$M_1 = 1.664 \quad M_2 = 0.485$

$M_3 = 0.2, \ 0.92.$

__/__/____

Three lines (hyperplane)

eq (1) $2x_1 + 3x_2 - 5 = 0$

β_1 β_2 β_0

Dataset

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

for $x_1 = 3, \ x_2 = 4$

$$d_1 = \frac{-5 + 2x_1 + 3x_2}{\sqrt{2^2 + 3^2}}$$

$$= \frac{-5 + 6 + 12}{\sqrt{13}} = \frac{13}{3.6} = 3.6$$

for $x_1 = 2, \ x_2 = 3$

$$d_2 = \frac{2(2) + 3(3) - 5}{3.6}$$

$$= \frac{4 + 9 - 5}{3.6} = \frac{8}{3.6} = 2.2$$

for $x_1 = 1, \ x_2 = -1$

$$d_3 = \frac{-2 + 3 - 5}{3.6} = \frac{10}{3.6} = 2.7$$

for $x_1 = -2, \ x_2 = +1$

$$d_4 = \frac{2(-2) + 3(1) - 5}{3.6} = \frac{-4 + 3 - 5}{3.6} = \frac{-6}{3.6} = 3.6$$

$\Rightarrow 1.6$

__/__/____

$M_1 = 1.66$

eq (2). $-x_1 + 4x_2 + 7 = 0$

for $x_1 = 3, x_2 = 4$

$$\frac{-3 + 4(4) + 7}{\sqrt{-1 + (4)^2}}$$

$$= \frac{-3 + 16 + 7}{\sqrt{-1 + 16}} = \frac{20}{3.8} = 5.2$$

for $x_1 = 2, x_2 = 3$

$$\frac{-2 + 4(3) + 7}{3.8}$$

$$= \frac{-2 + 12 + 7}{3.8} = \frac{17}{3.8} = 4.47$$

for $x_1 = 1, x_2 = -1$

$$\frac{-1 + 4(-1) + 7}{3.8}$$

$$= \frac{-4 + -4 + 7}{3.8}$$

$$= \frac{-1}{3.8} = 0.2.$$

for $x_1 = -2, x_2 = 1$

$$\frac{-(-2) + 4(1) + 7}{3.8}$$

$$= \frac{2 + 4 + 7}{3.8} \Rightarrow \frac{13}{3.8} = 3.42$$

__/__/____

Evaluation metrics

actual values
not predicted ones.

red circle → positive class
white circle → negative class

threshold = 0.5

Confusion Matrix

	Original +ve	org. −ve
pred +ve	9 (TP)	2 (FP)
pred −ve	1 (FN)	8 (TN)

T = True
P = +ve
N = −ve
F = False.

$$\text{Accuracy} = (TP + TN)/\text{Total}$$
$$= (9+8)/20$$
$$= 0.85.$$

$$F_1 \text{ score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{1}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$$

$$\text{precision} = \frac{TP}{TP + FP} \qquad \text{recall} = \frac{TP}{TP + TN}$$

$$\frac{2}{}$$

__/__/____

observation	Actual score	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

Compute the following :
1) Accuracy
2) Precision
3) Recall or sensitivity
4) Specificity
5) False Positivity
6) F1-score.
7) ROC plot
8) AUC.

threshold = 0.5

~~Accuracy = (TP + TN) / Total~~

=

Confusion Matrix

	orig +ve	orig -ve
pred +ve	✓ 4	3
pred -ve	1	✓ 2

Accuracy = (4+2) / 10

= 6/10

= 0.6

2) Precision $= \frac{TP}{TP + FP} = \frac{4}{4+3} = \frac{4}{7} = 0.57$

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

4) Specificity = Negative Recall $= \frac{TN}{TN+FP} = \frac{2}{2+3}$

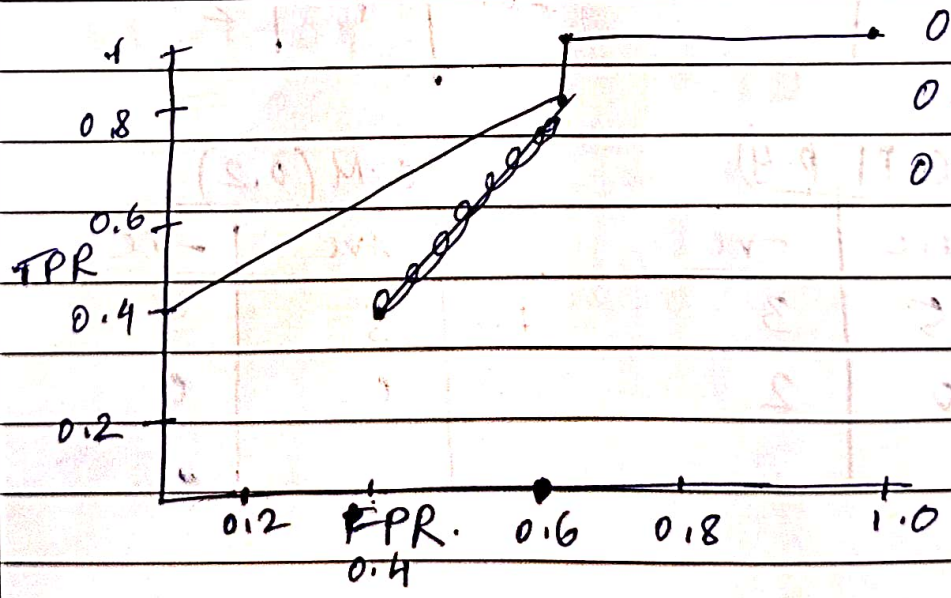
$= \frac{2}{5} = 0.4$

5) False Positivity =

6) F_1 score $= 2 \times \frac{p \times r}{p+1} \Rightarrow 2 \times \frac{0.57 \times 0.66}{0.5 + 0.66}$

$\Leftarrow \Rightarrow \frac{2(0.3762)}{1.16}$ $\Leftarrow$ 0.648.

