Date:

Problem.

- $\phi(x) =$

$x_1 x_1$
$x_1 x_2$
$x_1 x_3$
$x_2 x_1$
$x_2 x_2$
$x_2 x_3$
$x_3 x_1$
$x_3 x_2$
$x_3 x_3$

* commutativity not followed here.

To count transformed features :—

input → 4 features (n)

$X = [x_1 \; x_2 \; x_3 \; x_4)$

output → features mapped with degree 2 (d)

↙

all monomials (single terms) of input features where the sum of exponents is ≤ 2.

$\therefore$ Total count $= \binom{n+d}{d}$

- $x = [1, 2, 3]$

  $y = [4, 5, 6]$

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

$\phi(x)$ and $\phi(y)$ same order.

$\phi_1(x) =$

1
2
3
2
4
6
3
6
9

9×1

$\phi(y) =$

16	$y_1 y_1$
20	$y_1 y_2$
24	$y_1 y_3$
20	$y_2 y_2$
25	$y_2 y_1$
30	$y_2 y_3$
24	$y_3 y_3$
30	$y_3 y_1$
36	$y_3 y_2$

9×1

$p = 9$. here.

$\langle \phi(x), \phi(y) \rangle = 16 + 40 + 72 + 40 + 100 + 180 + 72 + 180 + 324$

$\langle \phi(x), \phi(y) \rangle = 1024.$

kernel trick lets us compute the dot product in higher dimension space without explicitly mapping to it.

papergrid

Date: / /

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

$= (4 + 10 + 18)^2 = (32)^2 = 1024$

→ able to mimic what happened in higher dimension space on lower dimension space.

- Types of kernel.

$\phi(x) = [\phi_1(x), \phi_2(x), \ldots, \phi_D(x)]$

1. linear kernel, $K(x, y) = x \cdot y$
2. Polynomial kernel, $K(x, y)$
3. Radial Basis function (RBF) ⇝ SVM
4. Laplacian kernel
5. Graph kernel
6. Fischer kernel

$1, x_1, x_2, x_1^2, x_2^2, x_1 x_2$ → captures curvature

$x_1 x_2$ → captures the interaction b/w features.

weight tells us how much $\phi(x)$ contributes to θ.

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

$\theta = \theta + \alpha \left(y^{(i)} - \theta^T \phi(x^{(i)}) \right) \phi(x^{(i)})$

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

feature map.

$\theta = \sum_{j=1}^{n} \beta_j \phi(x^{(j)})$

θ as the linear combination of vectors $\phi(x^{(1)}), \phi(x^{(2)})$

~~How is~~ correlation in kernel?

papergrid

Date: / /

11 | 03 | 2025.

SVM.

- Example.

→ Dimension of β is $(n \times 1)$

	x_1	x_2	y
	1	2	0
$n=2$	3	4	0

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

$$X_{2\times 1} = [x_1, x_2]^T$$

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

$\phi(x) \in \mathbb{R}^3$ → 3 dimension.

$\theta \in \mathbb{R}^3$

$$\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}_{2\times 2}$$

$$[\theta_1 \ \theta_2 \ \theta_3]^T$$

Find $\theta = \sum_{i=1}^{n} \beta_i \phi(x^{(i)})$ $\quad \beta_i \in \{\beta_1, \beta_2\}^T_{1\times 2}$

$\beta_1, \phi(x^{(1)}) = \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix}_{3\times 1}$, $\phi(x^{(2)}) = \begin{bmatrix} 9 \\ 16 \\ 12 \end{bmatrix}_{3\times 1}$

$$\theta = \underset{3\times 1}{\beta_1 \phi(x^{(1)})} + \underset{3\times 1}{\beta_2 \phi(x^{(2)})}$$

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

for all $i = 1, 2, \ldots, n$.

eqn (3) says that find the M which that datapoints should lie on or above M.

perpendicular distances

How to find distances from point to plane? How to calculate margin?

Perpendicular distance $(d) = \dfrac{|\beta_0 + \beta_1 x_1 + \beta_2 x_2|}{\sqrt{\beta_1^2 + \beta_2^2}}$ → shrinking coefficient; each plane weighted by the coefficient [normalization factor]

Three lines (hyperplane)

$2x_1 + 3x_2 - 5 = 0$ — (1)

$-x_1 + 4x_2 + 7 = 0$ — (2)

$5x_1 - 12x_2 + 10 = 0$ — (3)

Data

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

Q) find the optimal hyperplane?

