Q2

Solution

Hypothesis :-

$h_\theta(x^{i}) = \theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)}$

$x_0 = 1$

$\theta_0 = \theta_1 = \theta_2 = 0$

$h_\theta x^{(1)} = 0 \qquad h_\theta x^{(3)} = 0$

$h_\theta x^{(2)} = 0$

$J(\theta) = \frac{1}{2}\left[(0-3)^2 + (0-4)^2 + (0-5)^2\right]$

$= \frac{1}{2}[9+16+25] = \boxed{25}$

Gradient

i) $\frac{\partial J}{\partial \theta_0} = \sum_{i=1}^{3}(h_\theta(x^{(i)}) - y^{(i)})x_0^{(i)}$

$= (0-3)1 + (0-4)1 + (0-5)1$

$= -3-4-5$

$\frac{\partial J}{\partial \theta_0} = -12$

update parameters

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

$:= 0 - 0.1 \times (-12)$

$= 0 + 1.2 = 0.12$

ii) $\frac{\partial J}{\partial \theta_1} = (0-3)1 + (0-4)2 + (0-5)3$

$= -3-8-15$

$= -26$

update

$\theta_1 := \cancel{2.6}\ 0.26$

iii) $\frac{\partial J}{\partial \theta_2} = (0-3)2 + (0-4)1 + (0-5)3$

$= -6-4-15$

$= -25$

$\theta_2 := ~~2.5~~ \ (0.25)$ <u>update</u>

$J(\theta) = \frac{1}{2}$.

(2) $h_\theta(x)^{(1)} = ~~1.2~~ \ 2 \times 1 + ~~2.6~~ \times 1 + ~~2.5~~ \times 2 = ~~8.8~~ \ 0.88$

$h_\theta(x)^{(2)} = 1.2 \times 1 + 2.6 \times 2 + 2.5 \times 1 = ~~8.9~~ \ 0.89$

$h_\theta(x)^{(3)} = 1.2 \times 1 + 2.6 \times 3 + 2.5 \times 3 = ~~16.5~~ \ 1.65$

$J(\theta) = \frac{1}{2}\left[(~~8.8~~ \ 0.88 - 3)^2 + (~~8.9~~ \ 0.89 - 4)^2 + (~~16.5~~ \ 1.65 - 5)^2\right]$

$= \frac{1}{2}\left[4.5 + 9.67 + 11.22\right]$

$\boxed{J(\theta) = 12.695}$

<u>Gradient</u>

$\frac{\partial J}{\partial \theta_0} = (0.88 - 3)1 + (0.89 - 4)1 + (1.65 - 5)1$

$= -2.12 - 3.11 - 3.35$

$= -8.58$

$\frac{\partial J}{\partial \theta_1} = (0.88 - 3)1 + (0.89 - 4)2 + (1.65 - 5)3$

$= -2.12 - 6.22 - 10.05$

$= -18.39$

$\frac{\partial J}{\partial \theta_2} = ~~-2.12~~ - 4.24 - ~~6.22~~ \ 3.11 - 10.05$

$= -17.4$

$\theta_0 := 0.12 - 0.05 \times (-8.58)$

$= 0.1629$

$\theta_1 := 0.26 - 0.05 \times (-18.39)$

$= ~~0.4459~~ \ 1.1795$

$\theta_2 := 0.25 - 0.05 \times (-12.695)$

$:= 0.88475$

$h_\theta(x)^1 = 3.1129$ $h_\theta(x)^{(2)} = 3.40$ $h_\theta(x)^{(3)}$

$h_\theta(x)^{(4)} = 6.3556$

Bernoulli distribution :- [Discrete]

[simplest so we choose that]
parametrized by only θ

$$P(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \begin{cases} \theta & x_i = 1 \\ 1-\theta & x_i = 0 \end{cases}$$

$\{1,1,1,1,1,1,1,0,0$

when

$x_i = 1$ then;

$$P(1|\theta) = \theta^{1}(1-\theta)^{1-1}$$

$$P(1|\theta) = \theta$$

$$P(x_{1:n}|\theta) = \prod_{i=1}^{n} P(x_i|\theta)$$

SGD GD MLE

$X = \{1, 1, 0\}$

$P(x_{1:3} | \theta) = \theta^1 (1-\theta)^{1-1} *$

$\times \theta^1 (1-\theta)^{1-1} *$

$\times \theta^0 (1-\theta)^{1-0}$

$= \theta^2 * (1-\theta)$

$= \theta^2 - \theta^3$

$m \rightarrow$ no of success

$n \rightarrow$ no of trials.

probability of success

$P(x_{1:n} | \theta) = \theta^m (1-\theta)^{n-m}$

① $L(\theta) = P(x_{1:n} | \theta)$

$= \prod_{i=1}^{n} P(x_i | \theta) \Rightarrow$ since iid assumption

② compute log likelihood

$l(\theta) = \log L(\theta)$

$= \log \prod_{i=1}^{n} P(x_i | \theta)$

$= \log \prod_{i=1}^{n} [\theta^m (1-\theta)^{n-m}]$

$l(\theta) = m\log\theta + (n-m)\log(1-\theta)$

(3) Differentiate & equate to zero to find θ

$$\frac{d\,l(\theta)}{d\theta} = \frac{d}{d\theta}\left[m\log\theta + (n-m)\log(1-\theta)\right]$$

$$= \frac{m}{\theta} + \frac{n-m}{1-\theta}(-1)$$

$$= \cancel{\frac{m}{\theta}}$$

$$= \frac{m(1-\theta)-(n-m)\theta}{\theta(1-\theta)}$$

$$= \frac{m - \cancel{m\theta} - n\theta + \cancel{m\theta}}{\theta(1-\theta)}$$

$$0 = \frac{m-n\theta}{\theta(1-\theta)}$$

$$m - n\theta = 0$$

$$\boxed{\theta = \frac{m}{n}}$$

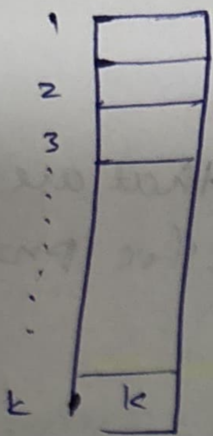
$$\boxed{\theta = \frac{7}{10}}$$

$$\boxed{\theta = 0.7}$$

→ Programme it in labs

24/01/25

Cross validation k fold :- [k fold cross validation]



