16/01/25 prev problem

Example ① $h_\theta(x^{(1)}) = 0$; $h_\theta x^{(2)} = 0$, $h_\theta x^{(3)} = 0$

Initialization :- $\theta_0 = 0$; $\theta_1 = 0$, $\theta_2 = 0$; $J = \frac{1}{2} \sum_{i=1}^{n} (h_\theta(x^{(i)} - y^{(i)})^2$

$J = \frac{1}{2} [(0-3)^2 + (0-4)^2 + (0-5)^2] = \frac{1}{2} [50] = 25$

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

$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \cdots$

Initializing $\theta_0 = \theta_1 = \cdots = 0$.

$\alpha = 0.1$ learning rate

① Compute the predictions

superscript → instance

$x_1, x_2 \cdots$ → input feat

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

$= 0*1 + 0*1 + 0*2$

$= 0$

$h_\theta(x^{(2)}) = 0*1 + 0*2 + 0*3$

$= 0$

$h_\theta(x^{(3)}) = 0*1 + 0*2 + 0*3$

$= 0$

Compute gradients

i) Compute errors / residuals

1st sample $e^{(1)} = h_\theta(x^{(1)}) - y^{(1)}$

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

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

$= -4$

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

$= -5$

ii) Calculate gradients

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

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

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

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

$= -3 * 1 + -4 * 2 + -5 * 3$

$= -26$

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

$= -3 * 2 + -4 * 1 + -5 * 3 \Rightarrow -6 - 4 - 15$

$= -25$

iii) Update parameters

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

(prev step limit to 0) $= 0 - 0.1 * (-12)$

$= 0 + 1.2$

$= 1.2$

$$\theta_1 = \theta_1 - \alpha \frac{\partial J}{\partial \theta_1}$$

$= 0 - 0.1 * -2.6$

$= 2.6$

$$\theta_2 = \theta_2 - \alpha \frac{\partial J}{\partial \theta_2}$$

$= 0 - 0.1 * -2.5$

$= 0 + 2.5$

$= 2.5$

at beginning

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$J = 25$

after 1st iteration

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

$J = 94.95$

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

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

$J(1) = \frac{1}{2} (0-3)^2$

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

$J(2) = (0-4)^2 = 16$

$J(3) = (0-5)^2 = 25$

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

$= \frac{1}{2} \times 50 = 25$

$J(\theta) = 25$

after 1st iteration

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

$= \frac{1}{2}$

After 2nd iteration

$\theta_0 = -1.02$; $\theta_1 = -2.41$; $\theta_2 = -2.6$

$J = 25678.56$

$\alpha = 0.01$ change (as we take lower steps)

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

$J = 10.8$

↓

After 1st iteration

$\theta_0 = 0.19$, $\theta_1 = 0.43$, $\theta_2 = 0.41$

$J = 5.4$

After 2nd iteration

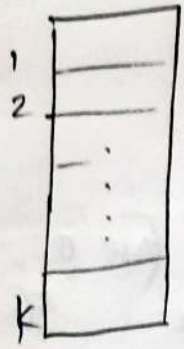
$\theta_0 = 0.25$, $\theta_1 = 0.55$, $\theta_2 = 0.5$

$J = 3$

If datasets are large we needn't use cross-val.
Train test split is sufficient.

24/01/26

K-fold cross valid^n



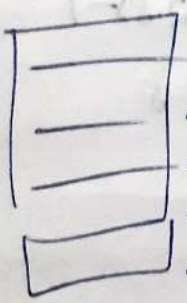
split as test
other as train

	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 CV
↳ print Acc
↳ "std dev
ignore O_0 (bias)

Fold 1:

$T_1\{O_1, O_2\} = \{-1.8, 2.8\}$

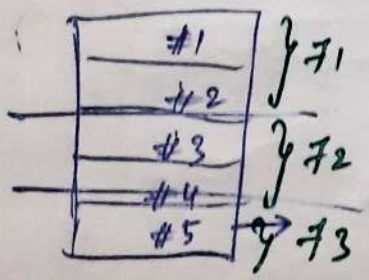


5 samples
divide into 3 parts

$T_2\{O_1, O_2\} = \{2.1, 3.1\}$

$T_3\{O_1, O_2\} = \{1.9, 4\}$

Compute Acc, Std dev^n



3 parts

Fold 1

Train set {#1, #2, ~~#3~~ #3, #4}
Test set {#5}

We can perform diff splits ⇓ sklearn

Fold 2

Train set {#3, ~~#4~~ / ~~#5~~, ~~#5~~}
Test set {#1, ~~#2~~}

X

Fold 3

Train set {#1, #2, #3}
Test set {#4, #5}

X

or

Train {#1, #3, #4, #5}
Test {#2}

Fold 2

Train set {#3, #4, #5}
Test set {#1, #2}

J_1 {-1.8, 2.8}

J_2 {2.1, 3.1}

J_3 {1.9, 4}

Fold 3

Train set {#1, #2, #5}
Test set {#3, #4}

$[J_1 \Rightarrow (1)(2)(3)(4)]$ (use θ_1, θ_2 of J_1)

$\frac{\partial}{\partial \theta_j} l(\theta) = [y - h_\theta x] x_j$

(1) $h_{\theta(1)} = \theta_1 x_1 + \theta_2 x_2$
$= -1.8 \times 2 + 2.8 \times 1$
$= -3.6 - 2.8$
$h_\theta = -6.4$

(2) $h_{\theta(2)} = \theta_1 x_1 + \theta_2 x_2$
$= -1.8 \times 0.5 + 2.8 \times 1.2$
$= -0.9 + 3.36 = 2.46$

(3) $h_{\theta(3)} = \theta_1 x_1 + \theta_2 x_2$
$= -1.8 \times 1 + 2.8 \times 2$
$= -1.8 + 5.6 = 3.8$

(4) $h(\theta)4 = \theta_1 x_1 + \theta_2 x_2$
$= -3 \times -1.8 + +2.8 \times (-2)$
$= 5.4 - 5.6 = -0.2$

$x_i \Rightarrow 1, 2, 3, 4 \quad \theta_1 = -1.8 \; ; \; \theta_2 = 2.8$

$$g(\theta^T x_1) = \frac{1}{1 + e^{-\theta(-1.8)x}}$$

$$= \frac{1}{1 + e^{-(-6.4)}} = \frac{1}{1 + 601.8450}$$

$$g(\theta_2^T x_2) = \frac{1}{1 + e^{-(2.46)}} = \frac{1}{1 + 0.0854}$$

threshold > 0.5

$$g(\theta_3^T x_3) = \frac{1}{1 + e^{-(3-8)}} = \frac{1}{1 + 0.223707}$$

$$g(\theta_4^T x_4) = \frac{1}{1 + e^{-(0.2)}} = \frac{1}{1 + 0.8187}$$

$= 0.5498$

$g(\theta^T x_1) = 0.0016 =$ compare with y $\hat{y} = 0$ (0)

$g(\theta^T x_2) = 0.92131 = \hat{y} = 0$ (0)

$g(\theta^T x_3) = 0.817890 \; (0.98711) = \hat{y} = 1$ (1)

$g(\theta^T x_4) = 05498 = \hat{y} = 1$ (1)

For test:-

$$h\theta = \theta_1 x_1 + \theta_2 x_2$$

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

$$= -7.2 + 0.28$$

$$= -6.92$$

$\hat{y} = 1$

Accuracy = 100%.