7) Roc curve → Threshold values → 0.7, 0.5, 0.4, 0.2



__/__/____

Actual score	Predicted score	0.7	0.5	0.4	0.2
1	0.85	1	1	1	1
0	0.60	0	1	1	1
1	0.70	1	1	1	1
1	0.40	0	0	1	1
0	0.55	0	1	1	1
1	0.50	0	1	1	1
0	0.65	0	1	1	1
0	0.35	0	0	0	1
1	0.60	0	1	1	1
0	0.20	0	0	0	1

C.M (0.7)

org / pred	+ve	-ve
+ve	2	0
-ve	3	5

C.M (0.5)

	+ve	-ve
+ve	4	3
-ve	1	2

C.M (0.4)

	+ve	-ve
+ve	5	3
-ve	0	2

C.M (0.2)

	+ve	-ve
+ve	5	5
-ve	0	0

0.2

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

TPR $= 5/(5+0) = 5/5 = 1$

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

0.7

FPR $= 1 - \text{Specificity}$

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

$= 1 - 1 = 0$

TPR = Sensitivity; Recall

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

$= 0.4$

0.5

FPR $= 1 - 0.4$

$= 0.6$

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

$= \frac{4}{4+1} = \frac{4}{5} = 0.8$

0.4

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

TPR $= \frac{TP}{TP+FN} = \frac{5}{5+0} = \frac{5}{5} = 1$

__/__/____

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

0.7 $\Rightarrow 0$

0.5 $\Rightarrow \frac{(0.6 - 0) \times (0.8 + 0.4)}{2}$

$= \frac{0.6 \times 1.2}{2} = \frac{0.72}{2} = 0.36$

0.4 $\Rightarrow (0.6 - 0.6) \times$

$= 0.$

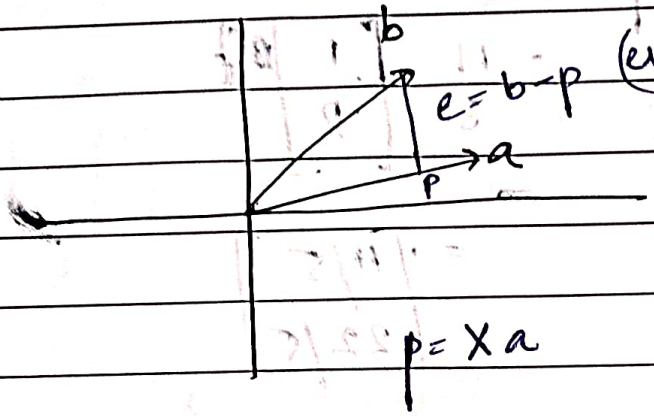
0.2 $\Rightarrow \frac{(1 - 0.6) \times (1 + 1)}{2}$

$= \frac{0.4 \times 2}{2} = \frac{0.8}{2} = 0.4$

$= 0.36 + 0.4 = 0.76$

:) a is perpendicular to e

$a \perp e$. $a^T e = 0$ $(e = b - p)$ __/__/____



(error time)

$p = Xa$

$a^T \cdot (b - p) = 0$

$a^T \cdot (b - Xa) = 0$

$a^T b = X a^T a$

$X = \frac{a^T b}{a^T a}$ → scaler

if a = unit vector

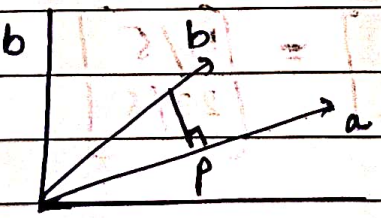
$a^T a = 1$

$p = Xa \Rightarrow \frac{a^T b}{a^T a} \cdot a$

If you want to project b onto a, do a projection matrix of a and multiply it with b.

$= \frac{a a^T b}{a^T a}$ → projection matrix.

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



$b = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$ $a = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

find p

$X = \frac{a^T b}{a^T a} \Rightarrow \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} \Rightarrow$

$\begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

3 + 8 = 11

1 + 4 = 5

~~3 + 8 = 11~~

~~1 + 4 = 5~~

$\begin{bmatrix} 3 & 6 \\ 4 & 8 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$

__/__/____

$X = \frac{11}{5}$ $\quad p = Xa$

$= \frac{11}{5} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$= \begin{bmatrix} 11/5 \\ 22/5 \end{bmatrix}$