Example :-

	x_1	x_2	y
1	2	-1	1
2	0.5	1.2	0
3	1	2	1
4	-3	-2	1
5	4	0.1	0

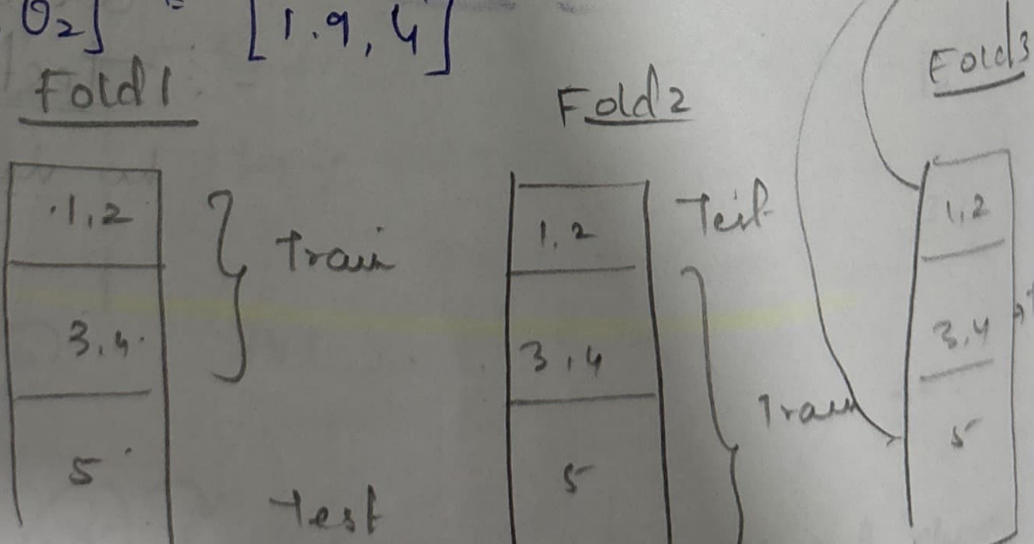
3 fold cross validation

Acc?
std?

F1 $[\theta_1, \theta_2] = [-1.8, 2.8]$

F2 $[\theta_1, \theta_2] = [2.1, 3.1]$

F3 $[\theta_1, \theta_2] = [1.9, 4]$



Test dataset

4	0.1	0

$\theta^T x = -1.8 \times 4 + .28 \times 0.1$

$\theta^T x = -6.92$

$g(z) = 0.0009$

$g(z)$

$\downarrow$ Threshold 0.5

$\hat{y} \quad [0]$

$y = [0]$

100%

we dont have to do it for training dataset cause parameters is already given, check for test dataset

Fold 2

Test-dataset

$$\theta^T x = \begin{bmatrix} 2 & -1 \\ 0.5 & 1.2 \end{bmatrix} \begin{bmatrix} 2.1 \\ 3.1 \end{bmatrix}$$

$$\theta^T x = \begin{bmatrix} 1.1 \\ 4.77 \end{bmatrix}$$

$$g(z) = \begin{bmatrix} 0.249 \\ 0.0431 \end{bmatrix}$$

$$\hat{y} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad y = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

50%

Fold 3

Test dataset

$$\theta^T x = \begin{bmatrix} 1 & 2 \\ -3 & -2 \end{bmatrix} \begin{bmatrix} 1.9 \\ 4 \end{bmatrix}$$

$$\theta^T x = \begin{bmatrix} 9.9 \\ -13.7 \end{bmatrix}$$

50%

$$g(z) = \begin{bmatrix} 0.999 \\ 0.0001 \end{bmatrix} \longrightarrow \hat{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad y = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

average $\Rightarrow$ $66.667\% = \text{mean} = \bar{x}$

$$s.d = \sqrt{\frac{(100-66.67)^2 + (50-66.67)^2 + (50-66.67)^2}{3-1}}$$

$$s.d = 35.35\%$$

General practical design consideration

Example walkthrough:-

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

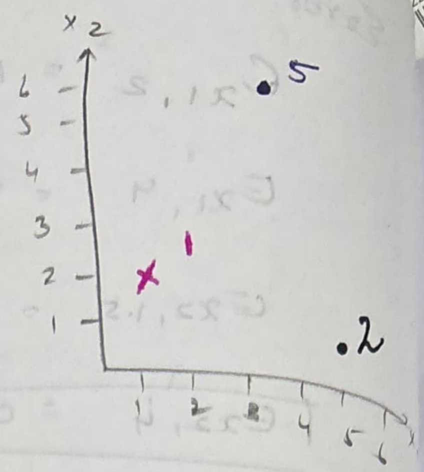
β

1

2

Neal

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



→ x_1 :- $K = \{2, 4\}$ $I = \{1, 2, 3\}$

$K = 2$

{

$I^{+}_{x_1, 2} = \{2, 3\}$

$I^{-}_{x_1, 2} = \{1\}$

$\hat{y}^{+} = \frac{2+5}{2} = 3.5$

$\hat{y}^{-} = \frac{1}{1} = 1$

$\Sigma = (3.5-2)^2 + (5-3.5)^2 + (1-1)^2$

$= 4.5$

$\boxed{\Sigma = 4.5}$

}

$\boxed{x_1 \geq 2}$

yes

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

$E = 4.5$

$k = 4$

{

$I^{+}_{x_1, 4} = \{3\}$

$I^{-}_{x_1, 4} = \{1, 2\}$

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

$\hat{y}^{-} = 3$

$E^{+} = (2-2)^2 = 0 \mid E^{-} = (3-3)^2 + (5-3)^2 =$

$\boxed{x_1 \geq 4}$

no yes

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

$E = 8$

$E^+ : 8$

$\boxed{E = 8}$

}

→ x_2 :- $K = \{1.5, 4\}$.

$K = 1.5$

{

$E^+_{x2,1.5} = \{1, 2\}$.

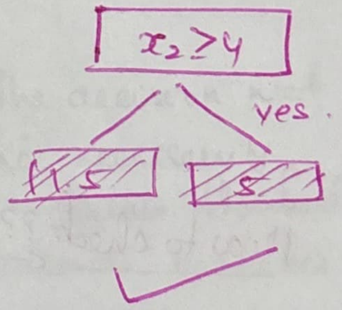
$I^-_{x2,1.5} = \{3\}$.

$y^+ = 3$

$y^- = 2$

$\boxed{E = 8 - 0 = 8}$

}



$K = 4$

{

$I^+_{x2,4} = \{2\}$

$I^-_{x2,4} = \{1, 3\}$.

$y^+ = 5$

$y^- = 1.5$

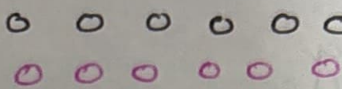
$\boxed{E = 0.5}$

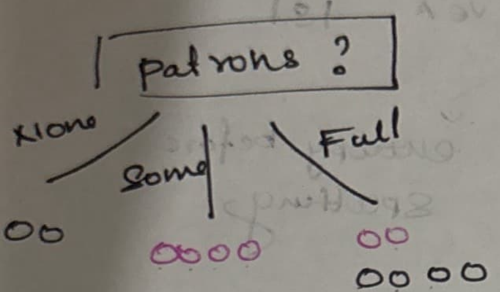
}

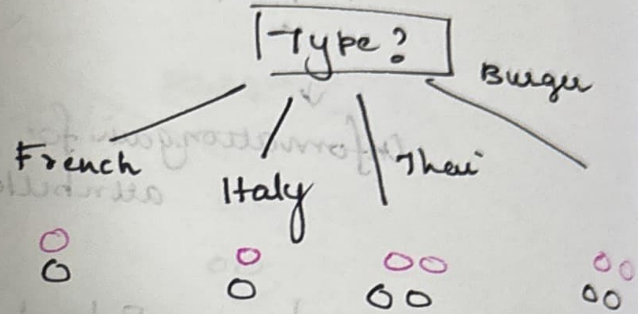
$j^*, s^* = \min\{E_{x1,2}, E_{x1,4}, E_{x2,1.5}, \underline{E_{x2,4}}\}$

