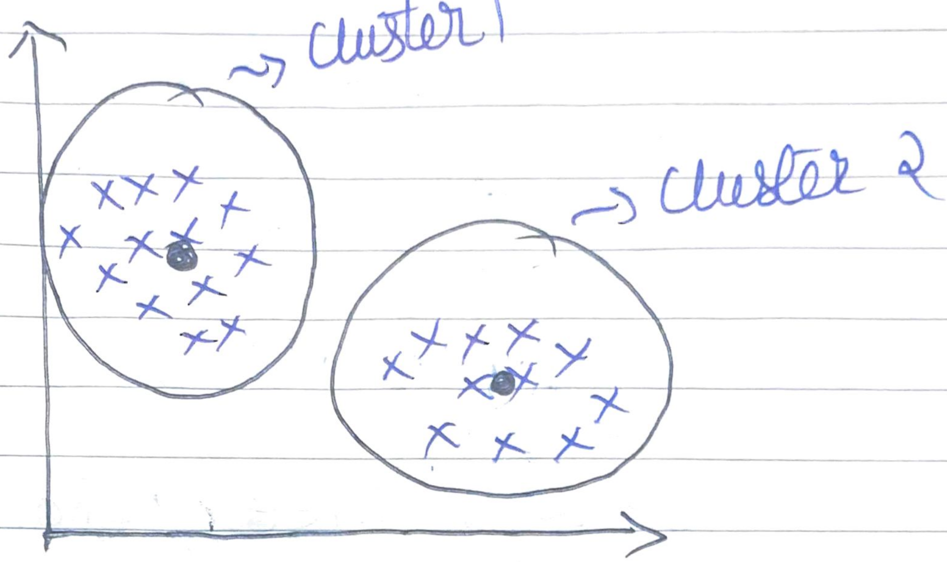
E.g → Blackbox → Random Forest,
ANN.

WhiteBox → Linear Regression,
Decision Tree.

Decision Tree	Random Forest

UNSUPERVISED M.L

K-MEAN CLUSTERING



K-Means → K ⇒ Centroids.

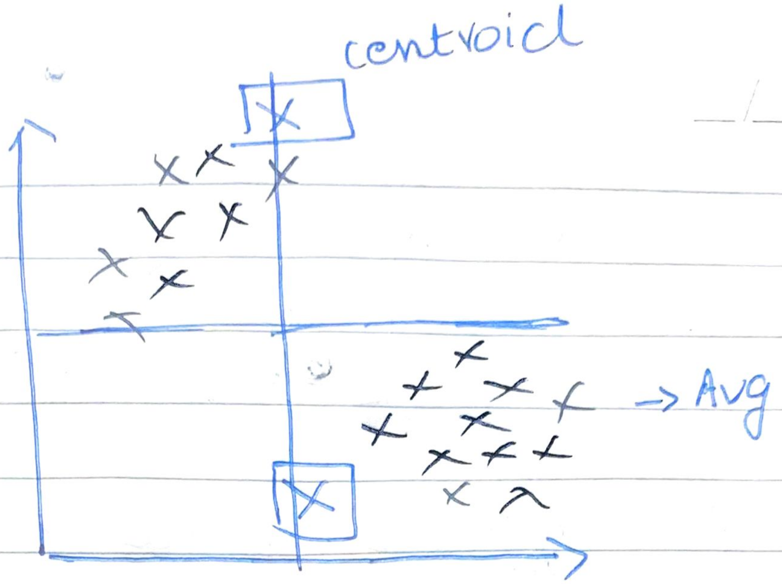
K = 2 → 2 centroids.
(●)