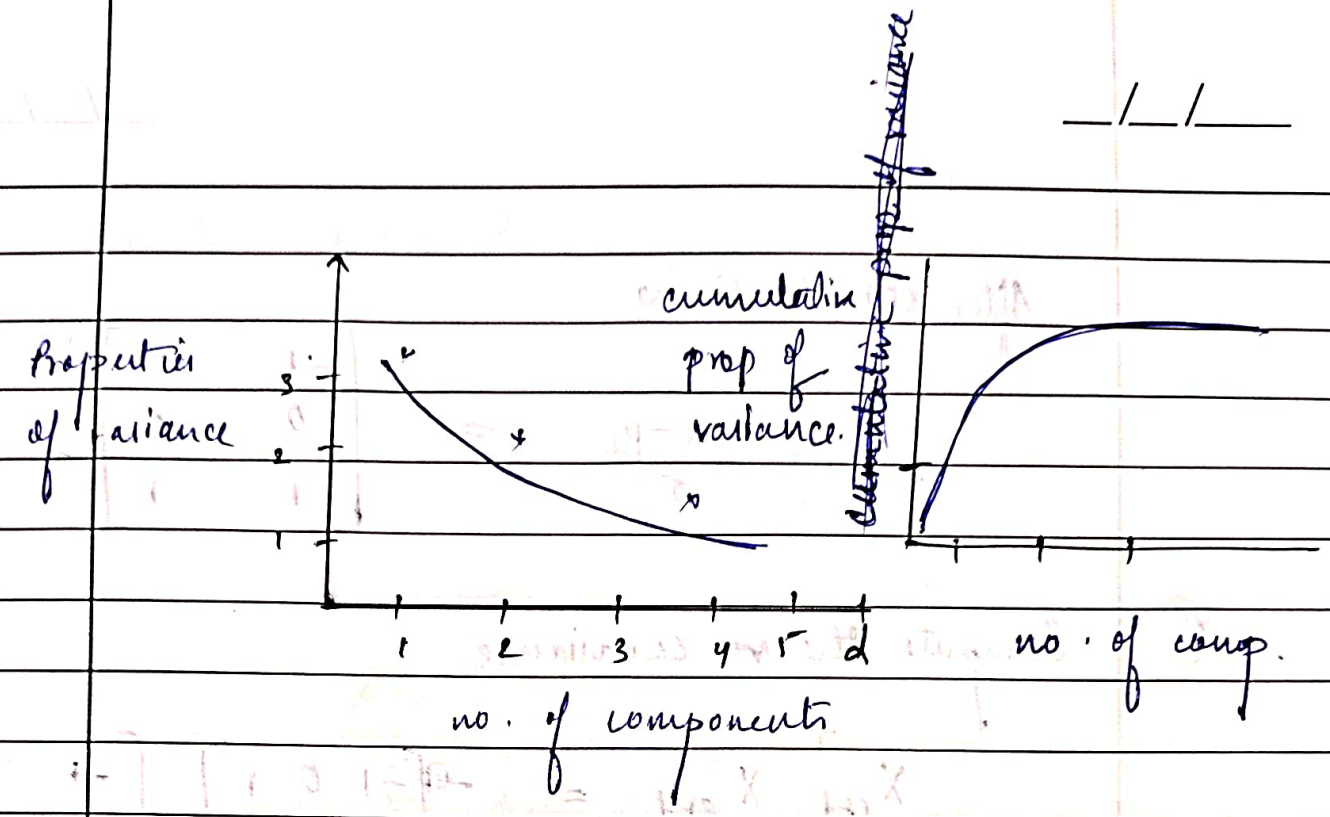
$= \frac{aa^T}{a^Ta} \cdot b \Rightarrow \frac{\begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix}}{\begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}} \cdot \begin{bmatrix} 3 \\ 4 \end{bmatrix}$

$\Rightarrow \frac{x+4}{x+4} = \frac{5}{5} \quad \frac{\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}}{5} \cdot \begin{bmatrix} 3 \\ 4 \end{bmatrix}$

$\Rightarrow \frac{\begin{bmatrix} 3+8 \\ 6+16 \end{bmatrix}}{5}$

$\Rightarrow \begin{bmatrix} 11 \\ 22 \end{bmatrix} = \begin{bmatrix} 11/5 \\ 22/5 \end{bmatrix}$

__/__/____



$$X = \begin{bmatrix} 2 & 100 \\ -4 & 200 \\ 6 & 300 \end{bmatrix} \xrightarrow[\text{PCA}]{\text{apply}} 2 = \begin{bmatrix} \quad \end{bmatrix}_{3*1}$$

(1) Standardize each column $\Rightarrow \frac{x - \mu_1}{\sigma}$

$2 + 4 + 6 = 12/3 = 4.$ $\quad \mu_2 = 200$

$\mu_1 = 4$ $\quad \sigma = \frac{1}{n}\sum_{i=1}^{n}(x-\mu)^2$

$\sigma_1 = \frac{8}{3\sqrt{2}} = 2\sqrt{4}$

$\sigma = \frac{24}{6} = \frac{24}{2 \cdot 6}$ $\quad = 2$

$\sigma_2 = 100$

$\mu_1 = 4$

$\mu_2 = 200$

$\sigma_1 = 2$

$\sigma_2 = 100.$

$(2-4)^2$

$= 4$

0

4

$8.$

$(100)^2 = 10000$

0

$(100)^2 \quad 10000$

20000

$\frac{2}{\sqrt{10k}}$

__/__/____

After standardising

$$\frac{x-\mu}{\sigma} = \begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}$$

(2) Compute the covariance.

$$X_{std}^{T} \; X_{std} = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1+0+1 & 1+0+1 \\ 1+0+1 & 1+0+1 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \; 1/n-1$$

$$\Rightarrow \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \qquad \frac{1}{2}$$

$A - \lambda I$ (3) Compute eigen values & eigen vectors.
needed

(4) Decide the % of variance and choose k principle components.

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

(6) Use these 2 to train a model.

__/__/____

$\lambda_1 = 0$, $\lambda_2 = 2$

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

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \qquad I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$A \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \lambda I = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

$$A - \lambda I = \begin{bmatrix} 1-\lambda & 1 \\ 1 & 1-\lambda \end{bmatrix}$$

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

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

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

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

$\lambda = 0 \quad \lambda = 2$

$$A - 2I = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} - 2\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

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

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

__/__/____

$-x + y = 0$

$x = y$

$\frac{1}{\sqrt{\text{length}^2}}$

$v_{n2} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $\quad v_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

(5) $Z_{pca} = X_{std} \times V_2$

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

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

$= \begin{bmatrix} \frac{-1}{\sqrt{2}} + \left(\frac{-1}{\sqrt{2}}\right) \\ 0. \\ 1/\sqrt{2} + \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \frac{-2}{\sqrt{2}} \\ 0 \\ 2/\sqrt{2} \end{bmatrix}$

Eigen vector capturing the largest eigen value will give the bigest variance.

