	x_0	x_1	x_2	y	
	1	1	2	3	→ sample 1
i ↓	1	2	1	4	2
	1	3	3	5	3

j →

(1) compute the predictions :-

for each sample: $h_\theta(x^{(i)}) = \theta_j x_j^{(i)} + \theta_j x_j^{(i)} + \theta_j x_j^{(i)}$

- $h(x^{(1)}) = \theta_0 x_0^{(1)} + \theta_1 x_1^{(1)} + \theta_2 x_2^{(1)}$
  $= 0*1 + 0*1 + 0*2$
  $= 0$
- $h(x^{(2)}) = 0*1 + 0*2 + 0*3$
  $= 0$
- $h(x^{(3)}) = 0*2 + 0*3 + 0*3$
  $= 0$

(2) compute gradient.

(i) compute errors or residuals.

$e^{(i)} = (h_\theta(x^{(i)}) - y^{(i)}) x_0^{(i)}$

for each sample:

$e^{(1)} = 0 - 3$
$= -3$

$e^{(2)} = 0 - 4$
$= -4$

$e^{(3)} = 0 - 5$
$= -5$

(ii) Calculate the gradients:-

$$\frac{\partial J}{\partial \theta_0} = \sum_{i=1}^{n} (h_\theta(x^{(i)}) - y^{(i)}) x_0^{(i)}$$

$$= -3*1 + (-4)*1 + (-5)*1$$

for each parameter

$$= \underline{-12}$$

$$\frac{\partial J}{\partial \theta_1} = -3*1 + (-4)*2 + (-5)*3$$

$$= \underline{-26}$$

$$\frac{\partial J}{\partial \theta_2} = -3*2 + (-4)*1 + (-5)*3$$

$$= -6 - 4 - 15$$

$$= \underline{-25}$$

(3) update parameters :- ($\alpha \to 0.1$)

$$\theta_0 = \theta_0 - \alpha \frac{\partial J}{\partial \theta_0}$$

$$\theta_0 = 0 - (0.1)(-12)$$

$$= \underline{1.2}$$

$$\theta_1 = 0 - (0.1)(-26)$$

$$= \underline{2.6}$$

$$\theta_2 = 0 - (0.1)(-25)$$

$$= \underline{2.5}$$

$$\begin{bmatrix} 1.2 \\ 2.6 \\ 2.5 \end{bmatrix}$$

$$J = \frac{1}{2} \sum_{i=1}^{n} (h_\theta(x) - y)^2 \quad \text{(error)}$$

$$= \frac{1}{2}((-3)^2 + (-4)^2 + (-5)^2)$$

$$= \frac{1}{2}(9 + 16 + 25)$$

M T W T F S S
Page No.:
Date:
YOUVA

$\theta_0 = 25$

⇒ 2nd Iteration :-

→ calculating the prediction

$h(x^1) = 1.2 * 1 + 2.6 * 1 + 2.5 * 2$

$= 1.2 + 2.6 + 5$

$= 8.8$

$h(x^2) = 1.2 * 1 + 2.6 * 2 + 2.5 * 1$

$= 1.2 + 5.2 + 2.5$

$= 8.9$

$h(x^3) = 1.2 * 1 + 2.6 * 3 + 2.5 * 3$

$= 1.2 + 7.8 + 7.5$

$= 17.5$

→ compute gradient.

i) error :-

$e^{(1)} = 8.8 - 3$

$= 5.8$

$e^{(2)} = 8.9 - 4$

$= 4.9$

$e^{(3)} = 17.5 - 5$

$= 12.5$

→ Calculate the gradient :-

$$\frac{\partial J}{\partial \theta_0} = \sum_{i=1}^{n} (h_\theta(x^{(i)}) - y^{(i)}) x_0^{(i)}$$

$= 5.8 * 1 + 4.9 * 1 + 12.5 * 1$

$= 5.8 + 4.9 + 12.5$

$= 23.2$

In scikit learn:-

- One hot encoder
- Ordinal encoder } for X
- Label encoder — for y

x_1	x_2	y
red	35.8	Y
blue	24.4	N

$$\hat{y} = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2$$

$x_{1,0}$	$x_{1,1}$	x_2	y
0	1	35.8	Y
1	0	24.4	N

$$\hat{y} = \theta_0 x_0 + \theta_1 (x_{1,0}^*) + \theta_2 (x_{1,1}^*) + \theta_3$$

(2) Data Standardization & Normalization

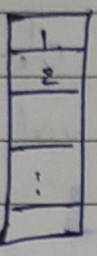
x_1 feet, x_2 kms, x_3 categorical variable

→ Normalized

$-\infty \leq x \leq +\infty$

24/1

→ K-fold cross validation:



Eg: Sample	x_1	x_2	y
1	2	-1	1
2	5	1.2	0
3	1	2	1
4	-3	-2	1
5	4	0.1	0

3-fold CV:- Acc = ? Std = ?

1 fold.

1 $[\theta_1, \theta_2] = [1.8, 2.8]$ — ①

2 $[\theta_1, \theta_2] = [2.1, 3.1]$

3 $[\theta_1, \theta_2] = [1.9, 4]$

⇒ 1st fold:

Train :- I II → {1,2} {3,4}

Test :- III → {5}

Acc 1) ~~-1.8×5+2.8~~ Std 1

2) 2

3) 3

2nd

$\begin{bmatrix} -1.8 \\ 2.8 \end{bmatrix}$ $\begin{bmatrix} 5 \end{bmatrix}$

Train :- I III → {1,2} {5}

test :- II → {3,4}

3rd

Train :- II III → {3,4} {5}

Test :- I → {1,2}

Acc 1)

2)

3)

Std 1)

2)

3)

fold 3 $[2 \quad -1]$ $[5 \quad 1.2]$

$[1.8 \quad 2.8] \begin{bmatrix} 2 \\ -1 \end{bmatrix}$

$[1.9 \quad 4] \begin{bmatrix} 2 \\ -1 \end{bmatrix}$

$\Rightarrow 0.45 \longrightarrow 0$

5th $\Rightarrow [1.9 \quad 4] \begin{bmatrix} 5 \\ 1.2 \end{bmatrix}$

$\Rightarrow 0.99 \longrightarrow 1$

$\Rightarrow$ fold 2:-

$[1 \quad -3]$

5th $[2.1 \quad 3.1] \begin{bmatrix} 2 \\ -1 \end{bmatrix}$

$\Rightarrow 4.2 + (-3.1)$

$= 1.1$

$k = n \longrightarrow$ LOOCV → done when dataset < 100

Leave One Out Cross Validation

10 → sample

10 → fold.

at each point 1 sample → test

9 → train

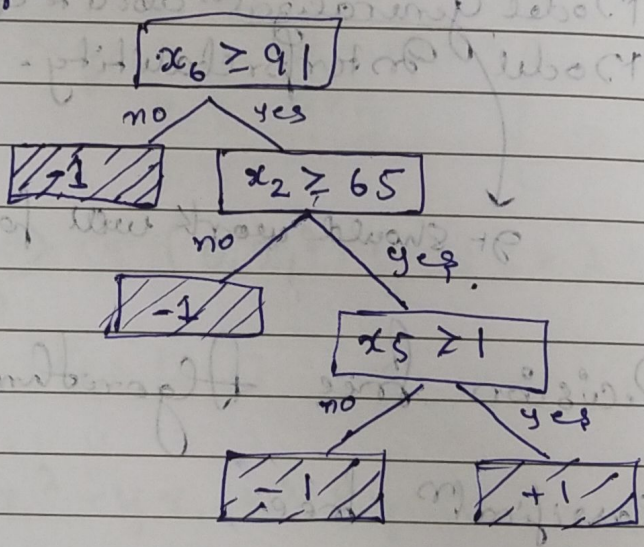
features :

x_1 : date
x_2 : age
x_3 : height
x_4 : weight
x_5 : Sinus tachycadia
x_6 : min systolic bP 24(hr)
x_7 : latest diastolic bp.

labels :

- high risk : +1
- low risk : -1

→ Graph:



→ testing / Inference / Predict phase

$x^{(i)}$ = 2020/11/17, 49, 172cm, 70.5 kg, 0, 115, 79.

low risk

IQ.	Social	verbal	Risk for Diabetes
100	79	86	Low
80	40	70	High
80	85	65	Low
90	82	45	High.

features:

x_1 : IQ

x_2 : Social

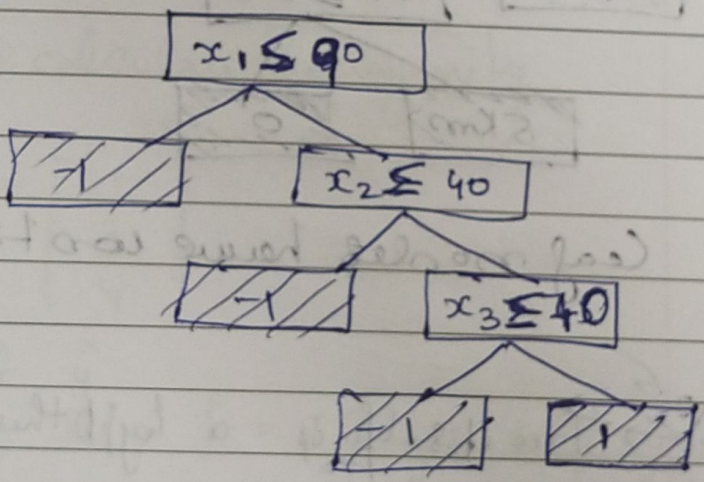
x_3 : verbal.

labels:

- high risk → +1

low risk → -1

→ Graph.