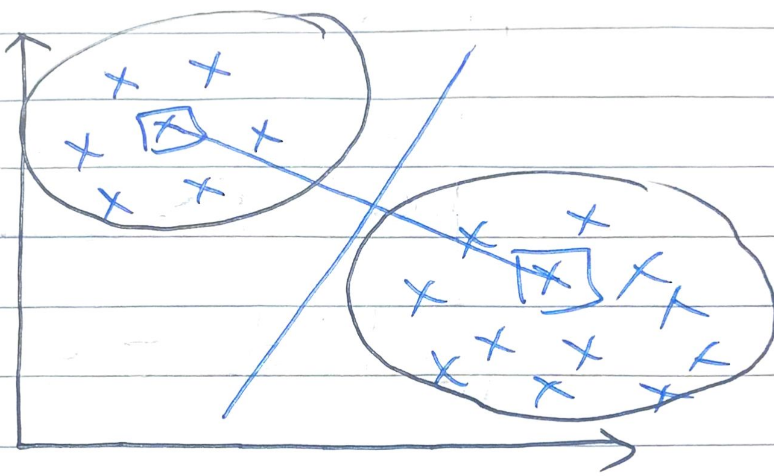
HOW TO DO?

(I) We try K values

(II) We intializes K no. of centroids.

(III) Compute the average to update centroid.



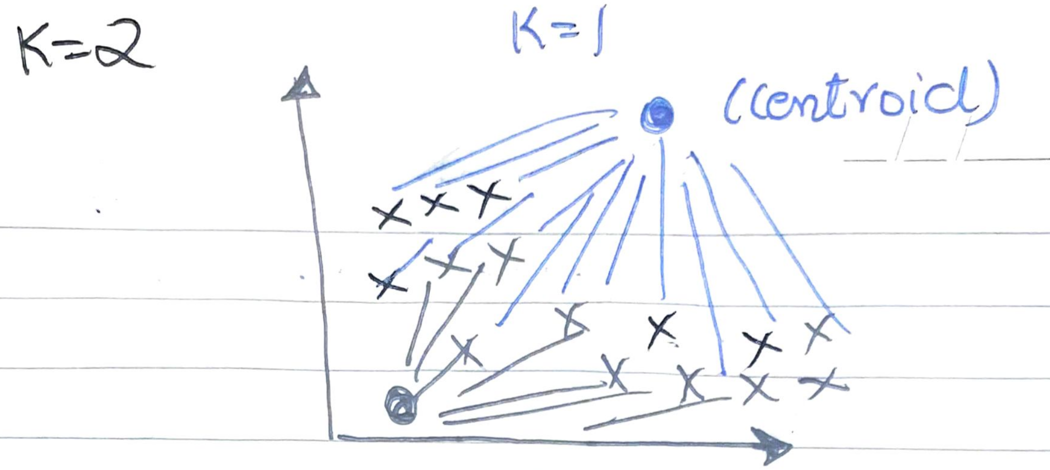


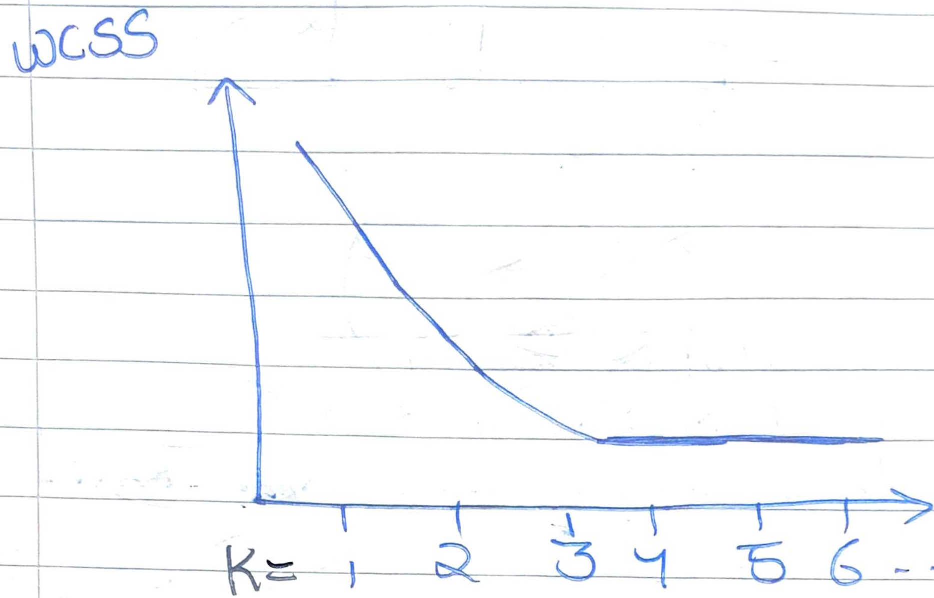
Q HOW TO DECIDE K-VALUE?