__/__/____

a) for each of K-cluster, compute the cluster centroid.

b) Assign each observation to the cluster whose centroid is closest (Euclidean distance).

sample no	x_1	x_2
1	1	1
2	1	2
3	2	2
4	8	8
5	8	9
6	9	8

K = 2 (sample no.) 2 dimension

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

$C_2 = \{1, 5\}$

Centroid d

~~$C_1 = \text{avg} \rightarrow 3.75$~~ $C_1 = \begin{bmatrix} 5 \\ 5 \end{bmatrix}$

take the sample no. → add them and ~~take the~~ divide it by the no. of sample. $C_2 = \begin{bmatrix} 4.5 \\ 5 \end{bmatrix}$

$d = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ ← 1st sample. $C_1 - d = \begin{bmatrix} 4 \\ 4 \end{bmatrix} = 5.6$

$C_2 - d = \begin{bmatrix} 3.5 \\ 4 \end{bmatrix} = 5.3$

After 1st iteration

$C_2 = \{1\}$.

$d = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ $C_1 - d = \begin{bmatrix} 4 \\ 3 \end{bmatrix} = 5$ $C_2 - d = \begin{bmatrix} 3.5 \\ 3 \end{bmatrix} = 4.6$

__/__/____

$C_2 = \{1, 2\}$

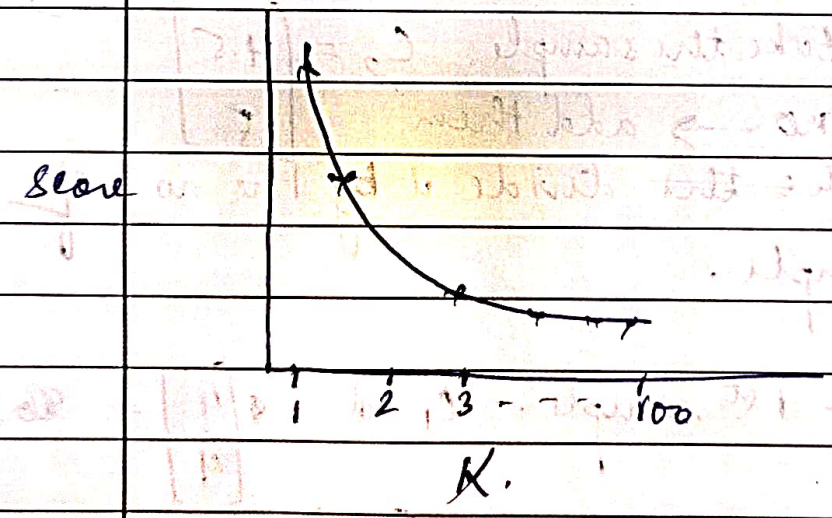
$d = \begin{bmatrix} 2 \\ 2 \end{bmatrix} \Rightarrow C_1 - d = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 4.2$

$C_2 - d = \begin{bmatrix} 2.5 \\ 2 \end{bmatrix} = 3.2$

$\therefore$

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

for practical → EDA insights to be written.

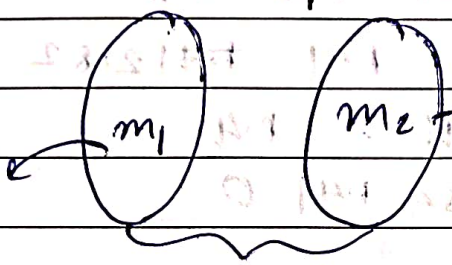


use the min. number to fuse

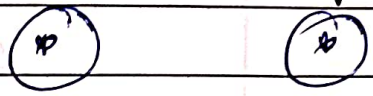
for Complete linkage max of two { ... }

min, max should be chosen on the basis of measure and data:

Average linkage



Centroid linkage



Dissimilarity between
dot product.

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

Single linkage

euclidian distance as the measure.

Step 1

$^nC_2 \Rightarrow {}^5C_2 \Rightarrow \frac{5!}{2!\,3!}$

$\Rightarrow \frac{120}{2 \times 6} = \frac{120}{12} = 10$

	A	B	C	D	E
A	0	1	3.6	5	6.4
B	1	0	2.8	4.2	5.6
C	3.6	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

$(x_2 - x_1)^2 = 1 - 2$

$A = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $B = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$

$C = \begin{bmatrix} 4 \\ 3 \end{bmatrix} = 3 = 9 + 4$

2

$D = \begin{bmatrix} 5 \\ 4 \end{bmatrix} \Rightarrow$

__/__/____

Merge A & B

New clusters are (AB) C D E

	AB	C	D	E
AB	0	2.28	4.62	5.6
C	2.28	0	1.41	2.82
D	4.62	1.41	0	1.41
E	5.6	2.82	1.41	0

$C - A = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ $\quad C - B = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$

$= 4+9$ $\quad = 4+4$

$= \sqrt{13}$ $\quad = \sqrt{8}$

$= \sqrt{13} + \sqrt{8} = 4.56 \Rightarrow 2.28$

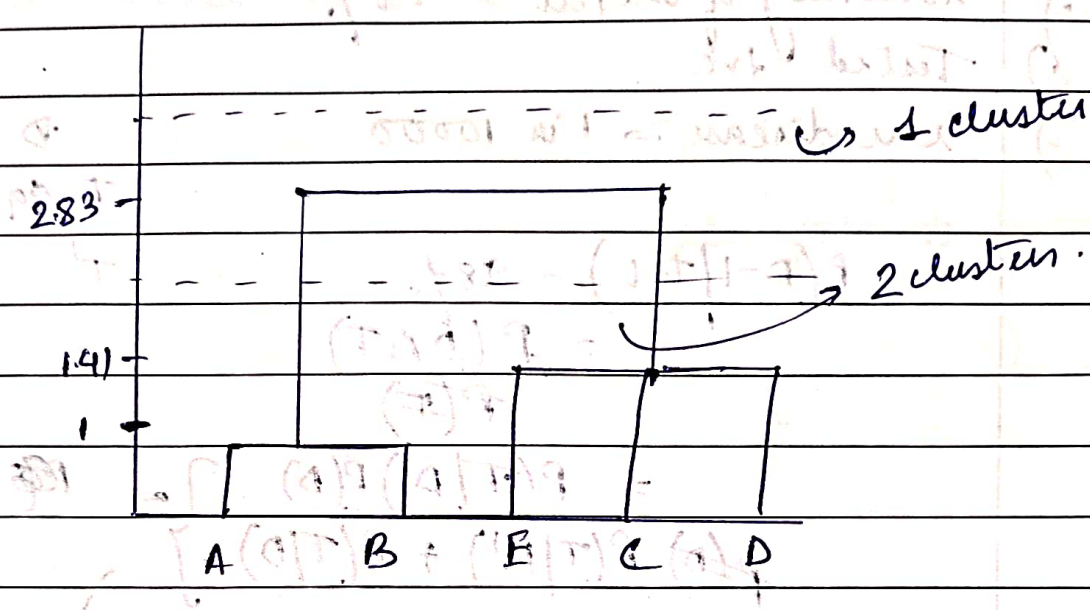
~~5+4.2~~ $D - A = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$ $\quad D - B = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$