$$g(\theta^T x_1) = \frac{1}{1 + e^{+6.92}} \rightarrow$$

$$= \frac{1}{1013.319}$$

$= 0.000986 =$

Fold 2 :- Test set is #1, #2

$h_\theta(x_1) = \theta_1 x_1 + \theta_2 x_2$

$= 2.1 \times (2) + 3.1 \times (-1)$

$= 4.2 - 3.1$

#1 $h_\theta(x_1) = 1.1$

#2 $h_\theta(x_2) = 2.1 \times (0.5) + 3.1 \times (1.2)$

$= 1.05 + 3.72$

$= 4.77$

$g_\theta(x_2) = \frac{1}{1+e^{-(4.77)}} = \frac{1}{1+0.008480}$

$= 0.9915 \quad \vec{y}_2 = 0$ Label

$g(\theta) x_1 = \frac{1}{1+e^{-(1.1)}} = \frac{1}{1+0.33287}$

$= 0.75026$

Label $\vec{y}_1 = 1$

Accuracy = 50%

Fold 3 :- Test set is #3 #4

$h_\theta(x_3) = \theta_1 x_1 + \theta_2 x_2$

$= 1.9 \times 1 + .9 \times 2 = 1.9 + 8$

$= 9.9$

$h\theta(x_4) = 1.9 \times (-3) + 4 \times (-2)$

$= -5.7 - 8$

$= -13.7$

$g(\theta)_3 = \frac{1}{1+e^{-4.9}} = \frac{1}{1+0.00005} = \frac{1}{1.00005} = 0.9995$

Label $\vec{y}_3 = 1$

$g(\theta)_4 = \frac{1}{1+e^{-(-13.7)}} = \frac{1}{1+890911.16597} = \frac{1}{890912.165}$

$= 0.000001124$

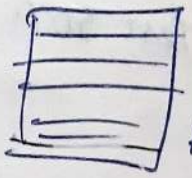
Label $\vec{y}_4 = 0$

Acc = 50%

$k=n$

n - no. of samples

k - folds ⇒ if $k=n$ → Leave One Out Cross Valid^n



99 Train
1 Test

computationally LOOCV is expensive

* When shd we use LOOCV?
* Should we use?

Whenever dataset <100,
use LOOCV

Guideline:- ~ extremely small dataset <100

[fold valid technique is not to improve accuracy just to increase confidence]

To train a ML Model:-

1) Cross valid^n
   - k-fold >100.
   - LOOCV <100
2) Train test split

Dataset

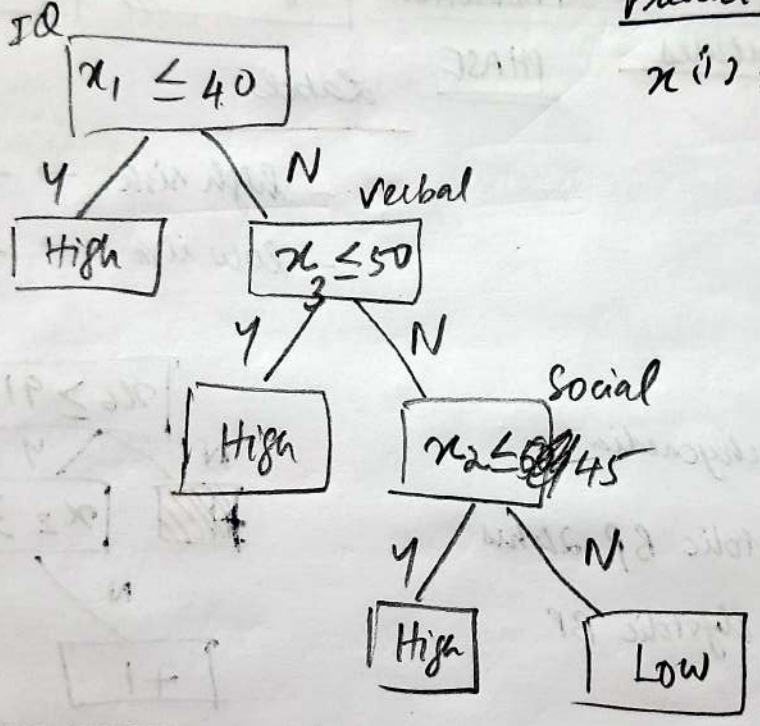
IQ (x_1)	Social (x_2)	Verbal (x_3)	Risk for Autism
100	79	86	Low
30	40	70	High
80	85	65	Low
90	82	45	High