$j^*, s^* = x2, 4$

Example:-







entropy of patrons:-

Expected entropy

total datapoints = 12

$$EH(\text{patrons}) = \frac{2}{12} H\left(\frac{0}{2}, \frac{2}{2}\right) + \frac{4}{12} H\left(\frac{4}{4}, \frac{0}{4}\right) + \frac{6}{12} H\left(\frac{2}{6}, \frac{4}{6}\right)$$

~~$EH(\text{patrons}) = -\frac{2}{12}(0.6931) - \frac{4}{12}(0.6931) - \frac{6}{12}$~~

[OR]

$$= \frac{2}{12}\left(-0 - 1\log_2(1)\right) + \frac{4}{12}\left(-1\log_2 1\right) + \frac{6}{12}\left(-\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6}\right)$$

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

$$\boxed{= H\left(\frac{6}{12}, \frac{6}{12}\right) = 1}$$

$$EH(\text{patrons}) = 0 + 0 + \frac{6}{12}(0.366 + 0.\ldots)$$

$$= 0 + 0 + \frac{1}{2}\left(\frac{1}{3} \times 1.5849 + \frac{2}{3} \times 0.5\ldots\right)$$

$$= \boxed{0.456}$$

IG = 1 − 0.456

$\boxed{IG = 0.544}$ → IG for patrons.

IG for type = 0

$\boxed{\text{patrons IG} > \text{type IG}}$

→ take patrons

on which feature should split

note:- if there is a tie, choose either one

Examples:-

	Age	Salary	Spend Pattern	Buy Product
1	16	0	High	Yes
2	30	90000	Low	No
3	31	90000	High	No
4	55	5,00,000	~~Low~~ Medium	No

→ B = 2.

B1 = {2, 3, 3, 4}

B2 = {1, 1, ~~4~~ 1, 3}

→ construct a tree

choose 2 features. {age, salary}

Age:-

for each value in feature

FG = FG & value. append

→ IG of Age = 1.2

IG of Sal = 1.8

→ Choose Salary as node

e.g.

Examples:-

	Age	Salary	Spend pattern	Buy Product
1	16	0	High	Yes
2	30	90000	Low	No
3	31	90000	High	No
4	55	5,00,000	~~Low~~ Medium	No

→ B = 2.

$B_1 = \{2, 3, 3, 4\}$

$B_2 = \{1, 1, \cancel{4}\ 1, 3\}$

→ construct a tree

choose 2 features. {age, salary}

Age:-

for each value in feature

IG = IG of value. append

→ IG of Age = 1.2

IG of Sal = 1.8

→ choose salary as node

mean metric . e. O

Step 1

f_1

$b = 1$

$\boxed{x_1 \geq 2}$

Yes No.

$r = y$

$f_1(x) \rightarrow r \cdot (y)$

$\frac{-1.5 + 8.1}{2}$ 3.5

$F(r; \theta) = 0$ 8.5

$+ 0.1$ $+3.3$

$f_0(r, 0)$

x_1	x_2	y	r
1	2	8	8.1
2	8	-1.5	-1.5
3	11	3.8	3.8

$\boxed{0 < \lambda < 1}$ → hyperparameter

λ is const across all models.

every f is a classifier

how much weight should I give the classifier

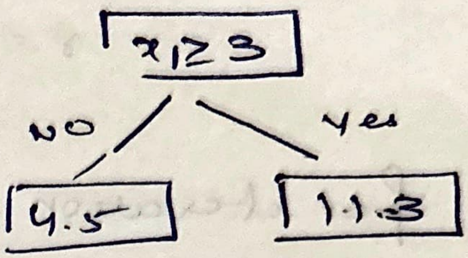
Examples:-

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

step1 :-

$r = y$

a) Fit a model.



b, c)

$\lambda = 0.1$

x_1	y	r	λf_b	$\hat{y} = \hat{y} + \lambda f_b$	r_new
1	3	3	0.1(4.5)	3.45	2.55
2	6	6	0.1(4.5)	6.45	5.55
3	8	8	0.1(11.3)	0.113	6.87
4	11	11	0.1(11.3)	0.113	9.87
5	15	15	0.1(11.3)	0.113	13.87

$\lambda \rightarrow$ is how slowly do I want to learn

very slow $\rightarrow$ overfitting

kamperh.com

1. Initialise $f(x) = 0$

$r = y$

2. for iteration 1 to B

{

a) Fit a model. $f_b(x; \theta) \to r$

b) update the final model, by adding the shrunken version

$$F(x;\theta) = f(x,\theta) + \lambda f_b(x;\theta)$$

c) update the residuals

$$r^{(1)} = r^{(1)} - \lambda f_b(x_i;\theta)_{i}$$

}

3. Final model as

$$F(x;\theta) = \sum_{i=1}^{B} \lambda f_b(x;\theta)$$

Gradient Boosting

Gradient boosted machines

How to apply this for classification setting?

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

$F(x) = 0$

Iteration 1

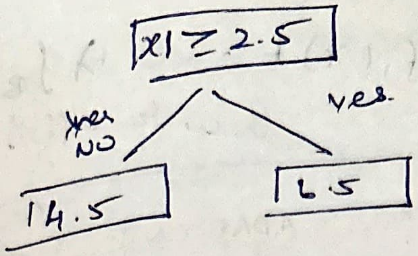
$r = y$

dataset :-

x_1	x_2	r
1	2	4
2	3	5
3	4	6

a) Fit a model $f_b(x, \theta) \rightarrow r$

$f_1(x)$



b) update the model $\cdot\, \lambda = 0.1$

$F(x) = 0 + \lambda f_1(x)$

c) update the residuals

$r(1) = r^{(1)} - \lambda f_1(x^{(1)}) = 4 - [0.1 \times 4.5] = 3.55$

$r(2) = r^{(2)} - \lambda f_1(x^{(2)}) = 5 - [0.1 \times 4.5] = 4.55$

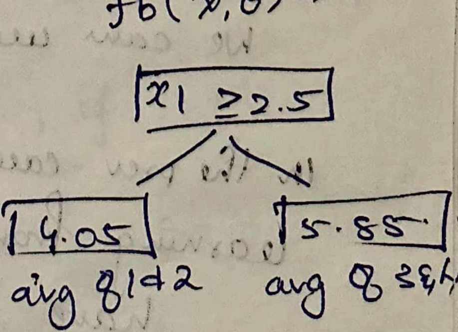
$r(3) = r^{(3)} - \lambda f_1(x^{(3)}) = 6 - [0.1 \times 6.5] = 5.35$

$r(4) = r^{(4)} - \lambda f_1(x^{(4)}) = 7 - [0.1 \times 6.5] = 6.35$

Iteration 2

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

a) Fit a model. $f_b(x, \theta) \rightarrow r$



b)

$F(x) = \lambda f_1(x) + \lambda f_2(x)$

c)