$= \sqrt{9+16}$ $\quad = \sqrt{9+9}$

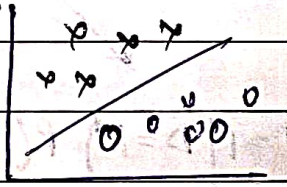
$= \sqrt{25}$ $\quad = \sqrt{18}$

$= 5 + 4.24 = 4.62$

__/__/____



Generative learning

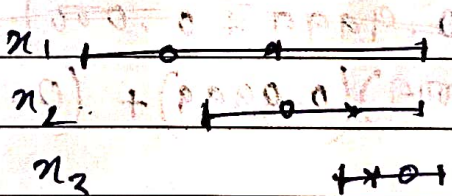


discriminative learning

discriminative models

Generative models.

capable of generating new data points (not so much into predictions)



studying objects of different classes.

__/__/____

1) Accuracy of the Test = 99%.
2) Tested +ve.
3) rare disease → 1 in 10000

	D	D'
T	99	
T'		

$$P(D=1|T=1) = 98\%.$$

$$= \frac{P(D \cap T)}{P(T)}$$

$$= \frac{P(T|D)P(D)}{P(D)P(T|D') + P(T|D)}$$

$$= \frac{(0.99)(0.0001)}{}$$

D = 0
[1 − 0.0001]

$$\frac{P(T=1|D=1)P(D=1)}{P(T=1|D=0)P(D=0) + P(T=1|D=1)P(D=1)}$$

$$\Rightarrow \frac{(0.99)(0.0001)}{(0.00009)(0.9999) + (0.99)(0.0001)}$$

$$= 0.0098/0.09$$

Naive Bayes Classifier

__/__/____

- Spam detection : deployed this model in production

vector representation has to be there for model.

If text is input then ~~[1.2, 8.4, ...]~~ ⇒

1.0
0.1
1.28
75
4.3
0
7.4
0
0
0
8.2

10×1

"To by this product"

Semantic distance between words.

a
able
buy
⋮
more
⋮
play

5000

2000

0
1
0
1
0
0
0
1
0
0

5000×1

are the spaces considered being here?

x_1	x_2	x_3	class
1	0	1	spam
1	1	0	spam
1	1	0	spam
0	0	1	no spam
0	0	0	no spam
1	1	1	no spam
0	1	0	spam
1	0	0	spam
0	1	0	spam
1	0	0	spam
1	1	0	No spam

__/__/____

preprocessing $2^3 = 8$ combinations

x_1 x_2 x_3	prob $y=1$ (spam)	prob $y=0$ (no spam)
0 0 0		
0 0 1		
0 1 0		
0 1 1		
1 0 0		
1 0 1		
1 1 0		
1 1 1	0	1/4
	0	1/4
thetas → parameters	2/7	0
	2/7	0
	1/7	0
parameters used for	2/7	1/4
bernoulli	0	1/4

learning from the data (prior)
like how many times 0 0 0 has been a spam in the previous table (probability)

this solution is not practical

→ theta and 1 − theta

a new sample has 1 1 0 (it is a spam or not spam)
it's a spam ← spam eg. 2/7 > 1/4

Naive Bayes assumption

x_i's are independent given y.
↳ conditional Independent assumption.

__/__/____

	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

	word "free"	word "win"	Label spam
1)	Yes	Yes	Spam
2)	No	Yes	spam
3)	No	No	not spam
4)	No	Yes	not spam
5)	Yes	No	not spam

Test message: Free win

Step 1: Calculate ~~prob~~.

probability of spam | test message
vs
no spam | test message.

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

__/__/____

** $n = 5$

spam = 2 $\quad P(\text{spam}=1) = \frac{2}{5} = 0.4$

not spam = 3 $\quad P(\text{spam}=0) = \frac{3}{5} = 0.6$

2. Compute likelihood :-

For spam class free appears 1 of 2 spam class

$$P(\text{free}=\text{yes} \mid \text{spam}=1) = \frac{1}{2}$$

$$P(\text{win}=\text{yes} \mid \text{spam}=1) = \frac{2}{2} = 1$$

For "not spam" class

$$P(\text{free}=\text{yes} \mid \text{not spam}) = 1/3$$

$$P(\text{win}=\text{yes} \mid \text{not spam}) = 1/3$$

Inference phase

Free = yes, win = yes

$$P(\text{spam} \mid X) = P(X \mid \text{spam})\ P(\text{spam}).$$

$$P(X \mid \text{spam}) = P(X_1, X_2 \mid \text{spam})$$

$$= P(X_1 \mid \text{spam})\ P(X_2 \mid \text{spam}).$$

__/__/____

$$P(\text{spam}|X) = P(X_1|\text{spam})\,P(X_2|\text{spam})\cdot P(\text{spam})$$

$$= P(\text{Free}=\text{yes}|\text{spam})\,P(\text{mins}=\text{yes}|\text{spam})\,P(\text{spam})$$