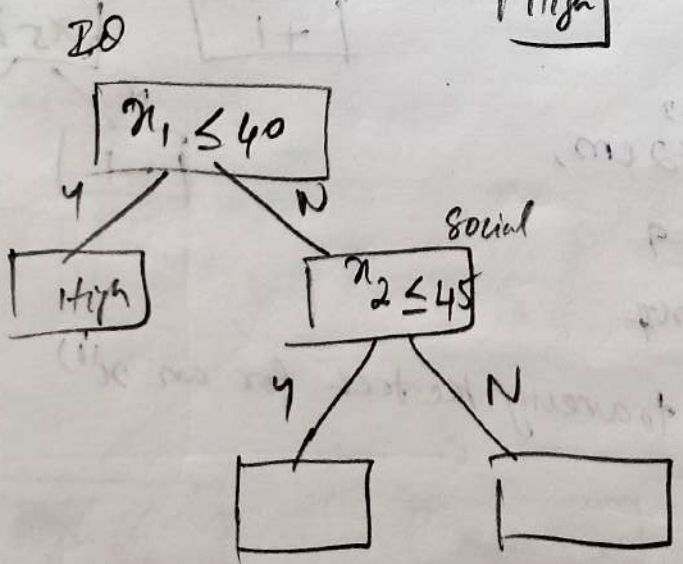
Predict for

$x(i)$ = IQ : 120
Social = 60
Verbal = 40

Predict

~~IQ~~
High





Leaf - Label

Internal nodes
⇓
feature, threshold value

Randomly init a tree (theta 100)

↓

Sum of sq. error for all samples

↓

Optimize (at every internal node what shd be the feature & what is the parameter)

We cannot apply grad. descent as the tree is rude

How do we construct trees using the algo? (my confusion?)

Tree

↓

Fix - cost function → annoy

↓

Optimize → keep constructing more trees

↓

Construct trees by varying param

↓

Take that tree where error is ↓

Example

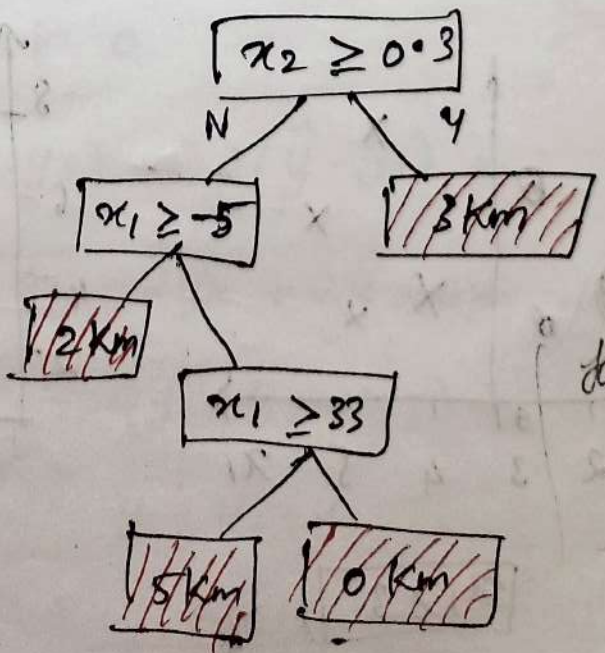
x_1 : Temperature (degree)

x_2 : Precipitation (cm/hr)