1st hyperplane

$2x_1 + 3x_2 - 5 = 0$

$\beta_1 = 2$

$\beta_2 = 3$

$\beta_0 = -5$

Sample 1: $d_1 = \dfrac{|\beta_0 + \beta_1 x_1 + \beta_2 x_2|}{\sqrt{\beta_1^2 + \beta_2^2}}$

dist. of pt. 1 from hyperplane

Date:

$$d_1 = \frac{|-5 + 2\times 3 + 3\times 4|}{\sqrt{(3)^2 + (3)^2}} = \frac{-5+6+12}{\sqrt{13}} = \frac{13}{\sqrt{13}}$$

$$d_2 = \frac{|-5 + 2\times 2 + 3\times 3|}{\sqrt{13}} = \frac{-5+4+9}{\sqrt{13}} = \frac{8}{\sqrt{13}}$$

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

$$d_4 = \frac{|-5 + 2\times(-2) + 3\times(+1)|}{\sqrt{13}} = \frac{|-5-4+3|}{\sqrt{13}} = \frac{+6}{\sqrt{13}}$$

Minimum distance for plane (1) $\rightarrow \{6/13\} \rightarrow \{d_3, d_4$

2nd hyperplane. $\boxed{-x_1 + 4x_2 + 7 = 0}$

$$d_1 = \frac{|\beta_0 + \beta_1 x_1 + \beta_2 x_2|}{\sqrt{\beta_1^2 + \beta_2^2}} = \frac{7 + (-1)\times 3 + 4\times(4)}{\sqrt{9+16}}$$

$$= \frac{7-3+16}{\sqrt{25}} = \frac{20}{\sqrt{17}}$$

$$d_2 = \frac{7 + (-1)\times 2 + 4\times(3)}{\sqrt{17}} = \frac{17}{\sqrt{17}}$$

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

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

Date: / /

~~min dist for plane 2~~ $= d_3 = \frac{2}{\sqrt{17}}$

3rd hyperplane. $5x_1 - 12x_2 + 10 = 0$,

$d_1 = \frac{|10 + 5(3) - 12(4)|}{\sqrt{25+144}} = \frac{10+15-48}{13} = \frac{23}{13}$.

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

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

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

min dist for plane 3 $\longrightarrow$ $d_4 = \frac{+12}{13}$

Step 2. Choose the min from computed 1st dist of all hyperplanes.

$\left[\frac{6}{\sqrt{13}}, \frac{20}{\sqrt{17}}, \frac{12}{13}\right] \Rightarrow$ Margin for hyperplanes.

Step 3. ! choose the maximal margin hyperplane from step 2.

$\therefore$ Max $[1.66, 0.485, 0.923] \Rightarrow 1.66$ i.e. hyperplane (1)

$(2x_1 + 3x_2 - 5)$.

$$p = \text{proj}(a) \cdot b$$

$$= \frac{a a^T}{a^T a} \cdot b$$

$$= \frac{\begin{bmatrix} 1 \\ 2 \end{bmatrix}_{2\times1} \begin{bmatrix} 1 & 2 \end{bmatrix}_{1\times2} \begin{bmatrix} 3 \\ 4 \end{bmatrix}_{2\times1}}{\begin{bmatrix} 1 & 2 \end{bmatrix}_{1\times2} \begin{bmatrix} 1 \\ 2 \end{bmatrix}_{2\times1}}$$

$$= \frac{\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}_{2\times2} \begin{bmatrix} 3 \\ 4 \end{bmatrix}_{2\times1}}{5}$$

$$= \begin{bmatrix} \frac{1}{5} & \frac{2}{5} \\ \frac{2}{5} & \frac{4}{5} \end{bmatrix}_{2\times2} \begin{bmatrix} 3 \\ 4 \end{bmatrix}_{2\times1}$$

$$= \begin{bmatrix} 3/5 + 8/5 \\ 6/5 + 16/5 \end{bmatrix} = \begin{bmatrix} 11/5 \\ 22/5 \end{bmatrix} = \begin{bmatrix} 2.2 \\ 4.4 \end{bmatrix}$$

Date: / /

Heuristics based approach.

Algo — random initialization.

K-means → Greedy optimisation.

① Randomly assign a number from 1 to K to each observations. 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

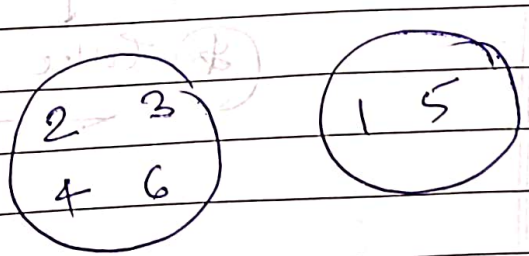
b) assign each observation to the cluster whose centroid is closest (Euclidean).