ELBOW METHOD (K-value)

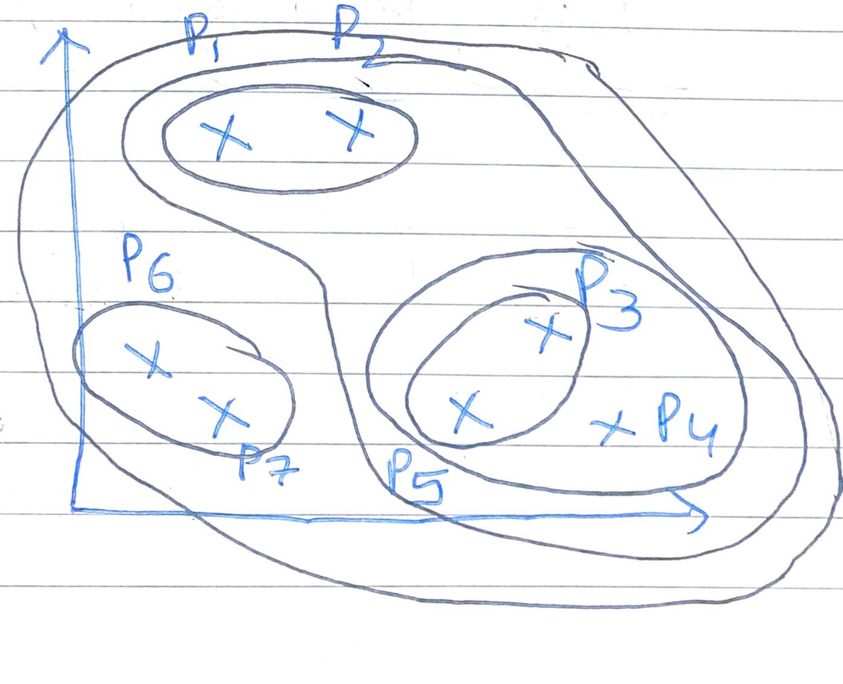
WCSS ⇒ within cluster sum of square

for i ⇒ 1, → 10

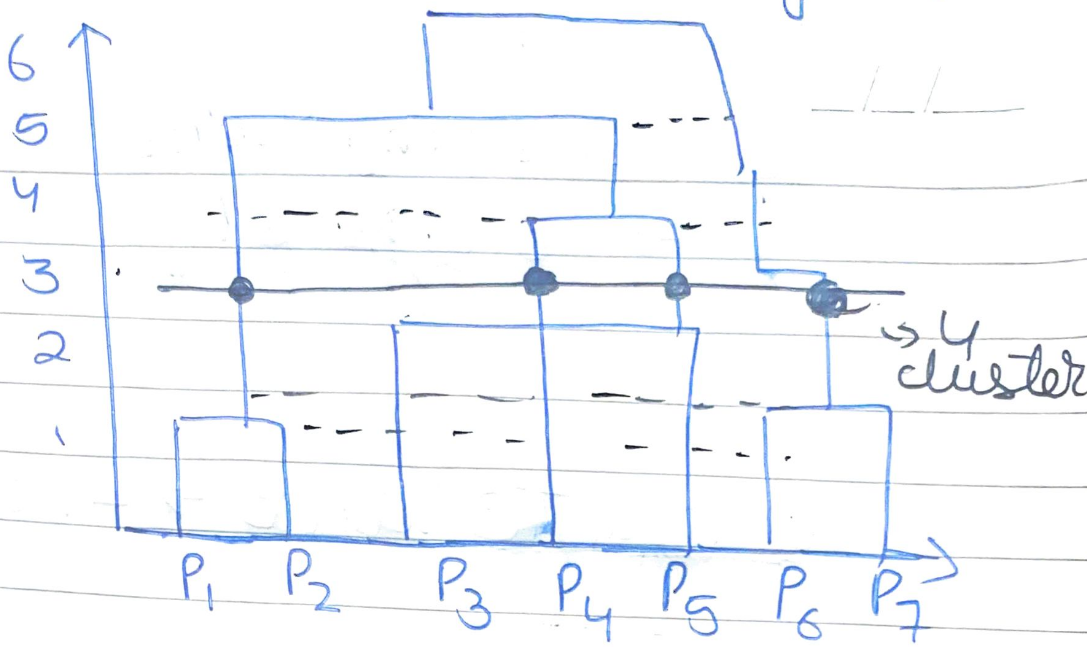




② HIERARICHAL CLUSTERING



→ Dendogram



→ The funda is that you need to find longest vertical line that has no vertical line passed through it.

Q Which take more time K-Means or Hierarichal clustering?

→ Max. time is taken by Hierarich-al clustering.

Q HOW DO WE VALIDATE CLUSTERING MODEL?