→ 1) feature
2) threshold.

Parameters - for every internal node
feature
threshold }

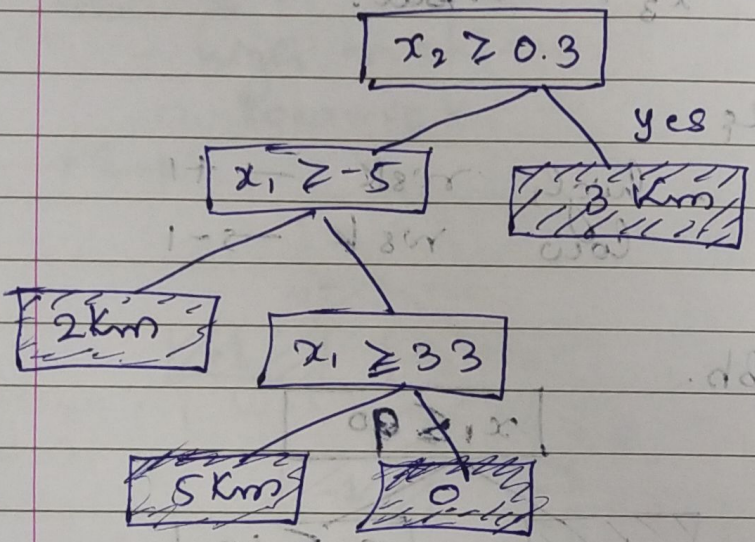
⇒ Regression setting :-

x_1 temperature - °C.

x_2 precipitation (cm/hr).

y : Walking - 1 Km
2 Km
3 Km

Regression tree :-



Leaf nodes have continuous values

Tree itself is a hypothesis function.

→

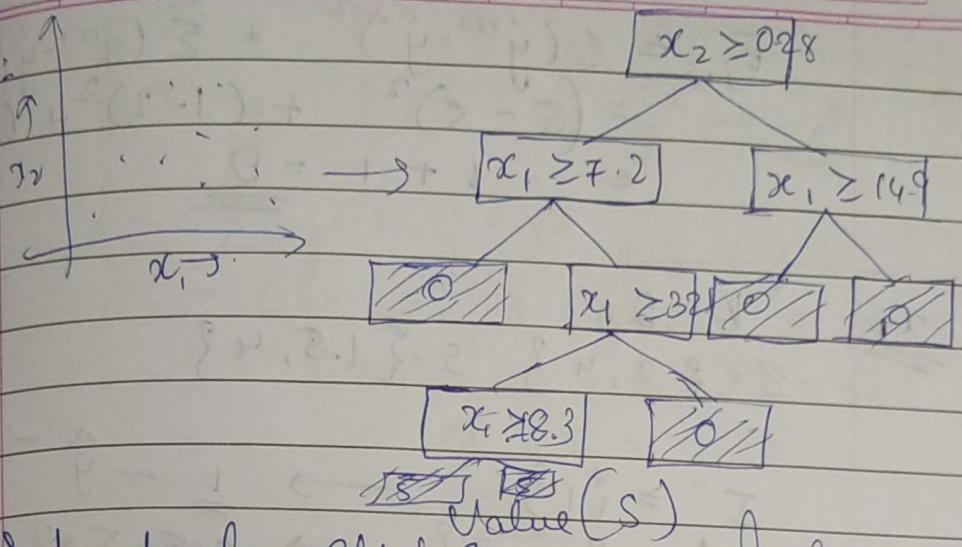
1) Split dim (feature)
2) Split value

Partition Space

feature

$[j^*, s^*]$ value

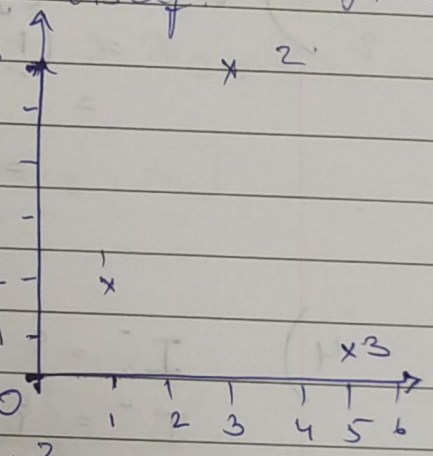
Eg:



Value (s)

⇒ to decide split ~~size~~ → people will usually take the middle of each feature or the data point itself. Point

#	x_1	x_2	y
1	1	2	1
2	3	6	5
3	5	1	2



$S = \{2, 4\}$ $S = \{1.5, 4\}$

x_1 $K = 2$.

2) $2 \rightarrow I^- = \{1\}$ $\leftrightarrow \hat{y}^-$

4× $I^+ = \{3, 5\}$ $\frac{8}{2} = 4$ $\rightarrow \hat{y}^+$

$$E_{j,s} = \Sigma (y^{(i)} - \hat{y}^+)^2 + \Sigma (y^{(i)} - \hat{y}^-)^2$$

$$= (1-1)^2 + (5-4)^2 + (2-4)^2$$

$$= 1 + 4 = 5$$

4) $I^- = \{1, 3\}$ $I^+ = \{5\}$

$\hat{y}^- = 2$ $\hat{y}^+ = 5$

$$E_{j,s} = \sum (y^{(i)} - \hat{y}^{+})^2 + \sum (y^{(i)} - \hat{y}^{-})^2$$
$$= (5-5)^2 + (1-2)^2 + (3-2)^2$$
$$= 1 + 1 = 2$$

---

x_1 $\quad k = 2$.

$S = \{2, 4\}$ $\quad s = \{1.5, 4\}$.

$I^- = \{1\} \longrightarrow 1 \Rightarrow \hat{y}^-$

$I^+ = \{2, 3\} \longrightarrow \frac{5+2}{2} = 3.5 \Rightarrow \hat{y}^+$

2) $E = \sum (y^{(i)} - \hat{y}^{+})^2 + \sum (y^{(i)} - \hat{y}^{+})^2$
$= (1-1)^2 + (5-3.5)^2 + (2-3.5)^2$
$= (1.5)^2 + (1.5)^2$
$= 2.25 + 2.25 = 4.5$

4) $I^- = \{1, 2\} \longrightarrow \frac{1+5}{2} = 3 \Rightarrow \hat{y}^-$

$I^- = \{3\} \longrightarrow 2 \Rightarrow \hat{y}^+$

$E = \sum (y^{(i)} - \hat{y}^{+})^2 + \sum (y^{(i)} - \hat{y}^{-})^2$
$= (5-3)^2 + (5-3)^2 + (2-2)^2$
$= 4 + 4$
$= 8$

---

x_2 $\quad S = \{1.5, 4\}$.