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

$k = 2$

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

$C_2 = \{1, 5\}$



After 1st iteration;

centroid$_1$ = centroid

$$= \left[\frac{1+2+8+9}{4}, \frac{2+2+8+8}{4}\right]$$

$$= [5, 5]$$

Centroid 2 : $\left[\frac{1+8}{2}, \frac{1+9}{2}\right]$

$= [4.5, 5]$

iteration (i=1)

S_1 [1 1]
C_1 [5 5]

S_1 [1 1]
C_2 [4.5 5]

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

$= \sqrt{16+16}$

$= \sqrt{32}$

$= 4\sqrt{2}$

$= 4 \times 1.41$

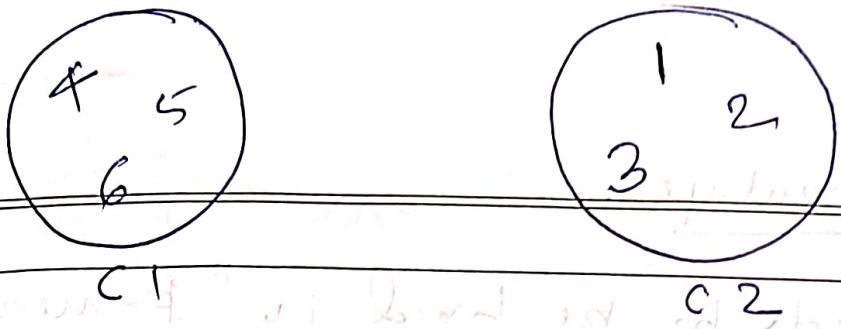
$= 5.64$

$d_{12} = \sqrt{(-3.5)^2+(4)^2}$

$= \sqrt{26.25}$

$d_{11} > d_{12}$

(*) Take min (d_{11}, d_{12}) ∴ S_1 [1] is in cluster 2.



3rd iteration

i=2

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

$= \sqrt{2r}$

$= 5$

$d_{22} = \sqrt{(1-4.r)^2 + (2-5)^2}$

$= \sqrt{(-3.r)^2 + (-3)^2}$

$= 4.60$

$\min(d_{21}, d_{22})$ → C2

i=3

$d_{31} = \sqrt{(-3)^2 + (-3)^2} = \sqrt{18}$

$d_{32} = \sqrt{(-2.r)^2 + (2-5)^2} = \sqrt{6.2r + 9}$

$= \sqrt{12.2r}$

i=4

$d_{41} = \sqrt{(3)^2 + (3)^2} = \sqrt{18}$

$d_{42} = \sqrt{(3.5)^2 + (3)^2} = \sqrt{12.2r + 9}$

$= \sqrt{21.2r}$

i=5

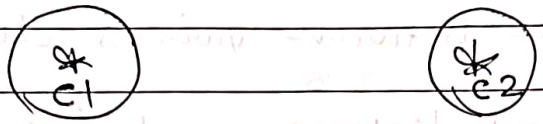
$d_{51} = \sqrt{(3)^2 + (4)^2} = \sqrt{2r}$

$d_{52} = \sqrt{(3.r)^2 + (4)^2} = \sqrt{12.2r + 16}$

$= \sqrt{28.2r}$

$\min(d_{51}, d_{52})$ → d_{51}

*. Centroid linkage



✓ Dissimilarity between dot product.

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 the measure ?

$^5C_2$ ~~of~~ pair-wise → 5
= 10

Step 1 compute the distance matrix.

	A	B	C	D	E
A	0	1	3.6	5	6.4
B	1	0	2.82	4.24	5.64
C	3.6	2.82	0	1.41	2.82
D	5	4.24	1.41	0	1.41
E	6.4	5.64	2.82	1.41	0

$$d(A-B) = \sqrt{(2-1)^2 + (1-1)^2} = \sqrt{(1)^2} = 1.$$

$$d(A-C) = \sqrt{(3)^2 + (2)^2} = \sqrt{13} = 3.6$$

$d(A-D) = \sqrt{16+9} = 5$

$d(A-E) = \sqrt{25+16} = 6.4$

$d(B-C) = \sqrt{(-2)^2 + (-2)^2} = \sqrt{8} = 2\sqrt{2}$
$= 2 \times 1.41$
$= 2.82$

$d(B-D) = \sqrt{(-3)^2 + (-3)^2} = \sqrt{18} = 3\sqrt{2}$
$= 4 \times 1.41$
$= 4.24$

$d(B-E) = \sqrt{+16+16} = \sqrt{32} = 4\sqrt{2} = 5.64$ (1.41)