↳ Silhouette CLUSTERING

Good Clustering Model
↳ $b(i) >> a(i)$

Silhouette Value → -1 to $+1$

$$\boxed{s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_I| > 1}$$

—x —x —x —x —x —x —x—x

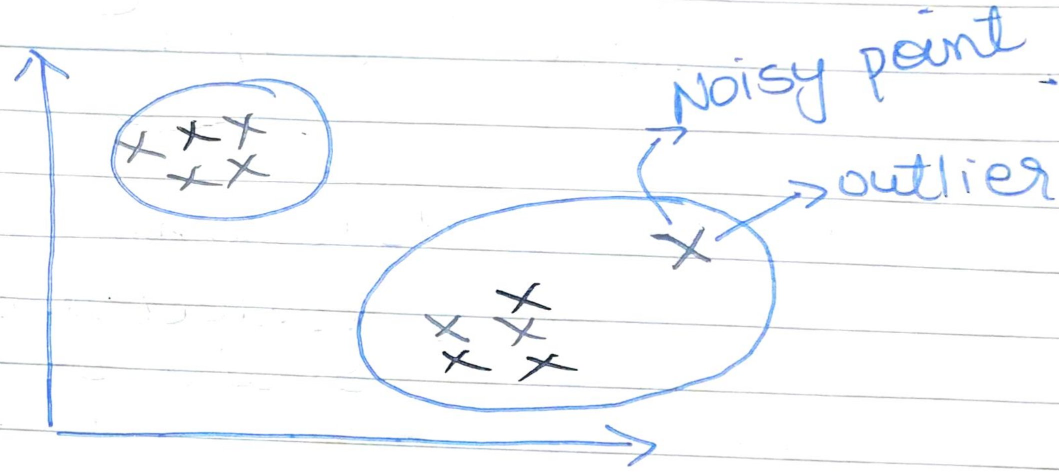
DBSCAN CLUSTERING

↳ Density-Based Spatial Clustering of Applications with Noise.

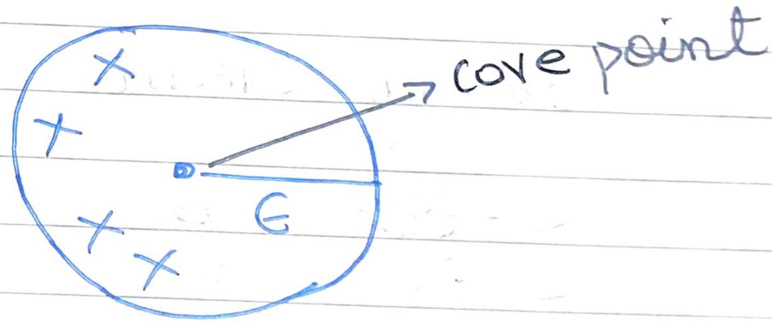
(I) Min points.
(II) Core points.
(III) Border Points.
(IV) Noise points.