$I^- = \{3\} \longrightarrow \hat{y}^- \Rightarrow 2$

$I^+ = \{1, 2\} \longrightarrow \hat{y}^+ \Rightarrow \frac{1+5}{2} = 3$

1.5) $E = \sum (y^{(i)} - \hat{y}^{+})^2 + \sum (y^{(i)} - \hat{y}^{-})^2$
$= (2-2)^2 + (1-3)^2 + (5-3)^2$
$= 4 + 4$
$= 8$

M T W T F S S
Page No.:
Date:
YOUVA

4) $I^- = \{1,3\} \Rightarrow \hat{y}^- \rightarrow \frac{1+2}{2} = 1.5$

$I^+ = \{2\} \Rightarrow \hat{y}^+ \rightarrow \frac{5}{1}$

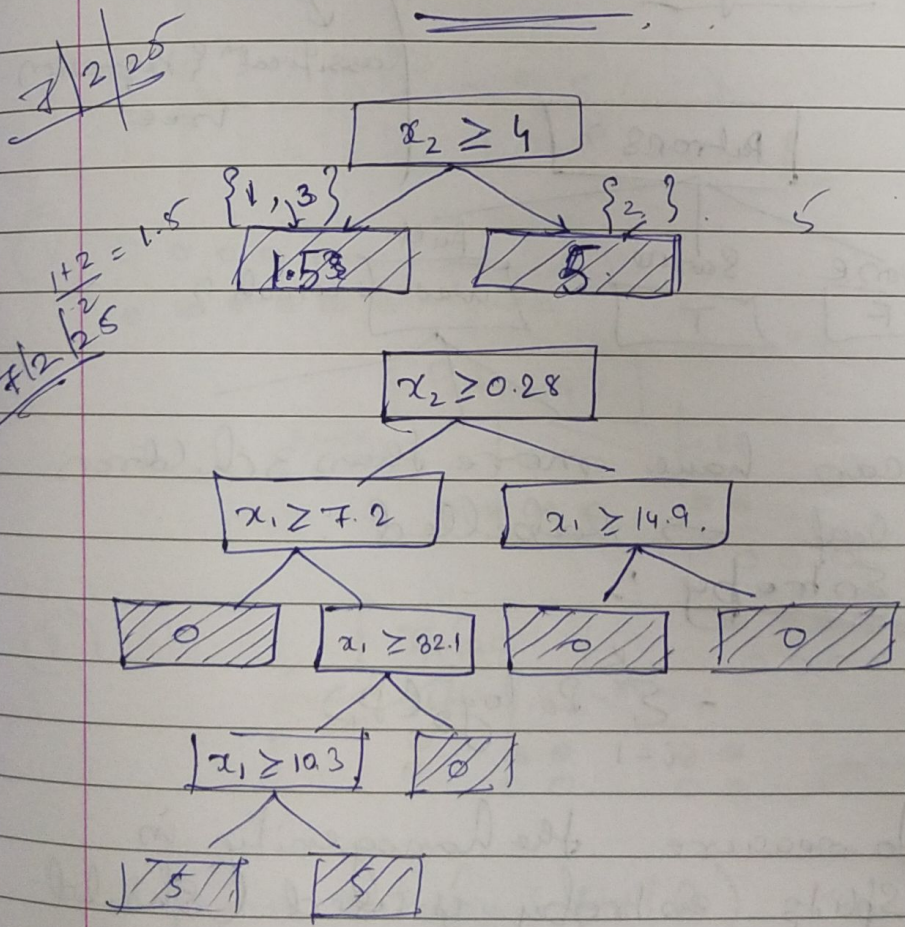
$$E = \Sigma(y^{(i)} - \hat{y}^+)^2 + \Sigma(y^{(i)} - \hat{y}^-)^2$$
$$= (1-1.5)^2 + (2-1.5)^2 + (5-5)^2$$
$$= (0.5)^2 + (0.5)^2 + 0$$
$$= 0.25 + 0.25$$
$$= 0.5$$

j^*, s^*

$\min = \{4.5, 8, 8, 0.5\}$

$= 0.5 \rightarrow x_2, 4.$

$j^*, s^* = [x_2, 4]$



IG = Information Gain.

no of +ve samples → negative samples.

$p = 6, \quad n = 6.$

base 2 algo

$$H\left(\frac{p}{p+n}, \frac{n}{n+p}\right) = -\frac{p}{p+n}\log$$

$$\Rightarrow -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{p}{p+n}$$

$$= -\sum_{i=1}^{k} -P_i \log_2 P_i$$

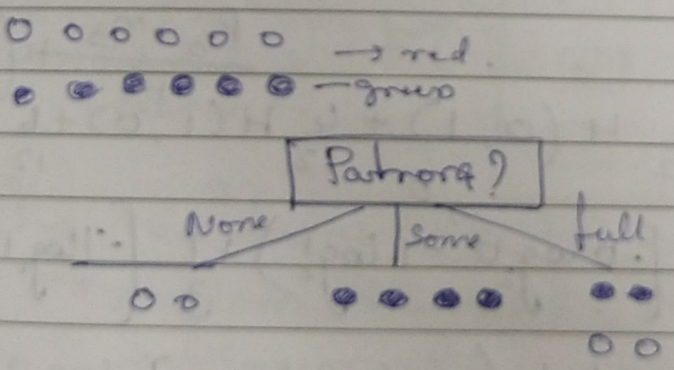
$$H(6,6) = -\frac{6}{12}\log_2\frac{6}{12} - \frac{6}{12}\log_2\frac{6}{12}$$

$$= -1/2 \log_2 0.5 - 0.5 \log_2 0.5$$

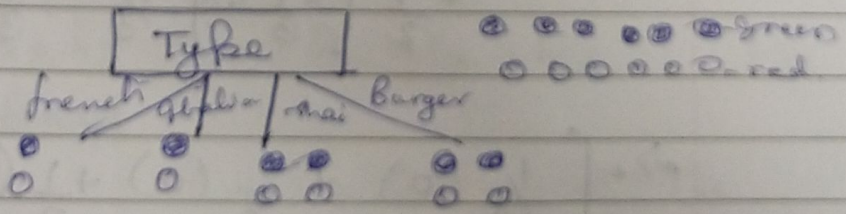
$$= -1/2(-1) - 0.5(-1)$$

$$= 1/2 + 0.5 = 1$$

(1)



(2)



Expected entropy : EH(A)

→

$$EH(A) = \sum_{i=1}^{k} -\left(\frac{p_i + n_i}{p+n}\right) H\left(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}\right)$$

maximize.

IG :

$$I(A) = H\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - EH(A)$$

1/2

$$H(\text{Patrons}) = H\left(\frac{p}{p+n}, \frac{n}{p+n}\right)$$

$$H\left(\frac{6}{12}, \frac{6}{12}\right) = \sum_{i=1}^{n} -p_i \log p_i$$

$$-\frac{p}{p+n} \log p_{frac}$$

$$= 1$$

$$EH(\text{Patrons}) = \sum_{i=1}^{3} -\frac{(p_i + n_i)}{p+n} H\left(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}\right)$$

$$\Rightarrow + \left[\frac{P_1 + n_1}{12} H\left(\frac{0}{2}, \frac{2}{2}\right) + \frac{P_2 + n_2}{12} H\left(\frac{4}{4}, \frac{0}{4}\right) + \right.$$

$$\left. \frac{P_3 + n_3}{12} H\left(\frac{2}{6}, \frac{4}{6}\right)\right]$$

$$\Rightarrow + \left[\frac{2}{12} H(0,1) + \frac{4}{12} H(1,0) + \frac{6}{12} H\left(\frac{1}{3}, \frac{2}{3}\right)\right.$$

$$\Rightarrow + \left[\frac{1}{6}\left[-0 \log_2 0 - 1 \log_2 1\right) + \frac{1}{3}\left[-1 \log_2 1 - 0 \log_2 0\right.\right.$$

$$\left. + \frac{1}{2}\left(-\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log 2/3\right)\right]$$

$$\Rightarrow + \left[\frac{1}{6}(0) + \frac{1}{3}(0) + \frac{1}{2}\left(+\frac{1.59}{3} + \frac{0.594}{3}\right)\right.$$

$$= + \left(\frac{1}{2}\right)(0.53 + 0.399)$$

$$= 0.464$$

$$I.G(\text{Patrons}) = H(\text{Patrons}) - EH(\text{Patrons})$$

$$= 1 - 0.46$$

$$= 0.54$$

⇒ Same calculation for $IG(\text{type}) = 0$

so patron split is more homogeneous.

Example:

x_1	y	r
1	3	3
2	6	6
3	8	8
4	11	11
5	15	15

a) $x_1 \geq 3$.

fit (x, r) $\boxed{x_1 \geq 3}$

$\boxed{4.5}$ $\boxed{4.3}$

$\frac{3+6+8}{3}$ $\frac{11+15+8}{2}$

x_1	r	$\hat{y}$	r_{new}
1	3	3.45	-0.45
2	6	6.45	-0.45
3	8	9.13	-1.13
4	11	12.13	-1.13
5	15	16.13	-1.13

$\hat{y} = r + \lambda$ (value)

$3 + 0.1 * 4.5 = 3.45$

$6 + 0.1 * 4.5 = 6.45$

$8 + 0.1 * 11.3 = 9.13$

x_1, r_{new}

⇒ Inference.

$x_1 = 5$.

$\lambda f_1(5) + \lambda f_2(5) + \lambda f_3(5) \ldots$

$\therefore \lambda f_b(5)$

$= 0.8$.