$d(D-E) = \sqrt{(-1)^2 + (-1)^2} = \sqrt{2} = 1.41$

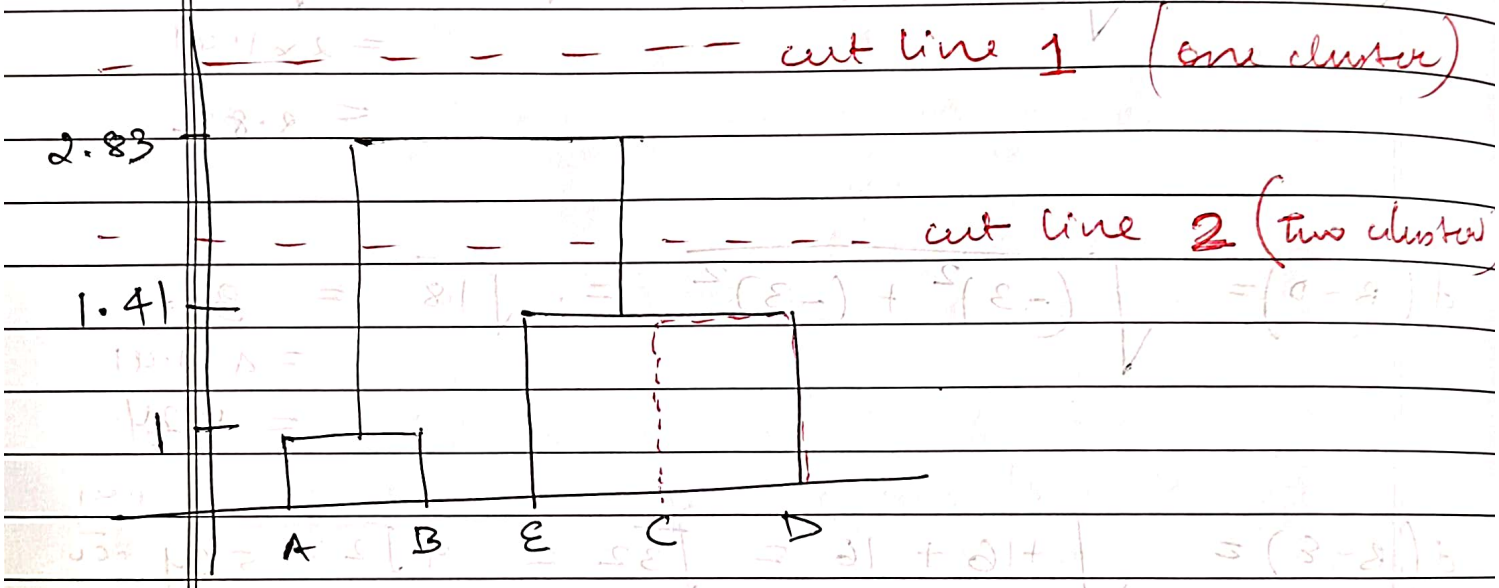
* Merge clusters A & B,

New cluster :- (AB) C D E

	AB	C	D	E
AB	0	2.82	4.24	5.64
C	2.82	0	1.41	2.82
D	4.24	1.41	0	1.41
E	5.65	2.82	1.41	0

$$\min\left[D(C,A),\ D(C,B)\right] = \min\left[3.6,\ 2.82\right]$$

$$= 2.82$$



(AB) C (DE) → (AB) (CDE)

	AB	C	DE
AB	0		
C		0	
DE			0

Date: / /

	x_1	x_2	x_3	class
1	1	0	1	Spam
2	0	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	not spam #

Total :
no spam = 4
spam = 7

↓ construct prob. table from training data. ↑ parameter.

Each of the row can be assigned a certain probability

total combination possible are $2^3 = 8$; as each feature can take 0 or 1 (Boolean values).

x_1	x_2	x_3	Prob(y=1) → spam	Prob(y=0) → not spam.
0	0	0	0	$\frac{1}{4}$ → 1 in 4 not spam.
0	0	1	0	$\frac{1}{4}$
0	1	0	$\frac{2}{7}$ → 2 had same pattern in 7 spam	0
0	1	1		
1	0	0	2/7	0
1	0	1	1/7	0
1	1	0	2/7	1/4
1	1	1	0	1/4

↓ can be explained using Binomial (Bernoulli's) distribution as the values are 0 & 1. →

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

feature choten.

	word "free"	word "win"	label
1)	Yes	Yes	spam.
2)	No	Yes	spam
3)	No	No	not spam
4)	No	Yes	not spam
5)	Yes	No	not spam

Step 1 : Calculate prior probability.

$$P(y=1 \mid x) = P(x \mid y=1)\, P(y=1)$$

$$\theta_{y=1} = \frac{\sum_{i=1}^{n} I\left(y^{(i)}=1\right)}{n} = \frac{2}{5} = P(\text{spam}=1) = 0.4$$

$$\theta_{y=0} = 1 - \frac{2}{5} = \frac{3}{5}$$

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

step 2 : compute the 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.

Free appears.

$$P(\text{Free} = \text{yes} \mid \text{not spam}) = \frac{1}{3}$$

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

Inference phase.

{Test message = Free Win.}

↓ ↓

Yes Yes.

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

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

$P(x_1 = \text{yes} \mid \text{spam} = \text{yes})$ · $P(x_2 = \text{yes} \mid \text{spam} = \text{yes})$

$$= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Free = Yes, Win = Yes

$$P(\text{spam} \mid X) = \frac{1}{4} \times P(\text{spam}) = \frac{1}{2} \times \frac{2}{5}$$

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

$P(\text{not spam} \mid x) = \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{3}{5} = \frac{1}{15} = 0.066$

papergrid

Date: / /

$0.2 > 0.066$ # compare both prob & assign the max.