→ where you want to skip the outlier, we can use DBScan

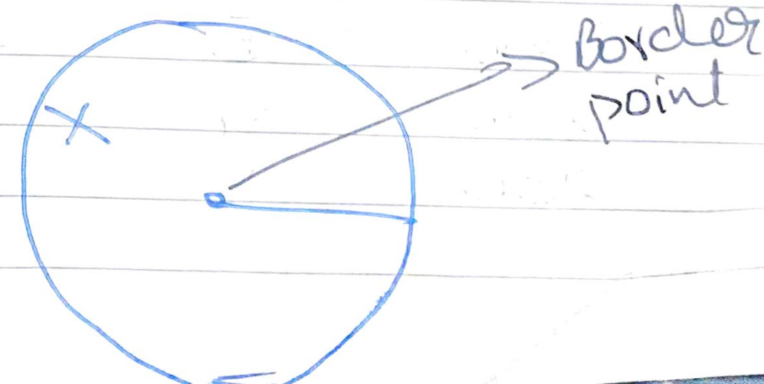
K-Means



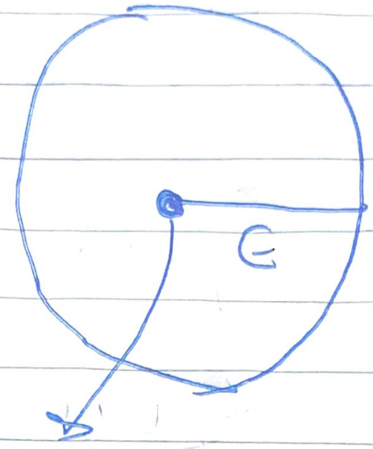
Core Point



BORDER POINT



Noisy Point



Noise point
(Treated as outlier
& will be neglected)

XgBoost CLASSIFIER

↳ Extreme Gradient Boosting

Salary	Credit	Approval	Residual (Appro. - pr.)
<=50	B	0	-0.5
<=50	G	1	0.5
<=50	G	1	0.5
>50	B	0	-0.5
>50	G	1	0.5
>50	N	1	0.5
<=50	N	0	-0.5

BASE MODEL $\rightarrow pr. = 0.5$

Steps

(I) Create a Binary Decision Tree using the feature.

(II) Calculate similarity weight.

$$= \frac{\Sigma (\text{Residual})^2}{\Sigma Pr(1-Pr) + \lambda}.$$

(III) Calculate Info. Gain

—x—x—x—x—x—x—x

(I) Salary $\Rightarrow 0.14$.

$<=50$: Similarity $\Rightarrow 0$

$[-0.5, 0.5, 0.5, -0.5]$

>50: Similarity $= 0.33$

$[-0.5, 0.5, 0.5]$

$\Downarrow$

(II)
$$\frac{(1-1)^2}{0.5(1-0.5)+\lambda + 0.5(1-0.5) + 0.5(1-0.5) + 0.5(1-0.5)} \Rightarrow 0.$$