$r(1) = r^{(1)} - \lambda f_2(x^{(1)}) = 3.145 = 3.55 - [0.1 \times 4.05]$

$r(2) = 4.145$, $r(3) = 4.765$, $r(4) = 5.765$

— after B iterations —

Final

$$F(x, \theta) = f_0(x) + \lambda f_1(x) + \dots + \lambda f_B(x)$$

how to apply this for

regression

setti

stump ①

example :-

	x_1	x_2	y	w	$\hat{y}_1$
1	1	2	+1	1	+1
2	2	3	+1	1	+1
3	3	3.	+1	1	~~+1~~ -1
4	4	5	-1	1	-1
5	5	5	-1	1	-1
6	6	6	-1	1	-1

No of weak learners = 5

1) iteration 1 :-

→ construct a tree $f_1(x_1, 3.5)$ $\boxed{x_1 \geq 2.5}$

No → $\boxed{+1}$; yes → $\boxed{-1}$

$\rightarrow \varepsilon = \frac{1 \times 1}{6}$

$= 0.1667$

$\rightarrow \lambda_1 = \frac{1}{2} \log \left(\frac{1 - 0.1667}{0.16667} \right) = 0.8045$

$\rightarrow w^3 = 1 \times e^{-0.8045}$ / $\rightarrow w^{\{1,2,4,5,6\}} = 1 \times e^{+0.8045}$

$\rightarrow \omega^3 = 0.447$

$\rightarrow \omega^{\{1,2,4,5,6\}} = 2.235$

x_1	x_2	y	ω	$\hat{y}_1$	ω
1	2	+1	1	+1	2.235
	3	+1	1	+1	2.235
2	3	+1	1	-1	0.447
3	3	+1	1	-1	2.235
4	5	-1	1	-1	2.235
5	5	-1	1	-1	2.235
6	6	-1	1	-1	2.235

disadv

→ noisy datase
- overfitti

→ interpretabi

↓

focus on Stump with nose

→ cant handle high dimen
- seconal data

→ end of interna[l]s ←

sum based on this

Sample	x_1	x_2	y_1	r
1	2	3	1	1
2	1	2	0	0
3	3	4	1	1
4	4	5	1	1
5	1	1	0	0
6	2	2	0	0

$y_1 = 6$

Step 1:-

initialisation

we assume $r = y$

(i.e) $\cdot 1 \cdot y$ -

$F_0(x) =$ log. oc

$p = 3/6$

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

$\frac{1}{1+e^{-F}}$ prob that it belongs to class 1

→ compute F_0 for first stump

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

get p values

$p = \frac{1}{1+e^{-F_0}}$

x_1	x_2	y
32	23	0
12	15	1

→ based first tree (doesn't exist yet

→ compute $y - p$

→ fit model to $y - p = y'$

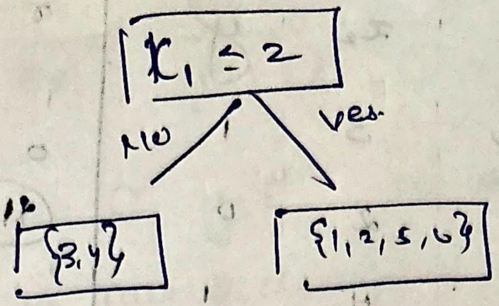
→ I left → w left
I righ → w right
labels

Step 2:-

$r_i = y_i - p_i$

Sample	True y	Initial 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



$$w = \frac{\Sigma(y-p)}{\Sigma P(1-P) + \lambda}$$ → residual sum

second deriv of loss function gives this

how slowly do I want to move?

↓

$$w = \frac{\text{sum of residuals}}{\text{sum}[\text{prob}(1-\text{prob})]}$$

prob values are not linearly additive so we can't take avg values

↓

so we can convert this to log-odds space → how?

Note:
Max lifting procedure is used in final step of building one tree from all stumps

$$w_{right} = \frac{0.5 - 0.5 - 0.5 - 0.5}{0.5(0.5) \times 4}$$

4 → 4 samples

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

$$= -1$$

$$\boxed{w_{right} = -1}$$

$$w_{left} = \frac{0.5 + 0.5}{(0.5 \times 0.5) \times 2}$$

$$= \frac{1}{0.5}$$

$$= \frac{\cancel{10}^{\,2}}{\cancel{5}}$$

$$\boxed{w_{left} = 2}$$

prob values from log odds

$$P_{right} = \cancel{\frac{1}{1+e^{x}}}$$

$$\cancel{1 \cdot \frac{1}{1+e^{n}}}$$

Kernel

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

↓

Kernel fn

$x = (x_1, x_2, x_3)$

$y = (y_1, y_2, y_3)$

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

Example:-

$$x = \begin{array}{ccc} x_1 & x_2 & x_3 \\ \end{array} \begin{bmatrix} 1, & 2, & 3 \end{bmatrix}$$

$$y = \begin{array}{ccc} y_1 & y_2 & y_3 \\ \end{array} \begin{bmatrix} 4, & 5, & 6 \end{bmatrix}$$

$$\phi(x) = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 2 \\ 4 \\ 6 \\ 3 \\ 6 \\ 9 \end{bmatrix} \quad \phi(y) = \begin{bmatrix} 16 \\ 20 \\ 24 \\ 20 \\ 25 \\ 30 \\ 24 \\ 30 \\ 36 \end{bmatrix} \quad \phi(x)^T \phi(y) = \begin{bmatrix} 16 + 40 + 72 \\ + 40 + 100 + 180 \\ + 72 + 180 + 324 \end{bmatrix}$$

$\phi(x)^T \phi(y) = 1024$

$\boxed{\langle \phi(x), \phi(y) \rangle = 1024}$

dot product of transformed vectors:- in a ∞ dimension space is computed

So, kernel solves this problem?

Can we perform some operations in original lower dim space that is comparable to doing the same in high dim space

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

$x = [1, 2, 3]$

$y = [4, 5, 6]$

$$= \left[[1, 2, 3] \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} \right]^2$$

$$= (4 + 10 + 18)^2$$

$$= [32]^2$$

$$= 1024$$

We are able to mimick what happened in high dim. in lower dim space with help of kernel function | So we can potentially do the the functionalities in ∞ dim. space in 1D space (lower dim space)

→ absolute ?
→ approximate ?
] depends on kernel func

Techniques / kernel methods

Support Vector Machines

11/03/24

Recap:- prev class

Kernel:-

Walkthrough example

x_1	x_2	y
1	2	0
3	4	0

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

$X = [x_1 \quad x_2]$ $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

$x \in \mathbb{R}^2 \quad \theta \in \mathbb{R}^2$

$\phi(x) \in \mathbb{R}^3 \quad | \quad \theta \in \mathbb{R}^3$

$x_1^2 \quad x_2^2 \quad x_1 x_2 \quad y_1^2 \quad y_2^2 \quad y_1 y_2$

X wrong

$$\theta_1 = \sum_{i=1}^{2} \beta_i \phi(x_1^{(i)})$$

$$\theta_i = \beta_1 \phi(x_1^1) + \beta_2 \phi(x_1^2)$$

$$\theta_1 = \cancel{\beta_1 [x^{(1)2} \quad x^{(2)2} \quad x^{(1)}x^{(2)}] + \beta_2}$$

$$= \beta_1 [x_1^2 \quad x_2^2 \quad x_1 x_2]^T + \beta_2 [x_1^2 \quad x_2^2 \quad x_1 x_2]^T$$

$$= \beta_1 [1 \quad 4 \quad 2]^T + \beta_2 [9 \quad 16 \quad 12]^T$$

$$= \beta_1 + 4\beta_1 + 2\beta_1 + 9\beta_2 + 16\beta_2 + 12\beta_2$$

$$\theta_1 = 7\beta_1 + 37\beta_2$$

correct

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

$$\theta = \beta_1 \phi(x^{(1)}) + \beta_2 \phi(x^{(2)})$$

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

n samples

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

$d \times p$

note:- β obtained by gradient descent

in the testing phase

$[\ ]_{d \times p}$

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

$[\ ]_{n \times d}$

$[\ ]_{n \times d}$

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

$[\ ]_{n \times d}$

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

testing $\to$ training set

$\to$ training set

for each testing set, we have to calculate for all samples

Computationally Expensive

uses subset of data instead of whole data

How to find : distances from point to plane?