↓

message belongs to spam class.

Date: 21/03/2025

observation	Actual labels	Predicted score	Predicted label (0.4)	Predicted label (0.7)	
1	1	0.85 (TP)	1	1	
2	0	0.60	0	1	* (FP)
3	1	0.70 (TP)	1	1	
4	1	0.40	1	0	* (FN)
5	0	0.55	0	1	* (FP)
6	1	0.50	0	1	
7	0	0.65	0	1	* (FP)
8	0	0.35 (TN)	0	0	
9	1	0.60 (TP)	1	0	
10	0	0.20 (TN)	0	0	

0.5 | 0.7 (0.5)

1) Accuracy
2) Precision
3) Recall or sensitivity or True positive Rate
4) Specificity
5) False positive Rate (1 - specificity)
6) F1-score
7) ROC plot
8) AUC.

Confusion matrix

	original positive (TP)	original -ve (TN)
predicted +ve	1,3,9 (3)	2,7,5 (3)
predicted -ve	4 (1)	8,10 (2)

① Accuracy = (TP + TN) / Total

= 8 / 10 = 0.8

② precision = TP / (TP + FP) = 4 / 7 = 0.57

③ Sensitivity = 4 / 4+1 = 4/5 = 0.8 = TPR,

④ specificity = 3/5 = 0.4

⑤ False positive rate (1 − 0.4) = 0.6.

⑥ F1-score = 2 · 0.57 · 0.8 / (0.57 + 0.8) = 0.66.

Threshold.

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

Q70.

	OP	ON
PP	2	0
PN	3	5

TPR = 2/5 = 0.4

specificity = FP / N = 5/5 = 1

FPR = 0.

	OP	ON
PP	5	3
PN	0	2

precision = $\frac{5}{8} = 0.4$

~~specificity~~

sensitivity = $\frac{5}{5} = 1$

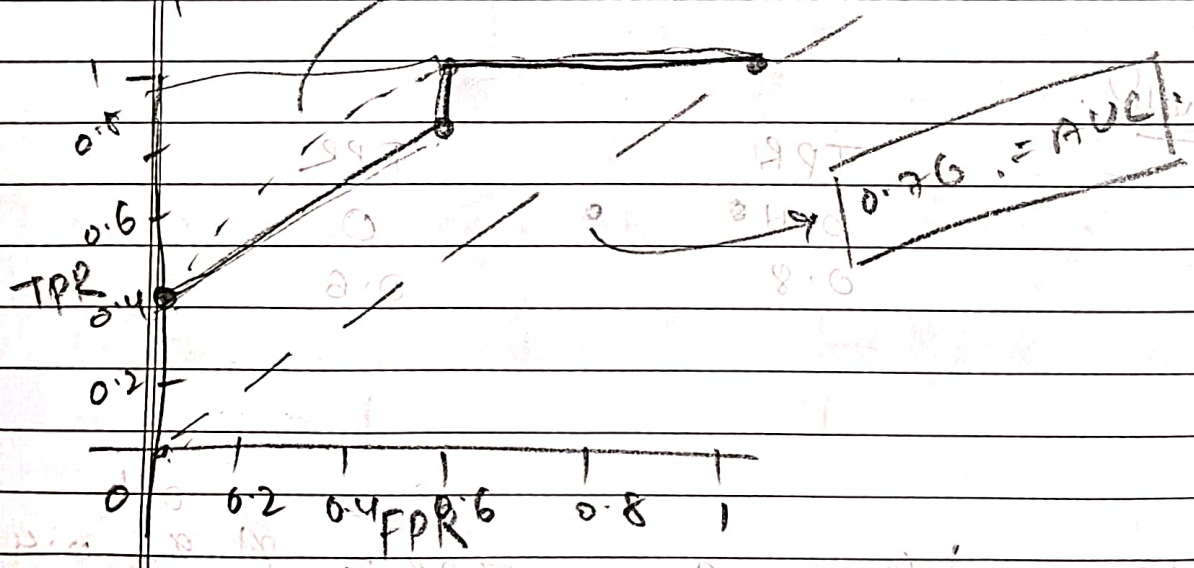
FPR specificity = $\frac{2}{5} = 0.4$

FPR = $1 - 0.4 = 0.6$

	OP	ON
PP	5	5
PN	0	0

TPR = 1

FPR = 1



$$AUC = \sum_{i=1}^{4} \frac{(FPR_i - FPR_{i-1}) * (TPR_i + TPR_{i-1})}{2} = 0.76$$

① $(0 - 0) * = 0$

② $\frac{(0.6)}{0.3} \times (0.8 + 0.4) = 0.3 * 0.12$

$= 0.36.$

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

$h_\theta(x) = x_0\theta_0 + x_1\theta_1 + \cdots$

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

$\alpha = 0.1$

Date: 16/01/2025

① compute the prediction. (for each & every sample).