$$= \left(\frac{1}{2}\right)\left(1\right)\left(\frac{2}{5}\right) = \frac{2}{10} = \frac{1}{5} = 0.2$$

$$P(\text{not spam}|X) = P(X_1|\text{not spam})\,P(X_2|\text{not spam})\,P(\text{not spam})$$

$$= P(\text{Free}=\text{no}|\text{not spam})$$

$$= P(X_1, X_2|\text{Not spam})$$

$$= P(X_1|\text{not spam})\cdot P(X_2|\text{not spam})$$

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

$$= 1/6$$

$$P(\text{not spam}|X) = \frac{1}{3}\times\frac{1}{3}\times 3/5$$

$$= \frac{1}{15} = 0.066$$

$$\nabla_\theta J(\theta) = 0$$

$$X^T X\theta - X^T \vec{y} = 0$$

$$X^T \vec{y} = X^T X\theta$$

$$\boxed{\theta = (X^T X)^{-1} X^T \vec{y}}$$

Steps

1) Compute the predictions (first go row wise) $\theta = 0$

$$h_\theta x^{(1)} = \theta_0 x_0^{(1)} + \theta_1 x_1^{(1)} + \theta_2 x_2^{(1)}$$
$$= 0\times 1 + 0\times 1 + 0\times 2$$
$$= 0 + 0 + 0$$
$$= 0$$

$$h_\theta x^{(2)} = 0\times 1 + 0\times 2 + 0\times 1$$
$$= 0 + 0 + 0$$
$$= 0$$

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

2) Compute gradients.

$$e^{(1)} = h_\theta(x^{(1)}) - y^{(1)} = 0 - 3 = -3$$
$$e^{(2)} = h_\theta(x^{(2)}) - y^{(2)} = 0 - 4 = -4$$
$$e^{(3)} = h_\theta(x^{(3)}) - y^{(3)} = 0 - 5 = -5$$

go column wise

3) Calculate gradients.

$$\frac{dJ}{d\theta_0} = \sum_{i=1}^{n} (h_\theta x^{(i)} - y^{(i)})\, x_0^{(i)}$$
$$= (-3\times 1) + (-4\times 1) + (-5\times 1)$$
$$= -3 + (-4) + (-5)$$
$$= -3 - 9 = -12$$

$\frac{9+16+25}{2}$

$$\frac{dJ}{d\theta_1} = \sum_{i=1}^{n} (h_\theta x^{(i)} - y^{(i)}) x_1^{i}$$

$$= (-3)\times 1 + -(4)\times 2 + (-5)\times 3$$

$$= -3 + (-8) + (-15)$$

$$= -3 + (-23)$$

$$= -26$$

$$\frac{dJ}{d\theta_2} = \sum_{i=1}^{n} (h_\theta x^{(i)} - y^{(i)}) x_2^{(i)}$$

$$= (-3)\times 2 + (-4)\times 1 + (-5)\times 3$$

$$= -6 + -4 + -15$$

$$= -25$$

3) Update Parameters

$$\theta_0 = \theta_0 - \alpha \frac{dJ}{d\theta_0} \Rightarrow 0 - (0.1)(-12)$$

$$= 0 - (-1.2)$$

$$= 1.2$$

$$\theta_1 = 0 - \alpha \frac{dJ}{d\theta_1} \Rightarrow 0 - (0.1)(-26)$$

$$= 0 - (-2.6)$$

$$= 2.6$$

$$\theta_2 = 0 - \alpha \frac{dJ}{d\theta_2} = 0 - (0.1)(-25)$$

$$= 0 - (-2.5)$$

$$= 2.5$$

To find J

$$J_\theta = \frac{1}{2} \sum_{i=1}^{n} (h_\theta x^{(i)} - y^{(i)})^2$$

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

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

$$= \frac{50}{2} = 25$$

For 1st iteration

$$\theta = \begin{bmatrix} 1.2 \\ 2.6 \\ 2.5 \end{bmatrix} \begin{matrix} -\theta_0 \\ -\theta_1 \\ -\theta_2 \end{matrix}$$

Step 1) Compute the predictions

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

$= 1.2 \times 1 + 2.6 \times 1 + 2.5 \times 2$

$= 1.2 + 2.6 + 5.0$

$= 8.8$

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

$= 1.2 + 5.2 + 2.5$

$= 8.9$

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

$= 1.2 + 7.8 + 7.5$

$= 16.5$

Step 2) Compute gradients

$e^{(1)} = h_\theta (x^{(i)}) - y^{(i)} = 8.8 - 3 = 5.8$

$e^{(2)} = 8.9 - 4 = 4.9$

$e^{(3)} = 16.5 - 5 = 11.5$

Step 3) Calculate gradients

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

$= (5.8 \times 1) + (4.9 \times 1) + (11.5 \times 1)$

$= 22.2$

$$\frac{dJ}{d\theta_1} = \sum_{i=1}^{n} (h_\theta x^{(i)} - y^{(i)}) x_1^{(i)}$$

$$= (5.8 \times 1) + (4.9 \times 2) + (11.5 \times 3)$$

$$= 5.8 + 9.8 + 34.5$$

$$= 50.1$$

$$\frac{dJ}{d\theta_2} = \sum_{i=1}^{n} (h_\theta x^{(i)} - y^{(i)}) x_2^{(i)}$$

$$= (5.8 \times 2) + (4.9 \times 1) + (11.5 \times 3)$$

$$= 11.6 + 4.9 + 34.5$$

$$= 51$$

3) Update Parameters

$$\theta_0 = \theta_0 - \alpha \frac{dJ}{d\theta_0} \Rightarrow 1.2 - (0.1)(22.2)$$

$$\Rightarrow -1.02$$

$$\theta_1 = \theta_1 - \alpha \frac{dJ}{d\theta_1} \Rightarrow 2.6 - (0.1)(50.1)$$

$$\Rightarrow -2.41$$

$$\theta_2 = \theta_2 - \alpha \frac{dJ}{d\theta_2} \Rightarrow 2.5 - (0.1)(51)$$

$$\Rightarrow -2.6$$

To find J

$$J_\theta = \frac{1}{2} \sum_{i=1}^{n} (h_\theta x^{(i)} - y^{(i)})^2$$

$$= \frac{1}{2} \left[(5.8)^2 + (4.9)^2 + (11.5)^2 \right]$$

$$= \frac{1}{2} \times 22.2 = 11.1$$

5

K-fold cross validation

	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

$F_1\{Q_1, Q_2\} = \{1.8, 2.8\}$

$F_2\{Q_1, Q_2\} = \{2.1, 3.1\}$

$F_3\{Q_1, Q_2\} = \{19, 4\}$.