Kamperh. corrnidafatutti

Boosted Regression Tree:-
(Ada Boost Regression)

Boosted Classification Tree
(Ada Boost classification)

1. Initialize $f(x) = 0$
$r = y$

2) for iteration 1 to B
{
a) fit a model, $f_b(x, \theta) \to r$
b) update the final model by adding shrunken version
$F(x, \theta) = f(x, \theta) + \lambda$
c) update the residuals.
$r^{(i)} = r^{(i)} - \lambda f_b(x; \theta)$
}
3 final model
$F(x, \theta) = \sum_{i=1}^{B} \lambda f_b(x, \theta)$

→ Eg:-

x_1	x_2	y
1	2	4
2	3	5
3	4	6
4	5	7.

$F(X) = 0$
$f_0(X) = 0$

Iteration 1 :-

x_1	x_2	y	$r = y$	r
1	2	4	4	3.55
2	3	5	5	4.55
3	4	6	6	5.35
4	5	7	7	6.35

a) fit model, $f_1(x) \rightarrow r$.

$x_1 \leq 2.5$

$\frac{4+5}{2}$ → 4.5

$\frac{6+7}{2}$ → 6.5

b) update the model $\lambda = 0.1$

$$F(x) = f_0(x) + \lambda f_1(x)$$

c) update the residuals.

$r_1 = r_1 - \lambda f_1(x) = 4 - 0.45 = 3.55$

$r_2 = r_2 - \lambda f_2(x) = 5 - 0.45 = 4.55$

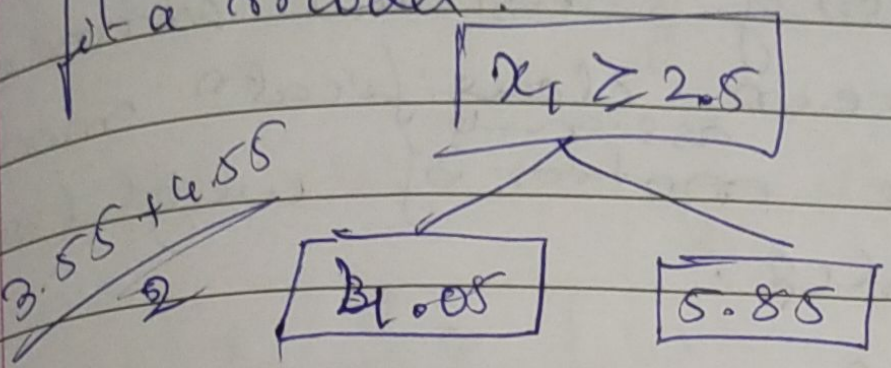
$r_3 = r_3 - \lambda f_3(x) = 6 - 0.65 = 5.35$

$r_4 = r_4 - \lambda f_4(x) = 7 - 0.65 = 6.35$

Iteration 2 :-

x_1	x_2	r
1	2	3.55
2	3	4.55
3	4	5.35
4	5	6.35

a) fit a model.



b) update the model.

$$f(x) = f_0(x) + \lambda f_1(x) + \lambda f_2(x)$$

c) update the residuals

$$r_1^{(1)} = r^{(1)} - \lambda f^{(1)}(x) = 3.55 - 0.405 = 3.145$$

$$r_2 = r^{(2)} - \lambda f^{(2)}(x) = 4.145$$

$$r_3 = r^{(3)} - \lambda f^{(3)}(x) = 4.765$$

$$r_4 = r^{(4)} - \lambda f^{(4)}(x) = 5.765$$

⋮

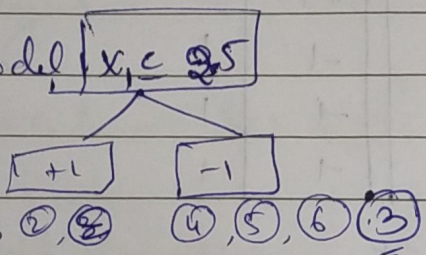
$\lambda \propto \frac{1}{\omega}$

Page No.: Date: YOUVA

Eg:-

x_1	x_2	y	ω
1	2	+1	1
2	3	+1	1
3	3	+1	1
4	5	-1	1
5	5	-1	1
6	6	-1	1

1) fit a model $x_1 \leq 2.5$



$\Rightarrow E = \frac{1}{6} = 0.1667.$

$\Rightarrow \lambda = \frac{1}{2} \log_e \left(\frac{1 - 0.1667}{0.1667} \right)$

$= \frac{1}{2} \log_e (4.998).$

$= \frac{1}{2} (1.6092)$

$= 0.805$

$\omega_i^n = \omega^n e^{-\lambda b}$

$\omega^1 = 1 \cdot e^{-0.805}$

$= 0.447 = \omega^2 = \omega^4 = \omega^5 = \omega^6$

$\omega^n = \omega^n e^{\lambda b}$

$\omega^3 = 1 \cdot e^{\lambda b} = e^{0.805} = 2.236$

x_1	x_2	y	w
1	2	+1	0.447
2	3	+1	0.447
3	3	+1	2.236
4	5	-1	0.447
5	5	-1	0.447
6	6	-1	0.447

M T W T F S S

Page No.: YOUVA

Date:

→ Gradient Boosting for classification.

Sample	x_1	x_2	y
1	2	3	1
2	1	2	0
3	3	4	1
4	4	5	1
5	1	1	0
6	2	2	0

Step 1 :-

Initialization :-

$F_0(x)$ = log. odds of positive class probability.

$$p = \frac{3}{6} = 0.5$$

$$F_0 = \log\left(\frac{p}{1-p}\right) = \log\left(\frac{0.5}{0.5}\right)$$

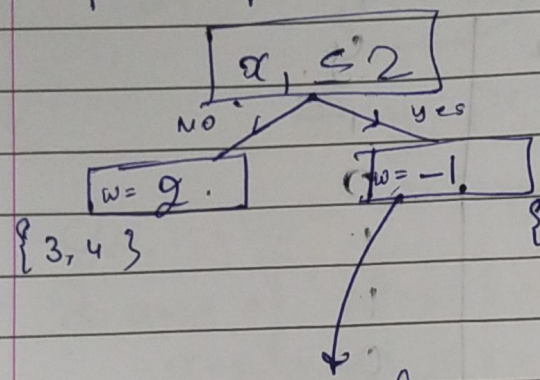
$$= 0$$

⇒ Step 2 :- Compute residuals.

$$r = y - p.$$

Sample	Tree y	intial P=0.5	Residual $r = y - p$
1	1	0.5	0.5
2	0	0.5	-0.5
3	1	0.5	0.5
4	1	0.5	0.5
5	0	0.5	-0.5
6	0	0.5	-0.5

Step 3: fit a decision Stump.



here for regression we put avg of ys but, in classification we can't use that because eg 0.9 + 0.3 → > 1 (Probability is not linearly additive, hence we use concept of log(odds).

hence,

$$w = \frac{\Sigma(y - P)}{\Sigma\left[P(1-P)\right] - \lambda}$$

regularization const.

2nd diff

log(odds) adjustment

$$= \frac{\text{Sum of residuals}}{\text{Sum}\left[\text{Prob} * (1 - \text{Prob})\right]}$$

for yes (1, 2, 5, 6)

$$w = \frac{0.5 + (-0.5) + (-0.5) + (-0.5)}{4\,(0.5\,(0.5))}$$

0.25 ← 4 samples

$$= \frac{-1}{1}$$

$$= -1$$

for left NO

$$w = \frac{1}{2(0.5(0.5))} = \frac{1}{0.5} = \frac{2}{=}$$

Logodds will be here.

→ for residuals also convert to P & substract.

→ $F = F_0(x) + (F_1(x)) + F_2(x)$.

for P from log(odds)

$$P = \frac{1}{1+e^{-z}}.$$

This is for → leaf nodes are having log(odds) → convert into probability & then give labels.

if $P \geq 0.5$ → label = 1

else → 0.

$h_\theta(x) = \theta^T x$

$$\theta = \theta + \alpha \sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)}) x^{(i)}$$

$\phi \to x \to x \in R^n$

$\in R^d$

$$\theta = \theta + \alpha \sum_{i=1}^{m} (y^{(i)} - \theta^T \phi(x^{(i)})) \phi(x^{(i)})$$

$$\phi(x) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_1^2 \\ x_1 x_2 \\ x_1 x_3 \\ \vdots \\ x_3^3 \\ x_1 x_2 x_3 \\ x_1^2 x_2 \end{bmatrix}$$

monomial of x with degree 3.

$d = 1000$

$p \approx O(d^3)$

$1000^3 = 1,000,000,000$

$p \approx 1$ billion.

for update steps we should compute 1 billion computations (1 billion times lower than normal)

Solution

Kernel fn :-

$k(x, y) = \langle \phi(x), \phi(y) \rangle$

dot product

transformed x & y.