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

$$= 0.1 + 0.1 + 0.2$$

$$= 0.$$

$$h_\theta(x^{(2)}) = 0\times1 + 0\times2 + 0\times1$$

$$= 0$$

$$h_\theta(x^{(3)}) = 0\times1 + 0\times3 + 0\times3$$

$$= 0$$

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

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

② compute the gradients.

i) compute errors / residuals.

$$e^{(1)} = \underbrace{h_\theta(x^{(1)})}_{\hat{y}} - y^{(1)} = 0 - 3 = -3$$

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

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

ii) compute Gradients.

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

$$e^1 \cdot x_0^1 + e^2 \cdot x_0^2 + e^3 \cdot x_0^3$$

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

$= -12$

$$\frac{\partial J}{\partial \theta_1} = \sum_{i=1}^{3} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_1^{(i)}$$

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

$= -26$

$$\frac{\partial J}{\partial \theta_2} = \sum_{i=1}^{3} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_2^{(i)}$$

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

$= -6 - 4 - 15$

$= -25$

(3) Update parameters

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

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

$= 0 + 1.2$

$\boxed{\theta_0 = 1.2}$

* $\theta_1 := \theta_1 - \alpha \frac{\partial J}{\partial \theta_0}$ $\quad 0 - 0.1 \times (-26)$

$\boxed{\theta_1 = 2.6}$

$\theta_2 := \theta_2 - \alpha \frac{\partial J}{\partial \theta_0} = 0 - 0.1 \times (-25)$

$\boxed{\theta_2 = +2.5}$

papergrid

Date: / /

∴ New parameters → $\theta_0 = 1.2$, $\theta_1 = 2.6$, $\theta_2 = 2.5$ } After 1st iteration.

Q. compute J value ? → $\boxed{J = 44.95}$ → missed the global minima.

due to the step size i.e. α.

After 2nd iteration

$\theta_0 = -1.02$, $\theta_1 = -2.41$, $\theta_2 = -2.6$.

$\boxed{J = 2568.56}$.

Reduce the learning rate α from 0.1 to 0.01 & iterate again.

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

$J = 10.81$

After 1st iteration: $\theta_0 = 0.19$, $\theta_1 = 0.43$, $\theta_2 = 0.41$

$\boxed{J = 5.4}$

After 2nd iteration $\theta_0 = 0.25$, $\theta_1 = 0.55$, $\theta_2 = 0.53$

$\boxed{J = 3}$

↳ decreases!

Date: 24 | 01 | 2025

k-fold cross-validation

1
2
K

	x_1	x_2	y
I	2	-1	1
II	0.5	1.2	0
III	1	2	1
IV	-3	-2	1
V	4	0.1	0

3-fold CV
Acc.
std.

Fold 1 $[D_1, D_2] = [-1.8, 2.8]$
Fold 2 $[D_1, D_2] = [2.1, 3.1]$
Fold 3 $[D_1, D_2] = [1.9, 4]$

Compute Avg. Acc & Avg. std deviation ?

$k = 3$.

Fold I → Train set: {I, II}
Test set: {III}

	x_1	x_2	y	
I	2	-1	1	
II	0.5	1.2	0	Acc = ? 100 %
III	1	2	1	
IV	-3	-2	1	
V	0.1		0	

Fold 2

Train set: {I, III}
Test set: {II}

Fold 3.

Train set: {II, III}
Test set = {I}

Avg = 66.67 %.

compute $\theta^T X$,

$$\theta = \begin{bmatrix} 2 \\ 0.5 \\ 1 \\ -3 \end{bmatrix} \quad X = \begin{bmatrix} -1.8 & -1 \\ 2.8 & +1.2 \\ 2 & 2 \\ 1 & -2 \end{bmatrix} \quad 4 \times 2$$

$$\theta^T X = \begin{bmatrix} 2 \times -1.8 + (-1) \times 2.8 \\ 0.5 \times -1.8 + (1.2) \times 2.8 \\ +1 \times -1.8 + 2 \times 2.8 \\ -3 \times -1.8 + (-2) \times 2.8 \end{bmatrix} = \begin{bmatrix} -6.4 \\ 2.46 \\ 3.8 \\ -0.2 \end{bmatrix}$$

$$g(z) = \frac{1}{1+e^{+6.4}} = 0.001658$$

$$g(z) = \begin{bmatrix} 0.001658 \\ 0.9218 \\ 0.978 \\ 0.45 \end{bmatrix}$$

Threshold = 0.5

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

Test fold. (I) $X = \begin{bmatrix} 4 & 0.1 \end{bmatrix}$ $\theta^T = \begin{bmatrix} -1.8 \\ 2.8 \end{bmatrix}$