split data into 3 splits

fold 1 : 100%

x_1	x_2	y
2	-1	1
-5	1.2	0

Train set = {#1, #2, #3, #4}

~~Train set = {2, -1, 1}~~

~~Test set = {-5, 1.2, 0}~~.

~~Acc~~. Test set = {#5}

fold 2 50%

x_1	x_2	y
1	2	1
-3	-2	1

Train set = {#3, #4, #5}

Test set = ~~{3, 2, 1}~~ {#1, #2}

~~Acc~~.

fold 3 50%

x_1	x_2	y
4	0.1	0

~~Train set = {4, 0.1, 0}~~

~~Test set = {}~~.

Train set = {#1, #2, #5}

Test set = {#3, #4}.

$h_\theta x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2$

Hypothesis $= 0 \times 1 + 0 \times 2 + 0 \times (-1)$ $(\theta = 0)$

$= 0 + 0 + 0$

$=$

$h_{\theta(1)} x = \theta_1 x_1 + \theta_2 x_2$

$= (1.8 \times 2) + (2.8 \times (-1))$

$= -3.6 + (-2.8)$

$= 3.6 - 2.8 \Rightarrow +6.4 \Rightarrow \frac{1}{1+e^{6.4}}$

$h_{\theta(2)} x = \theta_1 x_1 + \theta_2 x_2$

$= 1.8 \times (-5) + (2.8 \times 1.2)$

$= -9 + 3.36 = -5.64$

$= 9 + (3.36)$

$= 12.36 \Rightarrow \frac{1}{1+e^{12.36}}$

$h_\theta x = \theta_1 x_1 + \theta_2 x_2$

$= (1.8 \times 1) + (2.8 \times 2)$

$= -1.8 + (5.6)$

$\Rightarrow 3.8$

$(e^{-3.8} = 44.701) \Rightarrow \frac{1}{1+44.701} = \frac{1}{45.701} = 0.021$

$= \frac{1}{1+e^{-12.36}}$ $(e^{-12.36} = 4.2)$

$= \frac{1}{5.2} = 0.052$

$= \frac{1}{1+e^{0.6}}$

$= \frac{1}{1+e^{1.8}}$

$= \frac{1}{6.028}$

$= 6.028$

$h_\theta x = \theta_1 x_1 + \theta_2 x_2$

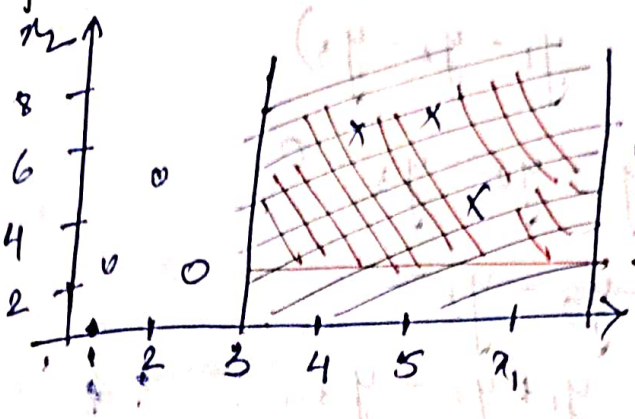
$= (1.8 \times -3) + (2.8 \times (-2)) \Rightarrow 5.4 + -5.6$

$\Rightarrow -0.2$

$= e^{-(-0.2)} = 1.22$

$\Rightarrow \frac{1}{1+1.22} \Rightarrow \frac{1}{2.22} \Rightarrow 0.022$

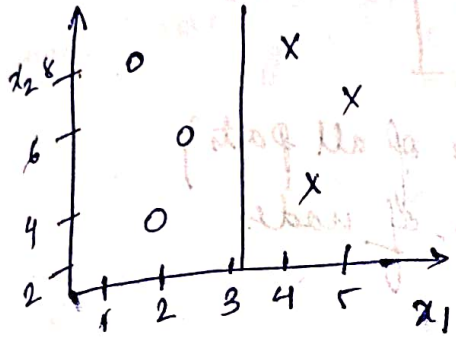
Partition Space



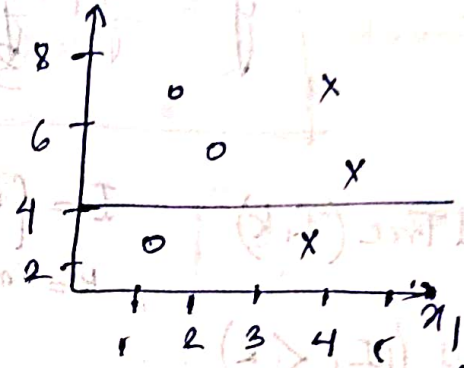
$x_1 \geq 3$ | 6 points

$x_2 \geq 2$

→ Building a Decision Tree.