Eg : $x = (x_1, x_2, x_3)$ 1st vector

$y = (y_1, y_2, y_3)$ 2nd vector (not target values)

attributes

$$\phi(x) = \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ x_1 x_3 \\ x_2 x_1 \\ x_2 x_2 \\ x_2 x_3 \\ x_3 x_1 \\ x_3 x_2 \\ x_3 x_3 \end{bmatrix}$$

$x = [1, 2, 3]$

$y = [4, 5, 6]$

→ compute → $\langle \phi(x), \phi(y) \rangle = ?$

$$\phi(x) = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 2 \\ 4 \\ 6 \\ 3 \\ 6 \\ 9 \end{bmatrix} \qquad \phi(y) = \begin{bmatrix} 16 \\ 20 \\ 24 \\ 20 \\ 25 \\ 30 \\ 24 \\ 30 \\ 36 \end{bmatrix}$$

$\langle \phi(x) \; \phi(y) \rangle =$

$16 + 40 + 72 + 40 + 100 + 180 + 72 + 180 + 324$

$= \underline{1024}$

But it is computationally expensive so, we

$K(x, y) = (\langle x, y \rangle)^2$

↓ in same space

$= [1, 2, 3] \; [4, 5, 6]$

$= 4 + 10 + 18$

$= (32)^2 = \underline{1024}$

⇒ Kernel methods.

1. Linear Kernel, $K(x, y) = xy$
2. Polynomial.
3. Radial Basis function
4. Laplacian Kernel
5. Graph Kernel.
6. Fisher Kernel.

Testing phase:-

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

$x^{(i)}$ → training sample

$$= \sum_{i=1}^{n} \beta_i \phi(x^{(i)}) \phi(x^{(t)})$$

$x^{(t)}$ → t^{th} instance of test sample

$$= \sum_{i=1}^{n} \beta_i K(x^{(i)}, x^{(t)})$$

dimension → $1 \times n$

Eg:-

x_1	x_2	y
1	2	0
3	4	0

$$\phi(x) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_1 x_2 \end{bmatrix}$$

$x = [x_1, x_2]^T$ ← 2×1 $\quad \beta \rightarrow \beta_1, \beta_2$

$x \in \mathbb{R}^2$.

$\phi(x) \in \mathbb{R}^3 \quad \theta \in \mathbb{R}^3 \rightarrow [\theta_1, \theta_2, \theta_3]$

$$\theta^{3} = \sum_{i=1}^{2} \beta_i \, \phi(x^{(i)})^{3}$$

$$= \beta_1 \left[\right.$$

$$\phi(x_1) = \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} \qquad \phi x_2 = \begin{bmatrix} 9 \\ 16 \\ 12 \end{bmatrix}$$

$$= \beta_1 \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} + \beta_2 \begin{bmatrix} 9 \\ 16 \\ 12 \end{bmatrix}$$

$$= \begin{bmatrix} \beta_1 \\ 4\beta_1 \\ 2\beta_1 \end{bmatrix} + \begin{bmatrix} 9\beta_2 \\ 16\beta_2 \\ 12\beta_2 \end{bmatrix}$$

$$\begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = \begin{bmatrix} \beta_1 + 9\beta_2 \\ 4\beta_1 + 16\beta_2 \\ 2\beta_1 + 12\beta_2 \end{bmatrix}$$

* But in Kernel method we need training set even when we are in Testing phase but in Linear or any other model we just had θ_3.

$$d = \frac{|\beta_0 + \beta_1 x_1 + \beta_2 x_2|}{\sqrt{\beta_1^2 + \beta_2^2}}$$

3 lines :- (hyperplanes)

$2x_1 + 3x_2 - 5 = 0$

$-x_1 + 4x_2 + 7 = 0$

$5x_1 - 12x_2 + 10 = 10$

Datasets

x_1	x_2	y
3	4	+1
2	3	+1
1	-1	-1
-2	+1	-1

→ distance b/w each points.

1) hyperplane

$2x_1 + 3x_2 - 5 = 0$.

$$d = \frac{|2x_1 + 3x_2 - 5|}{\sqrt{2^2 + 3^2}} = \frac{|2x_1 + 3x_2 - 5|}{\sqrt{13}}$$

(3,4) $d_1 = \frac{|2(3) + 3(4) - 5|}{\sqrt{13}} = \frac{13}{\sqrt{13}} = \sqrt{13}$ 3.6

(2,3) $d_2 = \frac{|2(2) + 3(3) - 5|}{\sqrt{13}} = \frac{8}{\sqrt{13}} = 2.218$

(1,-1) $d_3 = \frac{|2(1) + 3(-1) - 5|}{\sqrt{13}} = \frac{6}{\sqrt{13}} = 1.664$

(-2,1) $d_4 = \frac{|2(-2) + 3(1) - 5|}{\sqrt{13}} = \frac{6}{\sqrt{13}} = 1.664$

2) hyperplane $-x_1 + 4x_2 + 7 = 0$

$$d_4 = \frac{|-1x_1 + 4x_2 + 7|}{\sqrt{(-1)^2 + (4)^2}} = \frac{|-x_1 + 4x_2 + 7|}{\sqrt{17}}$$

(3,4) $d_1 = \frac{|-3 + 4(4) + 7|}{\sqrt{17}} = \frac{20}{\sqrt{17}} = 4.85$

(2,3) $d_2 = \frac{|-2 + 4(3) + 7|}{\sqrt{17}} = \frac{17}{\sqrt{17}} = 4.123$

(1,-1) $d_3 = \frac{|-1 + 4(-1) + 7|}{\sqrt{17}} = \frac{2}{\sqrt{17}} = 0.485$

(-2,1) $d_4 = \frac{|2 + 4(1) + 7|}{\sqrt{17}} = \frac{13}{\sqrt{17}} = 3.152$

3) hyperplane.

$$5x_1 - 12x_2 + 10 = 0$$

$$d = \frac{|5x_1 - 12x_2 + 10|}{\sqrt{5^2 + (-12)^2}} = \frac{|5x_1 - 12x_2 + 10|}{\sqrt{169}} = \frac{|5x_1 - 12x_2 + 10|}{13}$$

$(3,4)$ $d_1 = \frac{|5(3) - 12(4) + 10|}{13} = \frac{+23}{13} = 1.769$

$(2,3)$ $d_2 = \frac{|5(2) - 12(3) + 10|}{13} = \frac{16}{13} = 1.230$

$(1,-1)$ $d_3 = \frac{|5(1) - 12(-1) + 10|}{13} = 2.076$

$(-2,1)$ $d_4 = \frac{|5(-2) - 12(1) + 10|}{13} = 0.923$

$\Rightarrow \max(M_1, M_2, M_3) = \max\{1.664, 0.485, 0.92\}$

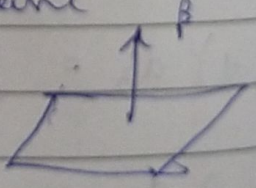
$M_2 = 0.485$ 1.664

Refresh linear algebra concepts:-

① projection
② Norm of a vector
③ vector addition
④ orthogonality
⑤ Subspaces

to prove that a point β_0 is always $\perp^{lar}$ to the plane

take two pts. x_1, x_2



$\beta \cdot x_1 = 0$

$\beta x_2 = 0$

21/3/25

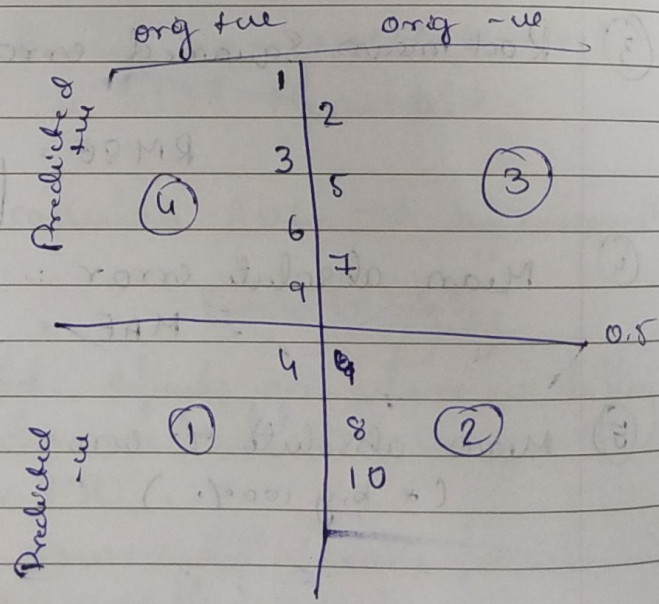
M T W T F S S
Page No.:
Date:
YOUVA

⇒ Eg: Evaluation metrics. Threshold → 0.5

observation	Actual labels	Predicted score	Predicted Label.
1	1	0.85	1
2	0	0.60	1 * (FP)
3	1	0.70	1
4	1	0.40	0 * (FN)
5	0	0.55	1 * (FP)
6	1	0.50	1
7	0	0.65	1 * (FP)
8	0	0.35	0
9	1	0.60	1
10	0	0.20	0