$\theta^T X =$

$= -1.8 \times 4 + 2.8 \times 0.1$

$= -7.2 + 0.28$

$= -6.92$

$g(z) = 0.0009$

$\hat{y} = 0 \quad | \quad y = 0$

Acc = 100%

(II) $X = \begin{bmatrix} 2 & -1 \\ 0.5 & 1.2 \end{bmatrix}$ $\theta^T = \begin{bmatrix} 2.1 \\ 3.1 \end{bmatrix}$

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

$= \begin{bmatrix} 1.1 \\ 4.72 \end{bmatrix}$

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

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

Acc = 50%

(iii)

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

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

2×2

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

$$g(z) = \begin{bmatrix} 0.999 \\ 0.0002 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$Acc = 50\%$$

* Avg. Acc $= \dfrac{50\% + 100 + 100}{3} = \dfrac{200}{3} = 66.67\%$

* Std. deviation

why did we not use SSE in logistic?

Entropy $-\sum_{i=1}^{k} p_i \cdot \log_2(p_i)$

A good split reduces entropy of the label. ∴ ↑ IG.

* In Restraunt data set, how to choose between Type & patron to be the starting node?

→ to measure the homogeneity for the tree, by measuring the entropy!

IG = Information Gain → always +ve, converges to zero at extreme thresholds (very low or very high)

$p = 6, n = 6.$ p → +ve samples, n → -ve samples.

→ 0, if both labels are of same

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

H → entropy; 2 → bits / encoding

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

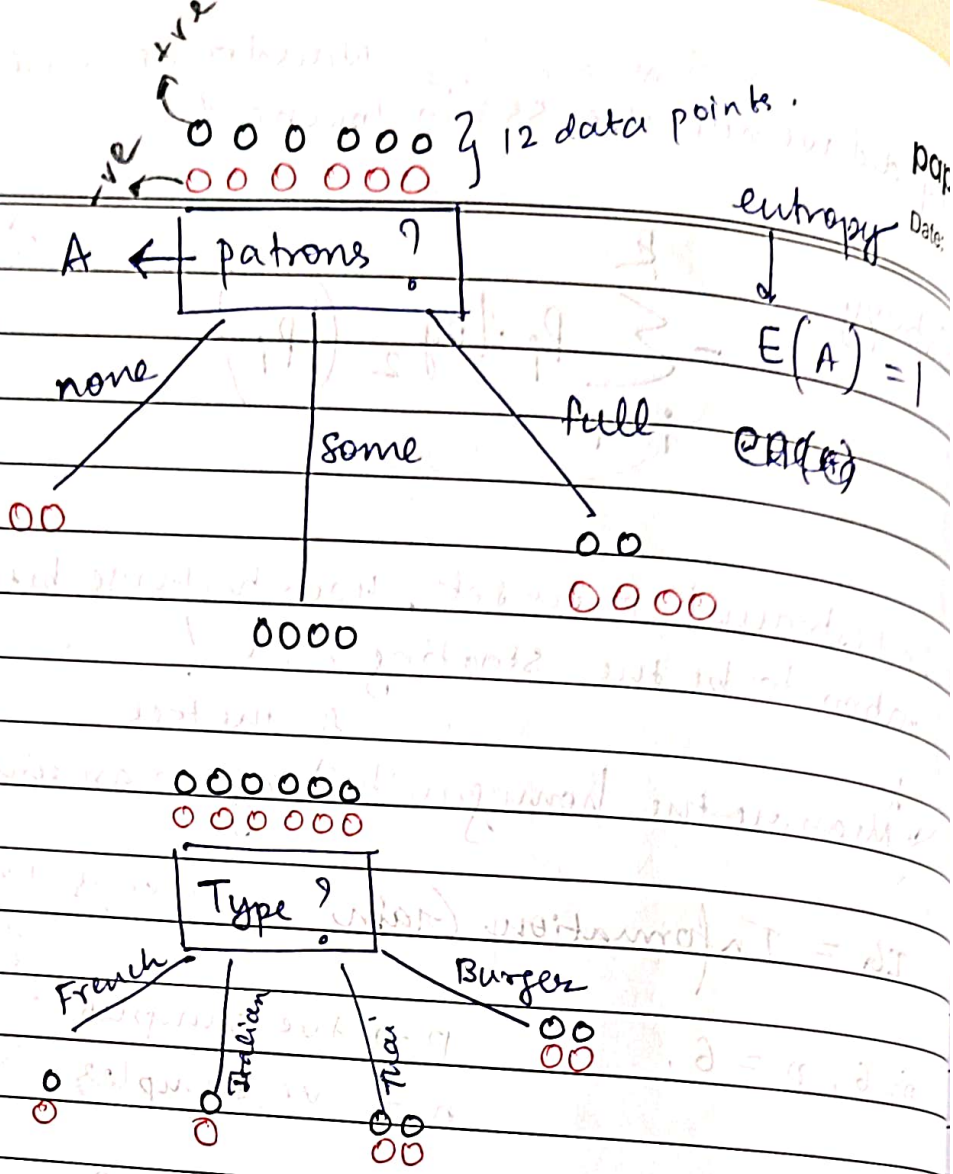
$$H\left(\frac{6}{12}, \frac{6}{12}\right) = -\frac{6}{12}\log_2\frac{6}{12} - \frac{6}{12}\log_2\frac{6}{12}$$

$$= -0.5\log_2(0.5) - 0.5\log_2(0.5)$$

$$= -0.5 \times (-1) - (0.5 \times -1)$$

$$= 0.5 + 0.5$$

$$= 1$$



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

↳ weighted factor.