How to calc Margin?

→ ⊥ distance

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

also shrinking each plane weights by the coeff [Normalization factor]

Three ones [hyperplane]

$2x_1 + 3x_2 - 5 = 0$

$-x_1 + 4x_2 + 7 = 0$

$5x_1 - 12x_2 + 10 = 0$

Dataset

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

find the optimal hyperplane

1st hyperplane $\quad 2x_1 + 3x_2 - 5 = 0.$

* coef are βs
* intercept is β_0

1st sample

$$d_1 = \frac{|2x_1 + 3x_2 - 5|}{\sqrt{2^2 + 3^2}} = \frac{|2\times3 + 3\times4 - 5|}{\sqrt{4+9}}$$

$$= \frac{|6+12-5|}{\sqrt{13}}$$

$$= \frac{13}{\sqrt{13}}.$$

$$d_2 = \frac{|2\times2 + 3\times3 - 5|}{\sqrt{13}}$$

$$= \frac{|4+9-5|}{\sqrt{13}} = \frac{8}{\sqrt{13}}.$$

$$d_3 = \frac{|2\times1 + 3\times-1 - 5|}{\sqrt{13}}$$

$$= \frac{6}{\sqrt{13}}.$$

$$d_4 = \frac{|2\times-2 + 3\times1 - 5|}{\sqrt{13}} = \frac{6}{\sqrt{13}}$$

min distance

-8
2

-9+3

2nd hyperplane

$-x_1 + 4x_2 + 7.$

$$d = \frac{-x_1+4x_2+7}{\sqrt{1+16}} = \frac{-x_1+4x_2+7}{\sqrt{17}}$$

3 4
2 3
1 1 -1
-2 1

$$d_1 = \frac{|3+16+7|}{\sqrt{17}} = \frac{20}{\sqrt{17}}$$

$$d_2 = \frac{|-2+12+7|}{\sqrt{17}} = \frac{17}{\sqrt{17}}$$

$$d_3 = \frac{|-1-4+7|}{\sqrt{17}} = \frac{2}{\sqrt{17}} \text{ min}$$

$$d_4 = \frac{|2+4+7|}{\sqrt{17}} = \frac{13}{\sqrt{17}}$$

3rd hyperplane

$5x_1 - 12x_2 + 10 = 0$

$$d = \frac{|5x_1-12x_2+10|}{\sqrt{25+144}} = \frac{|5x_1-12x_2+10|}{13}$$

$$d_1 = \frac{|5\times3-12\times4+10|}{13} = \frac{23}{13}$$

$$d_2 = \frac{|5\times2-12\times3+10|}{13} = \frac{16}{13}$$

$$d_3 = \frac{|5\times1-12\times-1+10|}{13} = \frac{27}{13}$$

$$d_4 = \frac{|5\times(-2)-12\times1+10|}{13} = \frac{12}{13} \text{ min}$$

-3 hyperplanes-

$$\max\left[\frac{6}{\sqrt{13}}, \frac{2}{\sqrt{17}}, \frac{12}{13}\right]$$

$$\max\left[1.66, 0.485, 0.923\right]$$

choose the 1st hyperplane

best hyperplane :- $2x_1 + 3x_2 - 5$

Reverse:-

① Projection

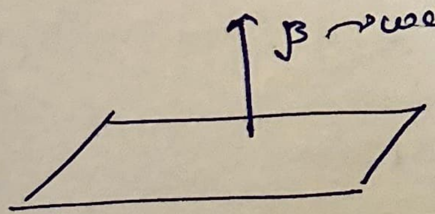
② Norm of a vector

③ vector Addition

④ Orthogonality.

⑤ Subspaces.

$\beta \rightarrow$ coeff is $\perp$ to the plane



$\beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots$

proof ?

we are trying to find margin in the direction of coeff vector [it's a part of direction vector] so that why we find unit vector $\perp$ to hyperplane.

21/03/25

Evaluation metrics - Sample Example

threshold - 0.5

Observation	Actual label	Pred. Label		Pred Score
1	1	1		0.85
2	0	1	FP	0.60
3	1	1		0.70
4	1	0	FN	0.40
5	0	1	FP	0.55
6	1	1		0.50
7	0	1	FP	0.65
8	0	0		0.35
9	1	1		0.60
10	0	0		0.20

Confusion Matrix

	original observed +ve	original observed -ve	
Predicted +ve	1, 3, 6, 9 (4) 0.85, 0.70, 0.50, 0.60	2, 5, 7 (3) 0.60, 0.55, 0.65	7
Predicted -ve	4 (1) 0.40	8, 10 (2) 0.35, 0.20	3
	5	5	

Accuracy :-

$\frac{TP}{TN}$ $\frac{TP+TN}{Total} = \frac{4+2}{10} = 0.6$

Precision :- (+ve pred value)

$\frac{TP}{TP+FP} = \frac{4}{7} = 0.57$

Positive pred value :- $\cancel{\frac{5}{7}} = \cancel{\frac{5}{7}}$

Recall :- Sensitivity, TPR

$\frac{4}{5} = 0.8$

FPR.

Specificity

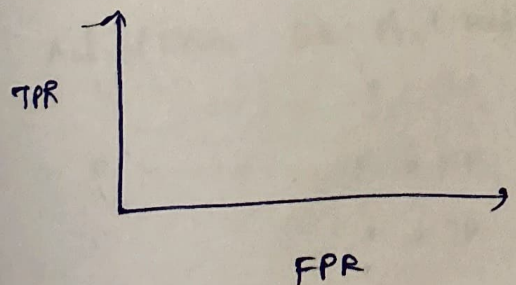
$\frac{2}{5} = 0.4$

FPR $= 0.6$. [1 - specificity = FPR]

$F_1 = 2 \cdot \frac{0.57 \times 0.8}{0.57 + 0.8} = 0.66.$

Threshold	FPR	TPR
0.5	0.6	0.8

ROC?



Plot for diff thresholds

Thresholds = $\begin{bmatrix} 0.7 \\ 0.5 \\ 0.4 \\ 0.2 \end{bmatrix}$

Obs	Actual Data	Pred Data		Pred Score
1	1	1	TP	0.85
2	0	0	TN	0.60
3	1	1	TP	0.70
4	1	0	FN	0.40
5	0	0	TN	0.55
6	1	0	FN	0.50
7	0	0	TN	0.65
8	0	0	TN	0.35
9	1	0	FN	0.60
10	0	0	TN	0.20

Threshold 0.7

Confusion matrix

	Org +ve	Org -ve
Pred +ve	TP 2	FP 0
Pred -ve	FN 3	TN 5
	5	5

Accuracy

$$\frac{5+2}{10} = 0.7$$

Precision :- 1

Sensitivity Recall :- TPR

$$\frac{2}{5} = 0.4$$

Specificity

0.1

FPR :- 0

Threshold	TPR
0.7	0.4

FPR

0.