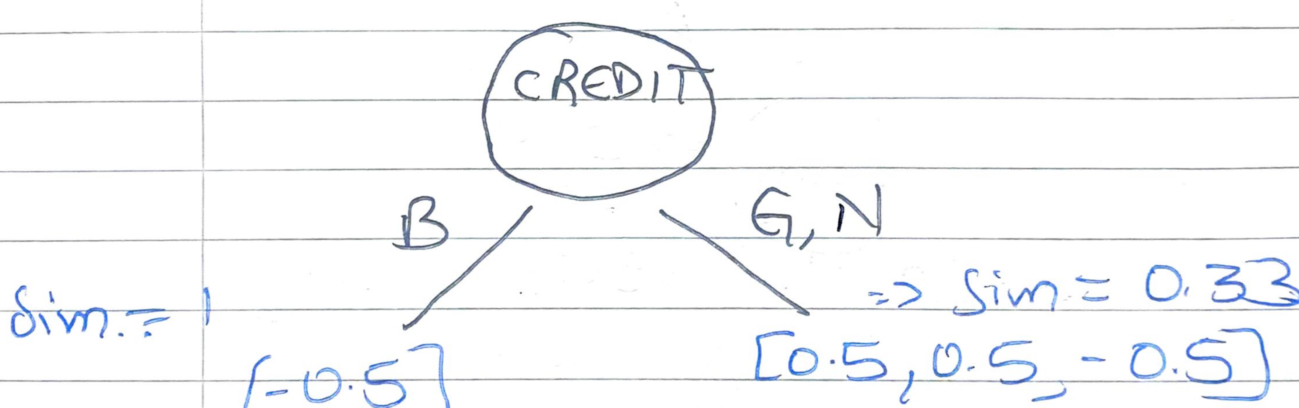
$\lambda = 0$ let us

11

$$\text{Similarity} = \frac{(1-0.5)^2}{0.5(1-0.5) + 0.5(1-0.5) + 0.5(1-0.5) + 0.5(1-0.5)} = 0.33$$