y: 1 Km
2 Km
4 Km
⋮

Regression tree

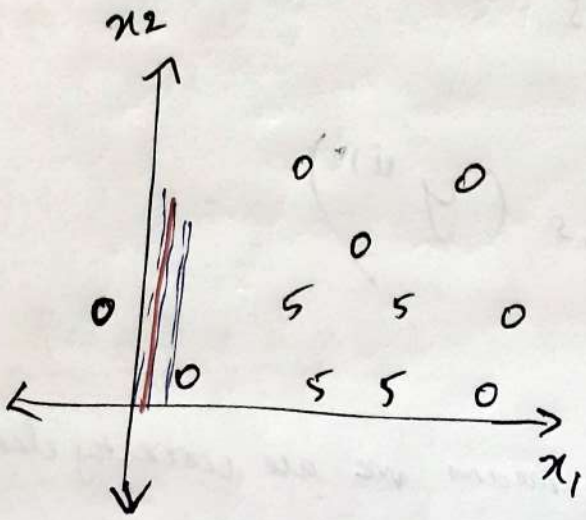
Leaves - y val

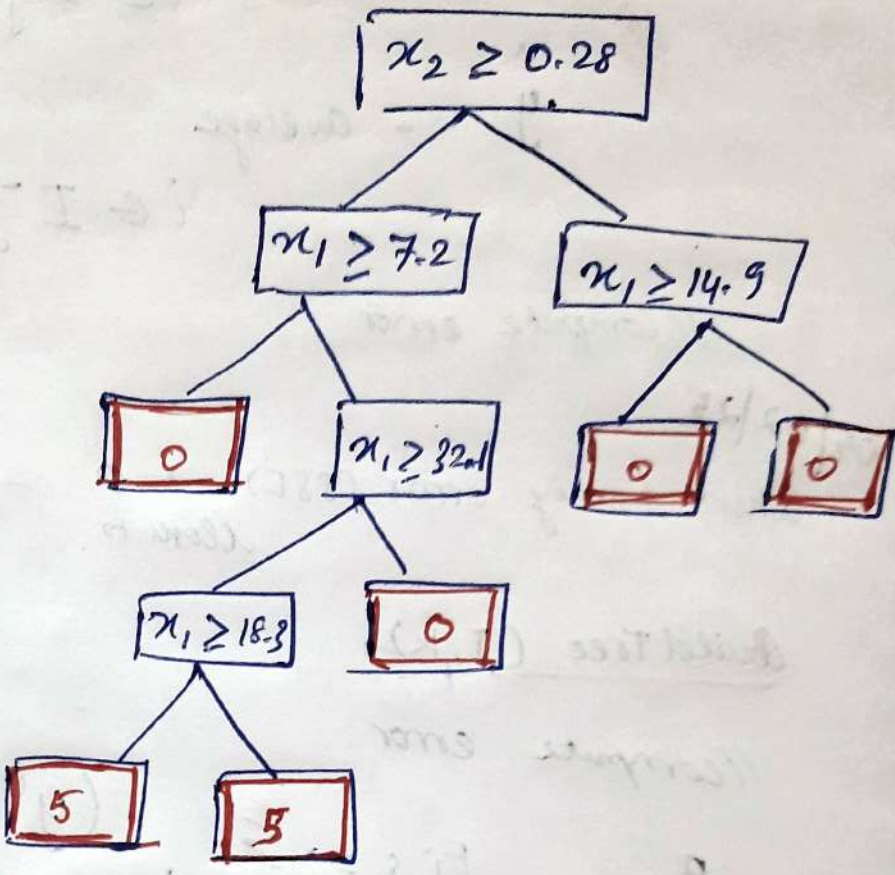


How to construct?
Heuristic based approach
Greedy

* It needn't be a binary tree..
* No imp/ order in the depth of the node.
* Pick the tree which has min train error

↓ 10 samples





The feature - finite set (can't be infinity)