Threshold : 0.4

obs	Actual Data	Pred Data	Pred Score
1	1	1 * TP	0.85
2	0	1 * FP	0.60
3	1	1 * TP	0.70
4	1	1 * TP	0.40
5	0	1 * FP	0.55
6	1	1 * TP	0.50
7	0	1 * FP	0.65
8	0	0 * TN	0.35
9	1	1 * TP	0.60
10	0	0 * TN	0.20

Confusion mat

	org +ve	org -ve	
pred +ve	5	3	8
pred -ve	0	2	2
	$\frac{1}{5}$	$\frac{1}{5}$	

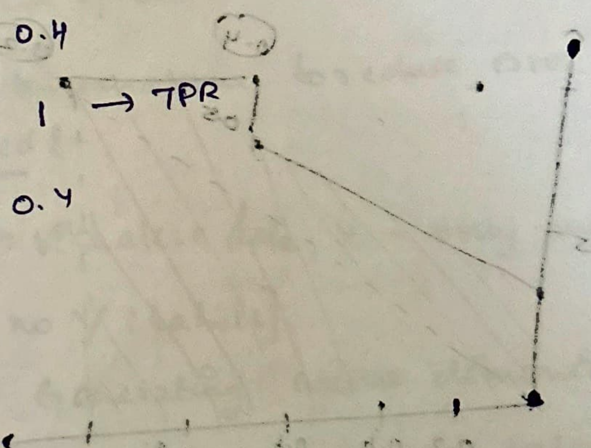
Accuracy : $\frac{7}{10}$

Precision : $5/8 = 0.4$

Sensitivity : $5/5 = 1 \rightarrow$ TPR

Specificity : $\frac{2}{5} = 0.4$

FPR $\rightarrow$ 0.6



Threshold = 0.2

obs	Actual Data	Pred Data	Pred Score
1	1	1 * TP	0.85
2	0	1 * FP	0.60
3	1	1 * TP	0.70
4	1	1 * TP	0.40
5	0	1 * FP	0.55
6	1	1 * TP	0.50
7	0	1 * FP	0.65
8	0	1 * FP	0.35
9	1	1 * TP	0.60
10	0	1 * FP	0.20

	org +ve	org -ve	
Pred +ve	TP 5	FP 5	10
Pred -ve	FN 0	TN 0	0
	5	5	10

Accuracy = 5/10 = 0.5

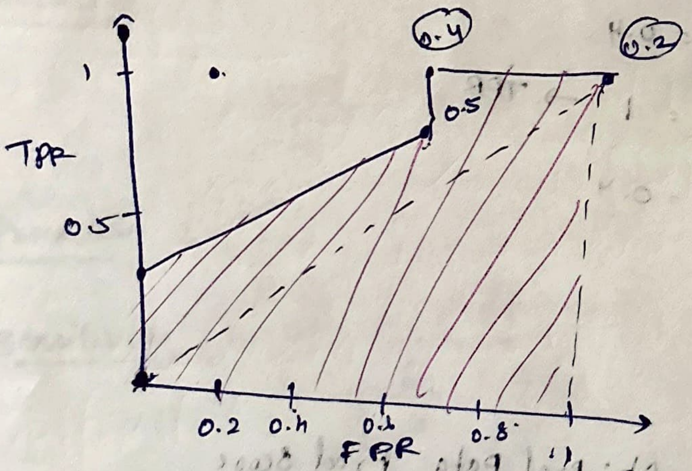
Precision = 5/10 = 0.5

Sensitivity = 5/5 = 1 ⇒ TPR

Specificity = 0/5 = 0 ⇒ F

FPR = 1

Threshold	TPR	FPR
0.7	0.4	0
0.5	0.8	0.6
0.4	1	0.6
0.2	1	1



ROC curve

Qualitative ?

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

$$= \sum_{i=1}^{4} \frac{(0.6-0) * (0.8+0.4)}{2} + \frac{(0.6-0.6) * (0.8+1)}{2}$$

$$+ \frac{(1.0-0.6) * 2}{2}$$

interpretation :
Model performance better than Random guess

Quantitative estimation : AUC curve
how good my model is

Area should be greater than 0.5

$$= \frac{0.6 \times 1.2}{2} + 0 + \frac{0.4 \times 2}{2}$$

$$= \boxed{0.76 = AUC}$$

Okayish Model

Example

$$X = \begin{bmatrix} 2 & 100 \\ 4 & 200 \\ 6 & 300 \end{bmatrix}_{3\times 2} \xrightarrow{\text{apply PCA}} Z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}_{3\times 1}$$

$$\mathbb{R}^2 \longrightarrow \mathbb{R}^1$$

① **Standardisation** of each column.

column 1 mean $\frac{2+4+6}{3} = \frac{12}{3} = 4$

2 mean $\frac{100+200+300}{3} = 200.$

$$SD = \sqrt{\frac{(2-4)^2 + (4-4)^2 + (6-4)^2}{2}}$$

$$SD_1 = \sqrt{\frac{4+4}{2}} = 2.$$

$$SD_2 = \sqrt{\frac{(100-200)^2 + (200-200)^2 + (300-200)^2}{2}}$$

$$= 100.$$

$$X_{scaled} = \begin{bmatrix} \frac{2-4}{2} & \frac{100-200}{100} \\ \frac{4-4}{2} & 0 \\ \frac{6-4}{2} & \frac{300-200}{100} \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}$$

② compute covariance matrix $X_{std}^T X_{std}$

$$\begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}_{2\times3} \begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}_{3\times2}$$

$$\frac{1}{n-1} \begin{bmatrix} +1+0+1 & +1+0+1 \\ +1+0+1 & +1+0+1 \end{bmatrix}_{2\times2} = \frac{\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}_{2\times2}}{2}$$

$$= \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}_{2\times2}$$

③ compute Eigen values & vectors of the covariance matrix

$AX = \lambda X$

* find eigen value.

$$|A - \lambda I| = 0$$

$$\left| \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = 0$$

$$\begin{vmatrix} 1-\lambda & 1 \\ 1 & 1-\lambda \end{vmatrix} = 0$$

$$(1-\lambda)^2 - 1 = 0$$

$$1 + \lambda^2 - 2\lambda - 1 = 0$$

$$\lambda^2 - 2\lambda = 0$$

$$\lambda(\lambda - 2) = 0$$

$$\boxed{\lambda = 2} \quad \boxed{\lambda = 0}$$

$\underline{\lambda = 2}$

$(A - \lambda I)v = 0$

$$\begin{bmatrix} 1-2 & 1 \\ 1 & 1-2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0$$

$$\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0.$$

$-x + y = 0$

$x - y = 0.$

$\boxed{x = y}$

$v_1 = t\begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$\sqrt{t^2+t^2} = 1$

$\sqrt{2t^2} = 1$

$\boxed{t = \frac{1}{\sqrt{2}}}$

$\boxed{t = 1/\sqrt{2}}$

$\underline{\lambda = 0}$

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0$$

$x + y = 0$

$x + y = 0$

$x = -y$

$x = -y$

$v_2 = t\begin{bmatrix} 1 \\ -1 \end{bmatrix}$

(4) choose k (Decide the needed % of variance explained & choose k principal component)

(5) Transform X into new space of principal components.