1) Accuracy ?
2) Precision ?
3) Recall or Sensitivity or True positive Rate
4) Specificity
5) False +ve rate (1 - Specificity)
6) F_1 - Score
7) ROC Plot
8) AUC

⇒ Accuracy



- TP = 4
- TN = 2
- FP = 3
- FN = 1

M T W T F S S

Page No.: YOUVA

Date:

1) $Acc = \frac{TP + TN}{Total} = \frac{4+2}{10} = \frac{6}{10} = 0.6$

2) Precision :- $\frac{TP}{TP+FP} = \frac{4}{4+3} = \frac{4}{7} = 0.571$

3) Sensitivity / Recall / TPR. $= TP / TP + FN = \frac{4}{4+1} = \frac{4}{5} = 0.8$

4) Specificity $= TN / (TN+FP) = \frac{2}{2+3} = \frac{2}{5} = 0.4$

5) False +ve ratio $= 1 - \text{specificity} = 1 - 0.4 = 0.6$

6) F1 Score $= 2 * \frac{\text{Precision} * \text{recall}}{\text{precision} + \text{recall}} = \frac{2 * 0.571 * 0.8}{0.571 + 0.8} = \frac{2 * 0.457}{1.371} = 0.666$

7) ROC

Thresholds	TP	TN	FP	FN	TPR	FPR
0.7	2	5	0	3	0.4	0
0.5	4	2	3	1	0.8	0.6
0.4	5	2	3	0	1	0.6
0.2					1	1

→ 0.7 → TP → 2, TN → 5, FN → 3, FP → 0

$TPR \Rightarrow \frac{TP}{TP+FN} = \frac{2}{2+3} = \frac{2}{5} = 0.4$

$FPR \Rightarrow 1 - \frac{TN}{TN+FP} = 1 - \frac{5}{5+0} = 0$

→ 0.4 → TP → 5, TN → 2, FN → 0, FP → 3

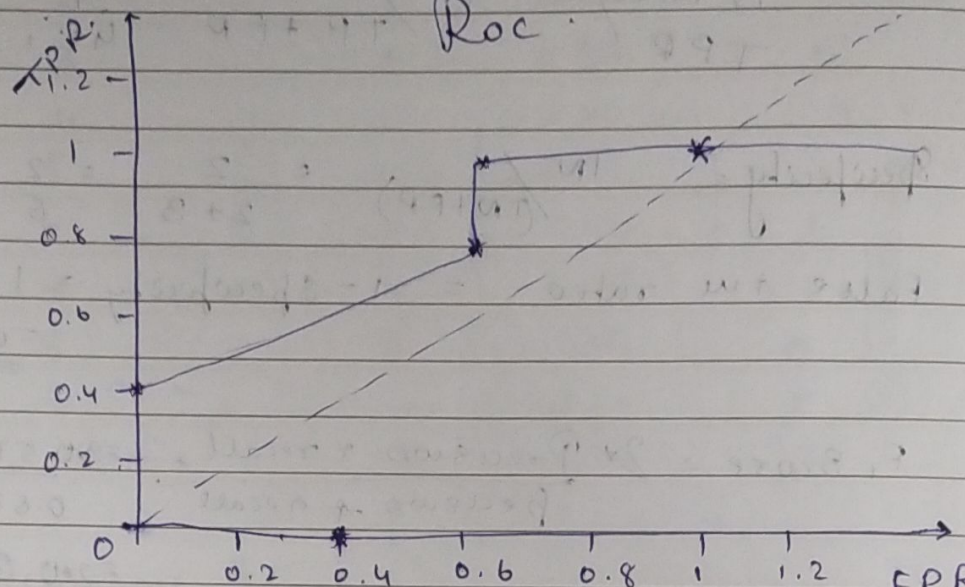
$TPR = \frac{5}{5} = 1$ $FPR = 1 - \frac{2}{2+3} = 1 - \frac{2}{5} = 0.6$

0.7
1 TP
0 TN
1 TP
0 FN
0 TN
0 FN
0 TN
0 TN
0 FN
0 TN

0.2 → TP = 0 , TN = 4 , FP = 5 , FN = 1

$TPR = \frac{TP}{TP+FN} = \frac{0}{0+1}$

$FPR = 1 - \frac{TN}{TN+FP} = 1 - \frac{4}{4+5} = 1 - 0.555 = 1 - 0$



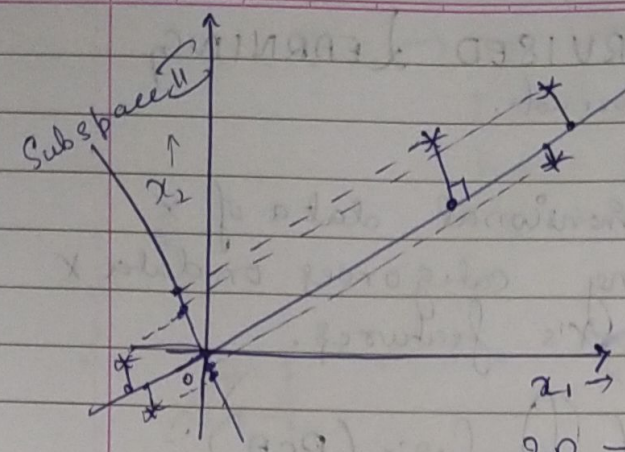
$$AUC = \sum_{i=1}^{4} (FPR_i - FPR_{i-1}) * \left(\frac{TPR_i + TPR_{i-1}}{2}\right)$$

$$= \left((0 - 0.6)*(0.4 - 0.8) + (0.6 - 0.6)(0.8 - 1) \right) + \frac{(0.6 - 0)(1 - 1)}{2}$$

$$= \frac{0.6 * 0.4}{2} = 0.$$

0.76

$$AUC = \sum_{i=1}^{n-1} (FPR_{i+1} - FPR_i) \cdot \left(\frac{TPR_{i+1} + TPR_i}{2}\right)$$



2D ⟶ 1D:

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2 \Rightarrow z = [z_1] \in \mathbb{R}^1$$

We are going to linearly project each point on that subspace (linear line). Orthogonal projection

- Eg: 2 subspace I & II.

A good projection is to maximize variance

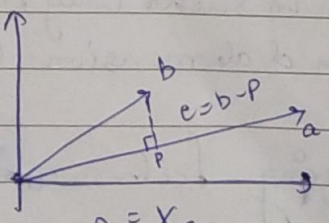
in subspace I ⟶ variance is more b/w the projected points as distance is more

but in subspace II ⟶ it's less.

So, subspace I is more better than II.
∴ It captures more info.

Find that subspace among all possible spaces where my projected points have longer variance

Linear algebra



we are trying to project b on a & to find P ⟶

- a ⊥ or to e.

$a^T . e = 0$ (Dot product)

$a^T . (b - P) = 0$

$$a^T(b - xa) = 0$$

$$a^T b = x a^T a$$

$$\boxed{x = \frac{a^T b}{a^T a}} \rightarrow \text{scalar.}$$

$$p = xa.$$

$$\boxed{p = \frac{a\, a^T b}{a^T a}} \rightarrow \text{projection matrix } \frac{a a^T}{a^T a}$$

$$\boxed{p = A(A^T A)^{-1} A^T b}$$

Eg: $b = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$ $\quad a = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ $\quad p = ?$

$$p = A(A^T A)^{-1} A^T b.$$

$$= \begin{bmatrix} 1 \\ 2 \end{bmatrix} \left(\begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right)^{-1} b$$

$$x = \frac{a^T b}{a^T a} \rightarrow = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 1+4 \end{bmatrix}^{-1} b$$

$$= \frac{\begin{bmatrix} 1 \\ 2 \end{bmatrix}_{2\times 1} \begin{bmatrix} 1 & 2 \end{bmatrix}_{1\times 2} \begin{bmatrix} 3 \\ 4 \end{bmatrix}}{5} = \frac{8}{5} \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

$$= \frac{\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}_{2\times 2} \begin{bmatrix} 3 \\ 4 \end{bmatrix}_{2\times 1}}{5} = \begin{bmatrix} 3+8 \\ 6+16 \end{bmatrix} = \begin{bmatrix} 11 \\ 22 \end{bmatrix} = \begin{bmatrix} 2.2 \\ 4.4 \end{bmatrix}$$

$$= \frac{\begin{bmatrix} 3+8 \\ 6+16 \end{bmatrix}}{5} = \frac{\begin{bmatrix} 11 \\ 22 \end{bmatrix}}{5}$$

→ continue PCA :-

u to be subspace of unit lengths

$\vec{x}$ → point to be projected.

b = projected pt of $\vec{x}$ = $\text{proj}(u) \cdot \vec{x}$

Algorithm : Kmean (K, dataset) {

1) Randomly assign a number from 1 to K to each observation. These serve as initial cluster assignments.