Similarity for Root node $\Rightarrow 0.14$

(III) $I.G = 0.33 - 0.14 = 0.19.$

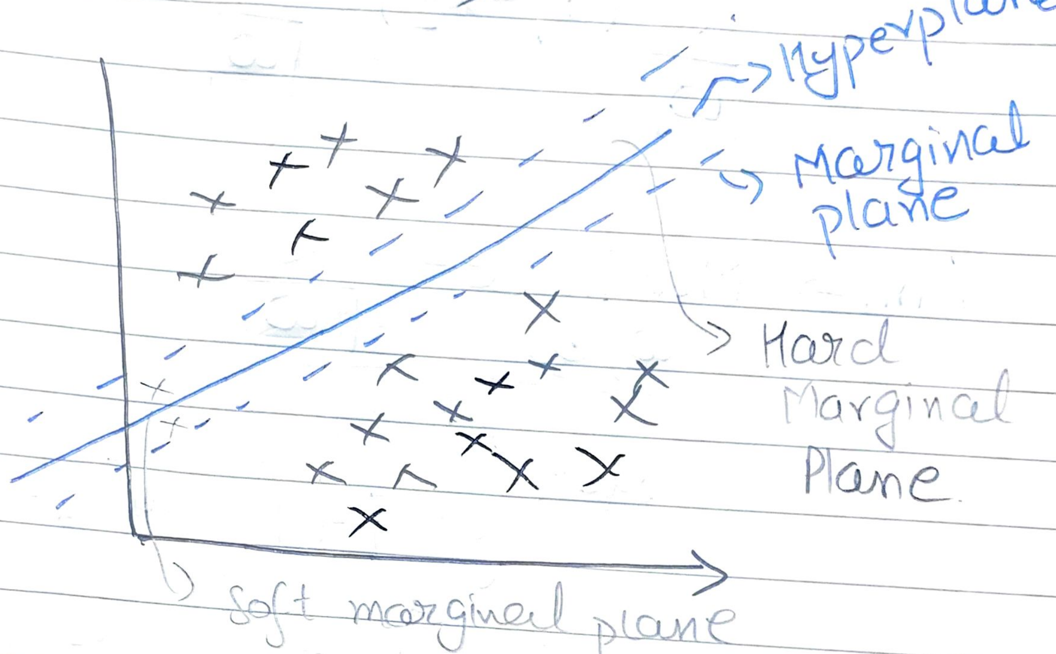


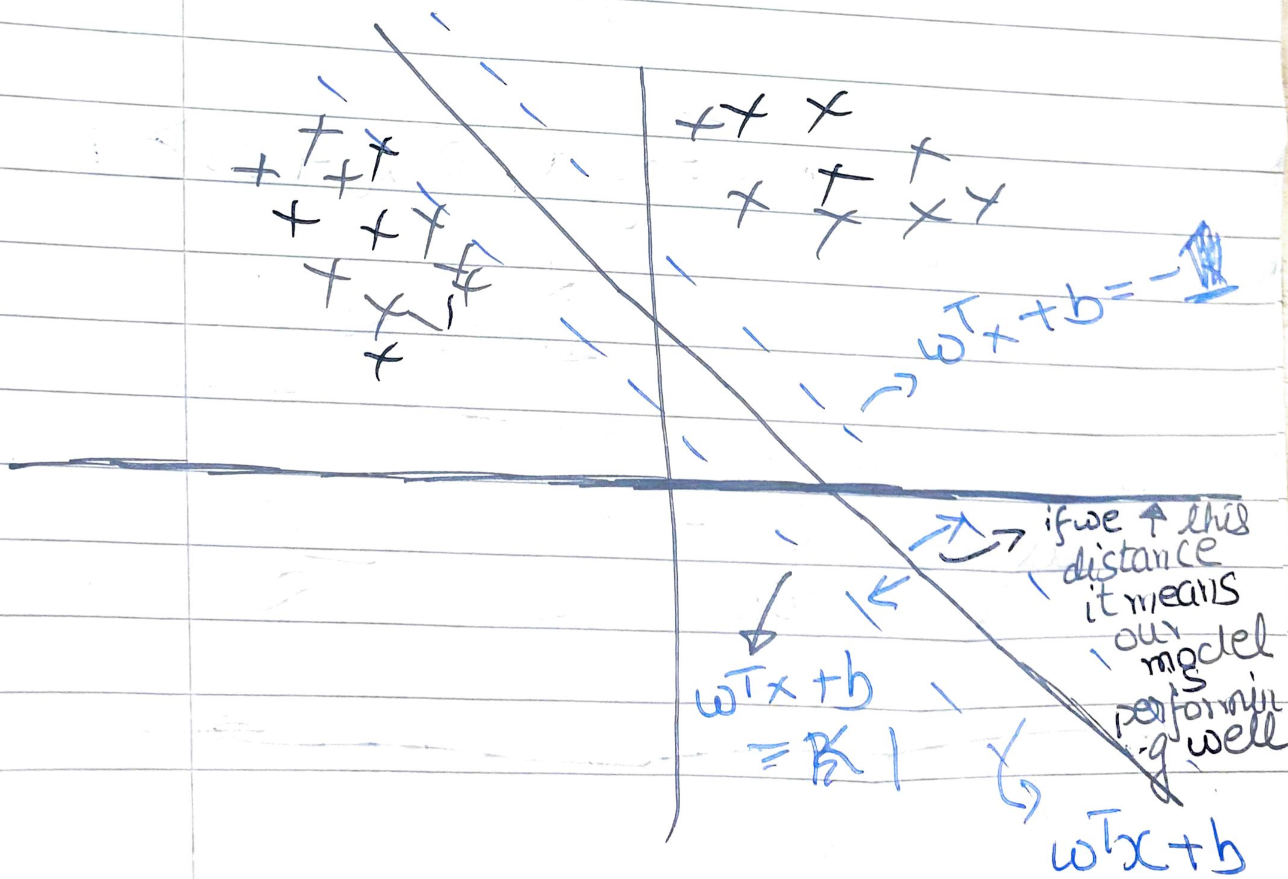
Sim = 1

Sim... = 0.33

$I.G = 1 + 0.33 - 0 = 1.33$

SUPPORT VECTOR MACHINE (SVM)





$$w^T x_1 + b = 1$$

$$w^T x_2 + b = -1$$

$$\frac{w^T}{|w|}(x_1 - x_2) = \frac{2}{|w|}$$

Aim → $\underset{(w,b)}{\text{Maximize}} \Rightarrow \frac{2}{\|w\|}$

such

$$y_i \begin{cases} +1 & w^T x + b \geq 1 \\ -1 & w^T x + b \leq -1 \end{cases}$$

Major Aim $\Rightarrow y_i * (w^T x_i + b) \geq 1$

$\forall$ correct point

$$\boxed{\text{Maximize } \frac{2}{\|w\|} \Rightarrow \underset{(w,b)}{\text{Min}} \ \frac{\|w\|}{2}}$$

Summation of the distance of wrong pts.

$$\boxed{\min_{(\omega, b)} \frac{\|\omega\|}{2} + C_i \sum_{i=1}^{n} \xi_i}$$

How many errors we can have