(6) use this Z to train a model

(4) $\left(\lambda_1 = 2 ;\ v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} ;\ \lambda_2 = 0, v_2 = \frac{-1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)$

Z_{pca} · choose (1) → 95% variance explained

$k = 1$

(5) $Z_{pca} = X_{std}^{T} \times v_1$

$$= \begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

$$\begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}_{2\times 3} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

$\lambda = 2$

$A = [\ \ ]$

$\lambda I = [\ \ ]$

retI $\begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}_{3\times 2} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}_{2\times 1}$

$$Z_{pca} = \begin{bmatrix} -1/\sqrt{2} - 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} + 1/\sqrt{2} \end{bmatrix} \cdot \begin{bmatrix} -2/\sqrt{2} \\ 0 \\ 2/\sqrt{2} \end{bmatrix} \begin{matrix} z_1 \\ z_2 \\ z_3 \end{matrix}$$

Example :-

# Sample	x_1	x_2
1	1	1
2	1	2
3	2	2
4	8	8
5	8	9
6	9	8

i) $C_1 = \{2, 3, 4, 6\}$

$C_2 = \{1, 5\}$.

$\text{centroid}_1 = \left[\frac{1+2+8+9}{4} \quad \frac{2+2+8+8}{4} \right]$

$\text{centroid}_1 = [5 \quad 5]$

$\text{centroid}_2 = \left[\frac{1+8}{2} \quad \frac{1+9}{2} \right]$

$\text{centroid}_2 = [4.5 \quad 5]$

$\theta = 1$

Iteration 1

for n = 1

S1 [1 , 1]

C1 [5 5]

$dist_{11} = \sqrt{(5-1)^2 + (5-1)^2}$

$= \sqrt{16 + 16}$

$= \sqrt{32}$

S1 [1 1]

C2 [4.5 5]

$dist_{12} = \sqrt{(4.5-1)^2 + (5-1)^2}$

$dist_{12} = \sqrt{28.25}$

put sample 1 in 2nd cluster

for n = 2

S2 [1 2]

C1 [5 5]

$dist_{21} = \sqrt{16 + 9}$

5

S2 [1 2]

C2 [4.5 5]

$dist_{22} = \sqrt{3.5^2 + 3^2}$

$= \sqrt{21.25} = 4.60$

dist 22 is less

put sample 2 in cluster 2

for n = 3

S3 [2 2]

C1 [5 5]

$dist_{31} = \sqrt{18}$

$= 4.24$

S3 [2 2]

C2 [4.5 5]

$dist_{32} = \sqrt{15.25}$

$= 3.9$

put sample 3 in cluster 2

for $n = 4$

$S4 = [8 \quad 8]$

$C1 = [5 \quad 5]$

$dist_{41} = \sqrt{9 + 9}$

$= \sqrt{18}$

put sample 4 in cluster 1

$S4 = [8 \quad 8]$

$C_2 \; [4.5 \quad 5]$

$dist_{42} = \sqrt{21.25}$

for $n = 5$

$S5 = [8 \quad 9]$

$C1 = [5 \quad 5]$

$dist_{51} = \sqrt{9 + 16}$

$= \sqrt{25}$

put sample 5 in cluster 1

$S5 = [8 \quad 9]$

$C2 = [\cancel{4}.5 \quad 5]$

$dist_{52} = \sqrt{(8-4.5)^2 + 16}$

$= \sqrt{28.25}$

for $n = 6$

$S6 = [9 \quad 8]$

$C1 = [5 \quad 5]$

$dist_{61} = \sqrt{25}$

put sample 6 in cluster 2

$S6 = [9 \quad 8]$

$C2 = [4.5 \quad 5]$

$dist_{62} = \sqrt{(4-4.5)} + 9$

$= \sqrt{29.25}$

Cluster 1	Cluster 2
1	4
2	5
3	6

Iteration 2:- Same.

convergence criteria?

Example:-

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

Single linkage

Euclidean distance as measure

A B C D E

(i) ${}^{n}C_2$ pairwise distances $= {}^{5}C_2 = \frac{5 \times 4}{2} = 10$

$$\boxed{d(A\,B) = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} = \sqrt{(2-1)^2 + (1-1)^2} = 1}$$

min

(8+)

① $d(AC) = \begin{bmatrix} 1 & 1 \\ 4 & 3 \end{bmatrix} = \sqrt{(4-1)^2 + (3-1)^2}$

$= \sqrt{9+4}$

$= \sqrt{13} = 3.60$

② $d(AD) = \begin{bmatrix} 1 & 1 \\ 5 & 4 \end{bmatrix}$

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

$= \sqrt{16+9}$

$= 5$

③ $d(AE) = \begin{bmatrix} 1 & 1 \\ 6, & 5 \end{bmatrix}$

$= \sqrt{(6-1)^2 + (5-1)^2}$

$= \sqrt{25+16}$

$= \sqrt{41} = 6.4$

→ A, B has min distance

	A	B	C	D	E
A	0	1	3.6	5	6.4
B	1	0	2.82	4.2	5.6
C	3.6	2.82	0	1.4	2.82
D	5	4.2	1.4	0	1.42
E	6.4	5.6	2.82	1.42	0

⑩ $d(D, E)$

$= \sqrt{1+1}$

$= \sqrt{2}$

④ $d(B.C) = \begin{bmatrix} 2 & 1 \\ 4 & 3 \end{bmatrix}$

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

$= \sqrt{4+4}$

$= \sqrt{8}$

≈ 2.82

⑤ $d(B, D) = \begin{bmatrix} 2, & 1 \\ 5 & 4 \end{bmatrix}$

$d(B.D) = 4.2$

$d(B, E) = \begin{bmatrix} 2, & 1 \\ 6, & 5 \end{bmatrix}$

⑥ $d(B, E) = 5.6$

⑧ $d(C, E)$

$= \sqrt{4+4}$

$= \sqrt{8}$

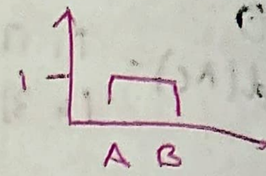
$= 2.82$

⑨ $d(C, D) =$

$= \sqrt{1+1}$

$= \sqrt{2}$

$\boxed{\text{Merge (A, B)}}$ distance value 1



	AB	C	D	E
AB	0	2.8	4.2	5.6
C	2.8	0	1.4	2.8
D	4.2	1.4	0	1.4
E	5.6	2.8	1.4	0

min (CA, CB)

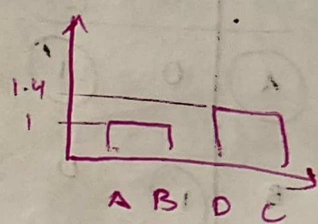
min (3.6, 2.8)