2) Iterate until the cluster assignments stop changing {
   a) for each of the K-clusters, compute the cluster centroid.
   b) Assign each observation to the cluster whose centroid is closest (based on Euclidean distance)
}}

Example:

Eg:	x_1	x_2
1	1	1
2	1	2
3	2	2
4	8	8
5	8	9
6	9	8

Given:
K = 2.

1) $C_1 = \{2, 3, 4, 6\}$
   $C_2 = \{1, 5\}$

Step 1:- $C_1 = \begin{bmatrix} 5.83 \\ 5 \end{bmatrix} \rightarrow$ $x_1 = \frac{23}{4} = 4.83$ 5
$x_4 = \frac{20}{4} = 5$

$C_2 = \begin{bmatrix} 4.5 \\ 5 \end{bmatrix}$ $x_1 = 9/2 = 4.5$
$x_2 = 10/2 = 5$

Step 2:- $d_{11} = ?$ {
$d_{12} = ?$

C_1 $d_{11} = \frac{1}{2}\sqrt{(5-1)^2 + (5-1)^2} = d_{12} = C_2$
$= \sqrt{32} = 5.65$

$d_{12} = \frac{1}{2}\sqrt{(4.5-1)^2 + (5-1)^2} = 5.31$ ⇒

$$d_{21} = \sqrt{(5-1)^2 + (5-2)^2}$$
$$= \sqrt{16+9} = 5$$

$$d_{22} = \sqrt{(4.5-1)^2 + (3.5-2)^2}$$
$$= \sqrt{} = 4.6$$

$$\rightarrow \quad d_{31} = \sqrt{(5-2)^2 + (5-2)^2}$$
$$= \sqrt{18} = 4.24$$

$$d_{32} = \sqrt{(4.5-2)^2 + (5-2)^2}$$
$$= 3.905$$

$$\rightarrow \quad d_{41} = \sqrt{(5-8)^2 + (5-8)^2}$$
$$= \sqrt{9+9} = 4.24$$

$$d_{42} = \sqrt{(4.5-8)^2 + (5-8)^2}$$
$$= 4.609$$

$$\rightarrow \quad d_{51} = \sqrt{(5-8)^2 + (5-9)^2} = 5$$

$$d_{52} \quad \sqrt{(4.5-8)^2 + (5-9)^2} = 5.31$$

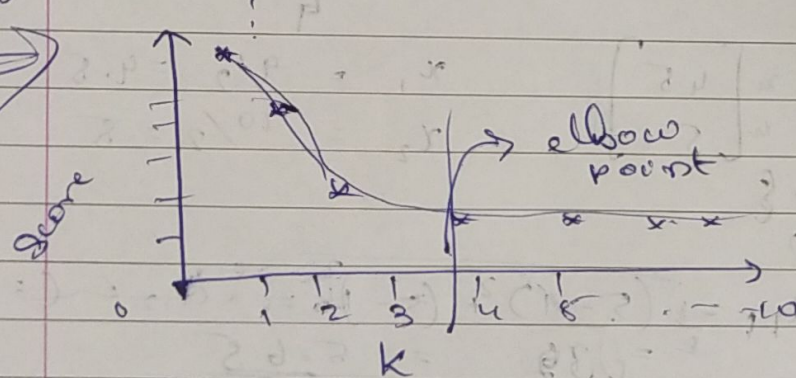
After 1st Iteration.

$C_1 = \{4, 5, 6\}$

$C_2 = \{1, 2, 3\}$

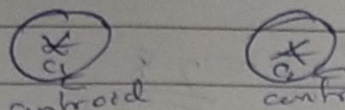
Cal centroid.

2/4/20



⇒ when to stop adding of k. is not making much difference in C_3.

→ Centroid linkage.



Dissimilarity b/w Dot product:

- Dot product → similarity
- 1 − Dot product → dissimilarity.

	x_1	x_2
A	1	1
B	2	1
C	4	3
D	5	4
E	6	5

Single linkage.
Euclidean distance.

Step 1:- compute the distance matrix

	A	B	C	D	E
A	0	1	3.61	5	6.4
B	1	0	2.8	4.2	5.6
C	3.61	2.8	0	1.41	2.8
D	5	4.2	1.41	0	1.41
E	6.4	5.6	2.8	1.41	0

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

$= \sqrt{(1)^2 + (0)^2} = 1$

$d_{BC} = \sqrt{(3)^2 + (2)^2} = \sqrt{13} = 3.61$

$d_{AD} = \sqrt{4^2 + 3^2} = \sqrt{25} = 5$

$d_{AE} = \sqrt{5^2 + 4^2} = \sqrt{42} = 6.4$

$d_{BC} = \sqrt{4 + 4} = \sqrt{8} = 2.8$

$d_{BD} = \sqrt{9 + 9} = \sqrt{18} = 4.2$

$d_{CD} = \sqrt{1+1} = \sqrt{2} = \underline{1.41}$

$d_{CE} = \sqrt{4+4} = \sqrt{8} = \underline{2.8}$

$d_{DE} = \sqrt{1+1} = \underline{1.41}$.

→ Minimum → AB.

merge	(AB)	C	D	E
(AB)	0	2.82	4.2	5.6
C	2.82	0	1.41	2.8
D	4.2	1.41	0	1.41
E	5.6	2.8	1.41	0

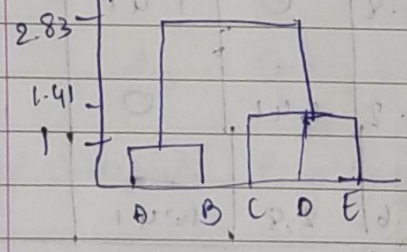
$AB \to C = \min(d_{AC}, d_{BC})$

$= \min(3.61, 2.8)$

$= \underline{2.8}$

repeat ↓

finally :→



4/4/25

M T W T F S S
Page No.: YOUVA
Date:

Naive Bayes Classifier :- (usually used in spam detection)

→ Input → Text modality.

Eg :- "To buy this product" ⇒ f (by looking @ context)

Some algo to convert so, that we can do further process

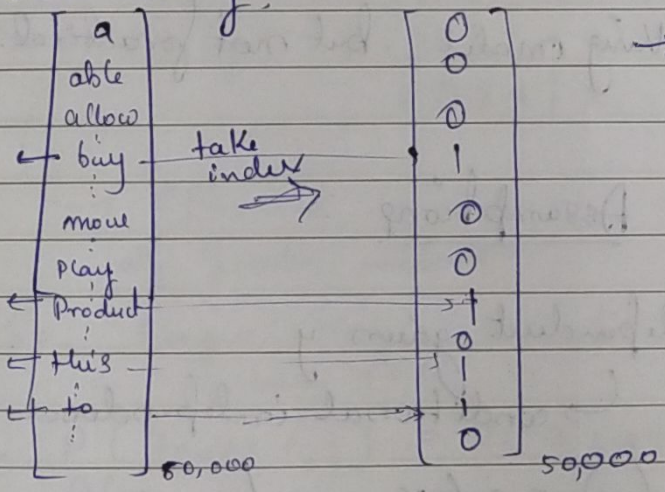
$$\begin{bmatrix} 1.0 \\ 7.8 \\ 4.3 \\ \vdots \\ \vdots \end{bmatrix}_{10 \times 1}$$

* It should capture the Semantic meaning of the sentence.

Eg :- King , empire

representation should be closer

dictionary.



→ One way to preprocess.

* take index from dictionary.
* Put 1 corresponding index of the words of sentences & rest all are zeros.

Eg :-	x_1	x_2	x_3	class
1	1	0	1	Spam
2	1	1	0	Spam
3	1	1	0	Spam
4	0	0	1	No Spam
5	0	0	0	No Spam
6	1	1	1	No Spam
7	0	1	0	Spam
8	1	0	0	Spam
9	0	1	0	Spam
10	1	0	0	Spam
11	1	1	0	No Spam.

2^3 → distinct combinations

x_1	x_2	x_3
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	1	0
1	0	1
1	1	1

→

	x_1	x_2	x_3	Prob spam	Prob no spam
1	0	0	0	0	1/4
2	0	0	1	0	1/4
3	0	1	0	2/7	0
4	0	1	1	2/7 0	0
5	1	0	0	2/4	0
6	1	0	1	1/7	0
7	1	1	0	2/7	1/4
8	1	1	1	0	1/4

$2^3 - 1 = 7$

Problem with this approach is if features → 1 million
possible combinations → $2^{1\text{million}}$.