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

information gain → how much impurity I can shed?

↓

we want maximum information!

↓

criteria to choose the feature ....

Date: 11/02/2025.

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

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

$$= 1$$

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

$$= -\left[\frac{p_1+n_1}{12} H\left(\frac{0}{2}, \frac{2}{2}\right) + \frac{p_2+n_2}{12} H\left(\frac{4}{4}, \frac{0}{4}\right) + \frac{p_3+n_3}{12} H\left(\frac{2}{6}, \frac{4}{6}\right)\right]$$

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

$$\left(-\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2}\right)$$

$$= -\left[\frac{1}{6}(-0\log 0 - 1\log_2 1) + \frac{1}{3}(-1\log_2 1 - 0\log_2 0) + \frac{1}{2}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right)\right]$$

$$= \frac{1}{6}\left[\frac{1}{3}\log_2(3) - \frac{2}{3}\log_2(0.67)\right]$$

$$\frac{1}{2}\left[\frac{1}{3} \times 1.5849 - \frac{2}{3} \times (-0.577)\right]$$

$$0.5283$$

$EH(\text{patrons}) = \cancel{0.546}.\ 0.456$

* $IG(\text{patrons}) = H(\text{patrons}) - EH(\text{patrons})$
$= 1 - 0.456$
$= 0.541.$

* $IG(\text{Type}) = 0.$

Gini index: → measure like Information Gain.

• If the Information gain is equal for two feat then we choose arbitrarily.

Ensemble methods.

Bagging (Bootstrap aggregation)

Boosting

→ multiple trees or model can be constructed fr given dataset. → individual tree or mode are referred to ~~as weak learner~~ → we tr to aggregate the learning using Ensemble method.

T ⇒ tree.

Dataset (D) → $x^{(1)}, y^{(1)}$ | $x^{(2)}, y^{(2)}$ | $x^{(3)}, y^{(3)}$ | $x^{(4)}, y^{(4)}$ | $x^{(5)}, y^{(5)}$ | $x^{(6)}, y^{(6)}$ → T_1

n=6.

Bootstrap sampling, (n, n) → possi

→ randomly draw n data points from dataset (D) & create a new dataset called as 'Bootstrapping'

(D') → x^3, y^3 | x^6, y^6 → T_2

Dataset:

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

i) initialize $F(x) = 0$

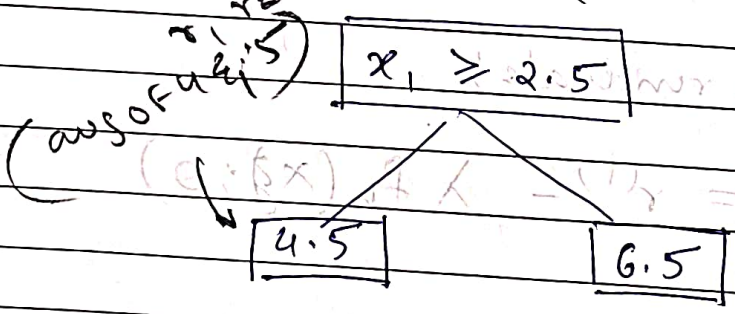
$f_0(x) = 0$

Iteration 1 $r = y$

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

$x_1 \geq 2.5$

a) Fit a model, $f_1(x) \rightarrow r$



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

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

assumed

$\lambda = 0.1$

c) update the residuals.

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

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

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

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

papergrid Date: / /

Iteration 2 $r = \begin{bmatrix} 3.55 \\ 4.55 \\ 5.35 \\ 6.35 \end{bmatrix}$

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.

$x_1 \geq 2.5$?

4.05 — # Avg of 3.55 & 4.55

5.85 — # Avg of 5.35 & 6.35.

(b) update the model.

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

(c) update the residuals

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

$r^{(2)} = r^{(2)} - \lambda f_2(x^{(2)}) = 4.145$

$r^{(3)} = r^{(3)} - \lambda f_2(x^{(3)}) = 4.765$

$r^{(4)} = r^{(4)} - \lambda f_2(x^{(4)}) = 5.765$

Residuals decreases !! ✓

∴

Final : $F(x, 0) = f_0(x) + \lambda f_1(x) + \lambda f_2(x) + \lambda f_3(x) + \lambda f_B(x)$.