→ choose min of this

Merge (C, D) distance value 1.4

~~D(AB, C)~~

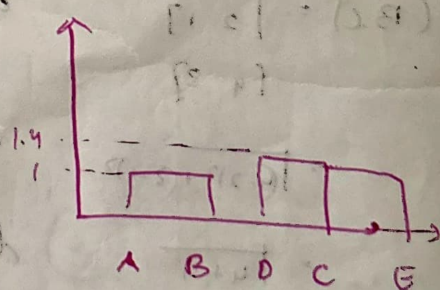
~~D(AB, D)~~



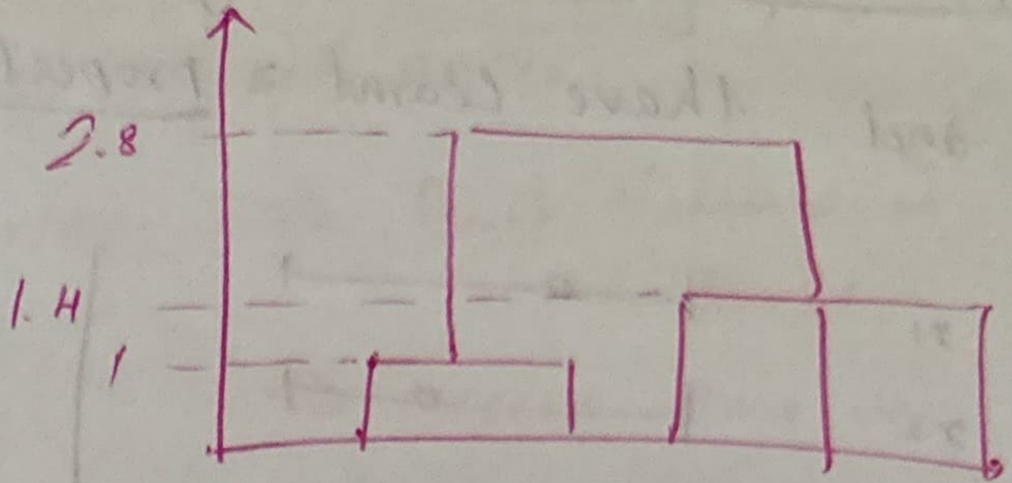
~~D(AB, E)~~

E(C, D)

	AB	CD	E
AB	0	~~2.8~~ 2.8	4.2
CD	2.8	0	1.4
E	4.2	1.4	0



	AB	CDE
AB	0	2.8
CDE	2.8	0



Dictionary Text modality

$$\begin{bmatrix} a \\ able \\ allao \\ \vdots \\ more \\ \vdots \\ Play \\ product \\ \vdots \end{bmatrix}$$ 50000 words

→ im going to map index of the words in the dictionary

S1 :- "To buy this product"

indx 2000	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$
indx 8000	"buy index"
idx 70000	"product"
idx 10000	"this"
	50000×1
	"to"

	x_1	x_2	x_3	class
S1	1	0	1	Spam
S2	1	1	0	spam
S3	1	1	0	Spam.
S4	0	0	1	No spam
S5	0	0	0	No spam.
S6	1	1	1	No spam
S7	0	1	0	spam
S8	1	0	0	spam
S9	0	1	0	spam
S10	1	0	0	spam
S11	1	1	0	no spam

notes:-

features → value & indices

each sentence → sample

we can assign prob. values for each row/sample

for 1 sample → uniquely identified by 3 features

x_1, x_2, x_3

$2 \times 2 \times 2$ → total of 8 possibilities map the prob values

How many times 000 led to spam in data

	x_1	x_2	x_3	prob [spam] $y=1$	prob [no spam] $y=0$
1	0	0	0	0	1/4
2	0	0	1	0	1/4
				2/7	0
3	0	1	0	0	0
4	0	1	1	2/7	0
5	1	0	0	1/7	0
6	1	0	1	2/7	1/4
7	1	1	0	0	1/4
8	1	1			

How many times 001 led to spam

discrete

|Spam| = 7 |no spam| = 4

→ Here it follows Bernoulli Distbn

we just have to learn $\boxed{\theta \cdot (p)}$ parameter.

for Gaussian it should be $\boxed{\mu, \sigma}$

Multinomial distribution:-

a vector of θ's

$\theta = [\theta_1, \ldots \ldots \theta_m]$

discrete value in range m

We should look at data & figure out the distbn

in inference stage:-

New Data [Email] with these three words.

lets say, "Good job" → [110]

max (2/7, 1/4) → 2/7

Example :-

X	Y
1. Free win now	Spam
2. Win a prize	Spam
3. Hello How are you	Not Spam
4. Let's win it	Not spam
5. Free lunch today	Not Spam

NB → Suited for Spam classifiers

← Training

Test
[Free win]

→ we have to Build NB classifier

we are only 2 features.

x_1 word free.	x_2 win	Label
1 Yes	yes 1	Spam
0 No	yes 1	Spam
0 No	No 0	Not Spam
0 No	Yes 1	Not Spam
1 Yes	No 0	Not Spam

Translate prior probs

$$\theta_{y=1} = \frac{\sum_{i=1}^{n} I(y^{(i)}=1)}{n}$$

$\boxed{\theta_{y=1} = 2/5}$ P(spam) = 2/5

$\boxed{\theta_{y=0} = 3/5}$ P(not spam) = 3/5

→ P(Free=yes | spam=1)

→ P(win=yes | spam=1)

$\boxed{\theta_{x_1|y=1} = \frac{1}{2}}$ $\boxed{\theta_{x_2|y=1} = 1}$

$\boxed{\theta_{x_1|y=0} = 1/3}$ $\boxed{\theta_{x_2|y=0} = 1/3}$

→ P(win=yes | spam=0)

→ P(~~win~~ free=yes | spam=0)

$$P(y=1|x_1) = \frac{P(x_1|y=1)\,P(y=1)}{P(x_1|y=1)P(y=1) + P(x_1|y=0)P(y=0)}$$

P(spam | x)

$$= \frac{1/2 \times 1}{1/2 \times 1/2 + 1/3 \times 3/5} = \frac{1/2}{1/4 + 1/5} = \frac{1/2}{9/20}$$

$$= \frac{1}{2} \times \frac{20}{9}$$

$$= 10/9$$

$P(x) = a \mid x$ Inference stage

$P(spam \mid x) = P(x \mid spam)\, P(spam)$ $\longrightarrow$ Denominator will ignored

$P(spam \mid x) = P(x_1, x_2 \mid spam)\, P(spam)$

$= P(x_1 \mid spam)\, P(spam)\, P(x_2 \mid spam)$ ...

$P(spam \mid x) = \frac{1}{2} \times \frac{2}{5} \times 1$

$\frac{2}{10} = 0.2$

$P(not\ spam \mid x) = P(x \mid not\ spam)\, P(non\ spam)$

$= P(x_1, x_2 \mid not\ spam)\, P(non\ spam)$

$= P(x_1 \mid not\ spam)\, P(x_2 \mid not\ spam)\, P(non\ spam)$

$= 1/3 \times 1/3 \times 3/5$

$\boxed{= \frac{1}{15} = 0.0667}$

Where are we using Test data?