∴ It is a working model, but not practical.

⇒ Naive Bayes Assumption :-

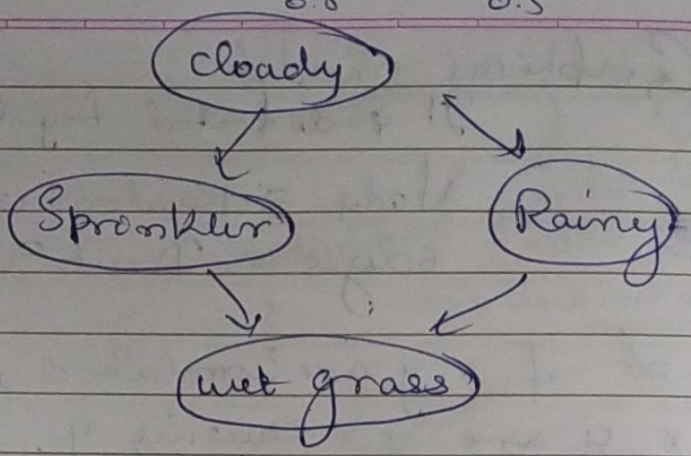
x_i's are independent given y

↳ conditional independence.

Eg :- $P(x_{2087}/y) = P(x_{2087}/y, x_{38193})$

↳ for given y, x_{38193} is not influencing x_{2087}

↳ Eg for given $y = 1$ (spam), "this" is not influencing "buy".



P(R/C)

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

P(S/C)

C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

P

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

12/4/25

Inference, phase:-

$$\Rightarrow P(S=1) = \sum_{C=0}\sum_{R=0}\sum_{W=0} P(C,R,W,S=1)$$

$$= \sum_{C}\sum_{R}\sum_{W} P(C)P(R/C)P(S=1/C)P(W/S,$$

$$= (0.5)(0.2+0.8)($$

$$= P(C=0)P(S=1/C=0)P(R=0/C=0)$$
$$P(W=0/S=1, R=0)$$
$$+ P(C=0)P(S=\phi/C=0)P(R=0/C=0)P($$
$$S=1$$

CRW
000
001
010
100
101
110
011
111

Continued 2 pgs later

as $C \uparrow^{Sc} \longrightarrow M \uparrow \longrightarrow$ ↓ Sensitivity.

$M \uparrow \longrightarrow$ ↓ bias ↑ variance

$M \downarrow \longrightarrow$ ↑ bias ↓ variance.

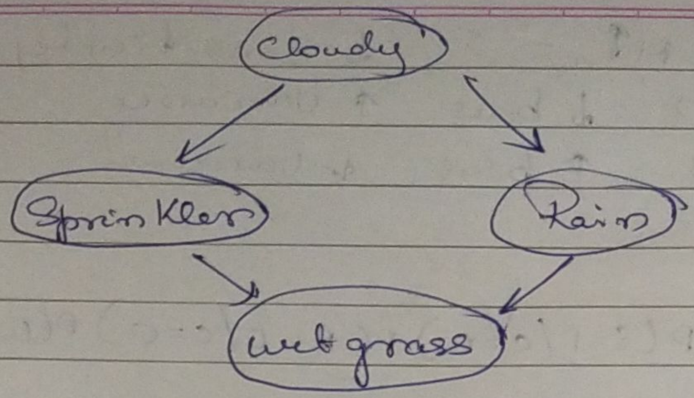
12/4/26

$$P(C=0)\,P(S=1/C=0)\,P(R=0/C=0)\,P(W=0/S=1,R=0)$$
$$+ P(C=0)\,P(S=1/C=0)\,P(R=0/C=0)\,P(W=1/S=1,R=0)$$
$$+ P(C=0)\,P(S=1/C=0)\,P(R=1/C=0)\,P(W=0/S=1,R=1)$$
$$+ P(C=1)\,P(S=1/C=1)\,P(R=0/C=1)\,P(W=0/S=1,R=0)$$
$$+ P(C=1)\,P(S=1/C=1)\,P(R=1/C=1)\,P(W=0/S=1,R=1)$$
$$+ P(C=1)\,P(S=1/C=1)\,P(R=0/C=1)\,P(W=1/S=1,R=0)$$
$$+ P(C=0)\,P(S=1/C=0)\,P(R=1/C=0)\,P(W=1/S=1,R=1)$$
$$+ P(C=1)\,P(S=1/C=1)\,P(R=1/C=1)\,P(W=1/S=1,R=1)$$

0.3

$$\Rightarrow (0.5 * 0.5 * 0.8 * 0.1) + (0.5 * 0.5 * 0.8 * 0.9) + (0.5 * 0.5 * 0.2 * 0.01) + (\ldots$$

Learning phase.



θ	1-θ

	F	T
F	α_1	$1-\alpha_1$
T	α_2	$1-\alpha_2$

	F	T
F	ν_1	$1-\nu_1$
T	ν_2	$1-\nu_2$

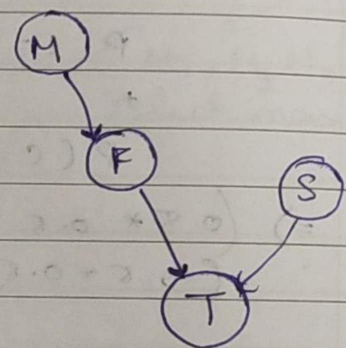
S	R	F	T
F	F	β_1	$1-\beta_1$
F	T	β_2	$1-\beta_1$
T	F	β_3	$1-\beta_2$
T	T	β_4	$1-\beta_3$

05/4/25

→ t-SNE - non linear approach like PCA (↓ linear)

⇒ write pseudocode - PCA, K-fold + Kmeans.

M (movie)	F (food)	S (study)	T (Twin)
1	1	0	0
1	0	1	1
1	1	0	1
1	1	1	1
0	0	0	0



$$P(M, F, S, T) = P(M)\,P(F/M)\,P(T/F,S)\,P(S)$$

build multiple model (graph/Tree). Select the model which gave highest joint probability.

⇒ Ways to learn parameters ($\alpha, \beta, \gamma, \theta$'s)

1) MLE approx
2) Bayesian learning approach

1) MLE

$P(M/\theta) = \theta^4(1-\theta)^1$ ← no of times $m=1$

$P(S/\alpha) = \alpha^2(1-\alpha)^3$

$P(F/M=0) = \gamma_1^0(1-\gamma_1)^1$

$P(F/M=1) = \gamma_2^3(1-\gamma_2)^1$

$P(T/F=0, S=0) = \beta_1^0(1-\beta_1)^1$

$P(T/$

↓

write joint probability → differentiate → equate to 0 → Maximize

Likelihood → find log

$\hat{\theta}_{MLE} = \frac{4}{5}$ γ_1

$\hat{\alpha}_{MLE} = \frac{2}{5}$ β

Bayesian learning approach's -

⇒ we use prior knowledge.

→ MLE → frequentist approach.

Step 1:- Given $x_{1\ldots n} = \{x_1; x_2; x_3; \ldots x_n\}$ write down the expression likelihood.

Step2:- Specify a prior, $P(\theta)$

Step3:- Compute posterior, $P(\theta/x_{1\cdots n}) = \frac{P(x_{1\cdots n}/\theta)}{P(x_{1\cdots n}}$

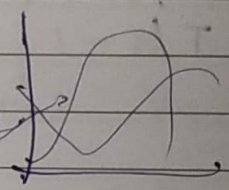
$\propto P(x_{1\cdots n}/\theta)$

$P(\theta)$

→ Beta distribution :- used to choose the parameter

for prior knowledge we can use beta distribution to capture it

Eg:- B(1,1) or B(2,2).



Let $p(\alpha) = \text{Beta}(10, 1)$

$$P(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

↑ parameters of Beta(θ)

$P(\alpha) = \text{Beta}(10,1) = \alpha^{10-1}(1-\alpha)^{1-1}$

Prior rule,

$= \alpha^9(1-\alpha)^0$

$P(\alpha/S) \propto P(S/\alpha)\,P(\alpha)$

$\approx \alpha^2(1-\alpha)^3\,[\alpha^9(1-\alpha)^0]$

$P(\alpha/S) \approx \boxed{\alpha^{11}(1-\alpha)^3} \Rightarrow E[\alpha/S] = \frac{11}{11+3} = \frac{11}{14}$

≈ 0.7

So, the α value obtained by MLE was 0.4,

Bayesian → 0.7 (for B(10,1))