$x_1 \geq 3$



$x_2 \geq 27$

This is better because it's splitting into homogenous set of cases.

$$y = \begin{bmatrix} -2 \\ 3 \\ 4 \end{bmatrix} \begin{matrix} y_1 \\ y_2 \\ y_3 \end{matrix}$$

$$f = \sum_{i=1}^{n} (\hat{y} - y^{(i)})^2$$

$$\hat{y} \cdot \alpha \quad f = (\hat{y} - y_1)^2 + (\hat{y} - y_2)^2 + (\hat{y} - y_3)^2$$

$$\min_{\hat{y}} \frac{df}{d\hat{y}} = 0 \qquad \hat{y} = 2$$

$$\frac{df}{d\hat{y}} = \frac{d(\hat{y} - y_1)^2}{d\hat{y}} + \frac{d(\hat{y} - y_2)^2}{d\hat{y}} + \frac{d(\hat{y} - y_3)^2}{d\hat{y}}$$

$$= 2(\hat{y} - y_1) + 2(\hat{y} - y_2 + 2(\hat{y} - y_3)$$

$$= 2\hat{y} - 2y_1 + 2\hat{y} - 2y_2 + 2\hat{y} - 2y_3$$

$$\Rightarrow 6\hat{y} - 2y_1 - 2y_2 - 2y_3$$

$$\Rightarrow 2(3\hat{y} - y_1 - y_2 - y_3)$$

$$3\hat{y} - y_1 - y_2 - y_3 = 0$$

~~y_1 y_2 y_3~~

$$3\hat{y} = y_1 + y_2 + y_3$$

$$\hat{y} = \frac{y_1 + y_2 + y_3}{3}$$

BuildTree(I, k) I = {indices of all parts} k = min no. of nodes

```
{
  If (|I| ≤ k)
    Set ŷ = Average(y^(i))
    return leaf(label = ŷ)
  else
    for each split-dim j → feature
    {
      for each split value S → value
      { // inner loop
        I+_{j,s} = {i ∈ I | x_j^(i) ≥ s}
        I-_{j,s} = {i ∈ I | x_j^(i) < s}
        ŷ+ = average of (y^(i))   i ∈ I+_{j,s}
        ŷ- = average of (y^(i))   i ∈ I-_{j,s}
```

Based on the chosen feature and value we will do partition.

mean on both sides

(↑ the chance of overfitting is less.

$$C_\alpha(T) = \sum_{i=1}^{n} L\left(T(x^{(i)}, y^{(i)})\right) + \boxed{\alpha |T|}$$

→ Build the procedure

leaves need not match with the training data

$|T|$ → No. of leaves

Regularization cont. → hyperparameter

Pruning of tree → removing leaves.

Classification Tree

CART

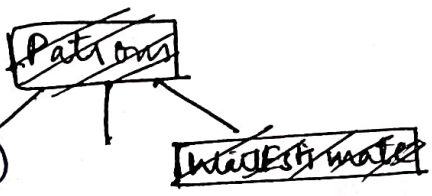
Leaves are always be the target values

↓

labels

Entropy

$$\sum_{i=1}^{K} p_i \log_2(p_i)$$



To measure homogenity we can measure the entropy and ensure if the tree is homogenous or not.

$p = 6, \; n = 6$

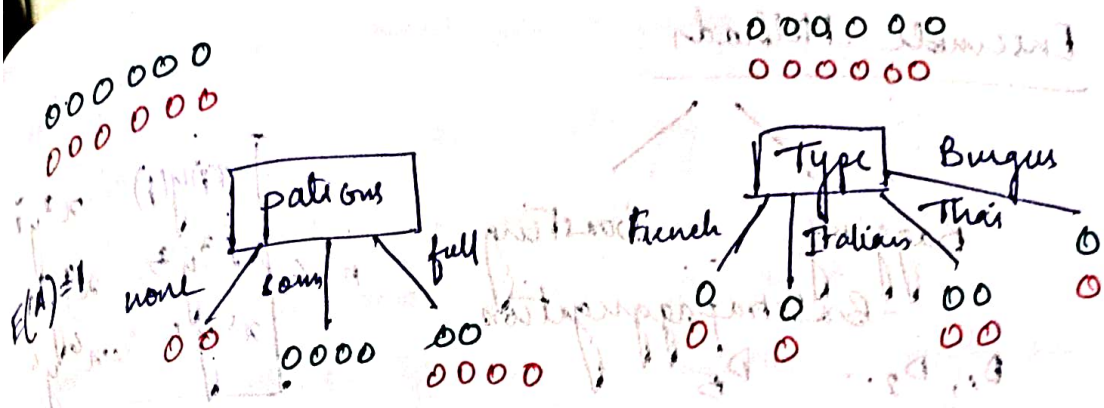
$$H\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = \frac{-p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

$$= \sum_{i=1}^{K} -p_i \log_2 p_i$$

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

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

= 1



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

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

To maximize $I(A)$ lower $EH(A)$

(*) $p_i n_i$

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

$$\frac{2}{12}(-0\log_2 0 - 1\log_2 1) + \frac{4}{12}(1\log_2 1 - 0\log_2 0)$$

$$+ \frac{6}{12}\left(\frac{2}{6}\log\frac{2}{6} - \frac{4}{6}\log\frac{4}{6}\right)$$

$$= 0.455$$

$$IG(\text{Patrons}) = H(\text{Patrons}) - EH(\text{patrons})$$

$$= 1 - 0.455$$

$$IG(\text{Patrons}) = 0.541$$

Kernel

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

transformed x into higher dimensional X and y respectively.

Dot prod $\langle \rangle$ ← symbol for multiplication.

$x = (x_1, x_2, x_3)$ ← is a vector not a feature

$y = (y_1, y_2, y_3)$

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

$x = [1, 2, 3]$

$y = [4, 5, 6]$

$\langle \phi x, \phi y \rangle$

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

$\phi x * \phi y = 16 + 40 + 72 + 40 + 100 + 180 + 72 + 180 + 324$

$= 1024$

$K(x, y) = \langle x, y \rangle^2$ (squaring on original data)

$= (1024)^2 = 1048448$ $32 \times 32 = 1024$

gradient descent on higher dimensional space is computationally expensive hence Kernel helps to compute on a lower dimensional space.