The split - can have real values (continuous) $(-\infty$ to $+\infty)$

↓ problem

appx split value - ~~point~~ value b/w 2 datapoints

⇒ infinite possibilities split val. ⟶ why to eliminate the other possibilities

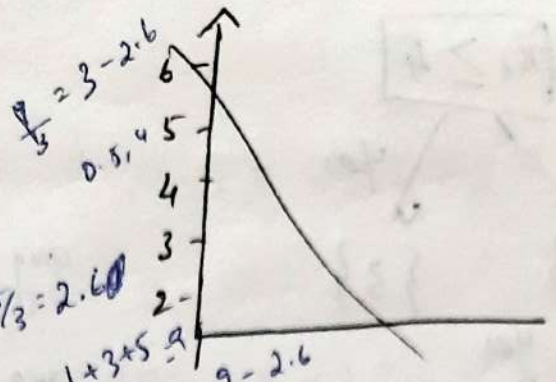
$I = \{1, 2, 3, \ldots, 11\}$

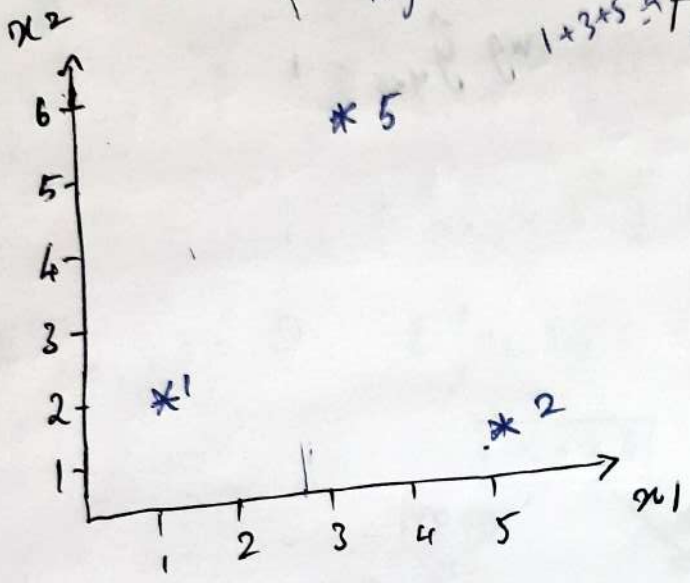
$|I|$ ⇒ element in the index

06/02/25

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

Avg: $8/3 = 2.6$





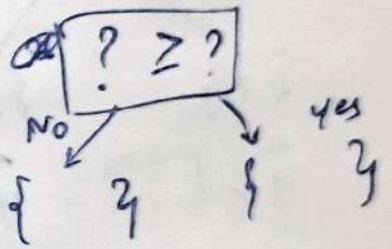
$*5$

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

$K = 2$

Construct 1 node

$\hat{y}$ = Avg $y^{(i)}$



midpoints $j = x_1, x_2$

↳ for x_1

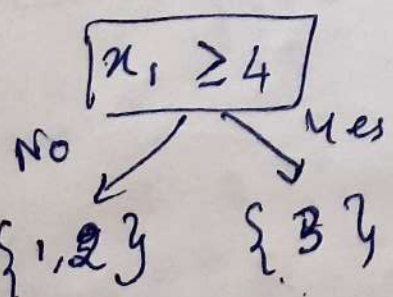
$S = \{2; 4\}$

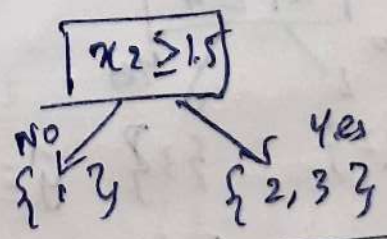
↳ for x_2

$S = \{1.5, 4\}$

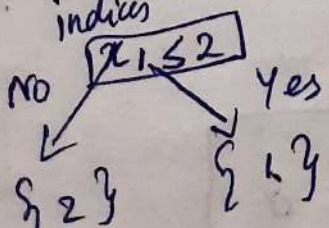
$S(4,2)$

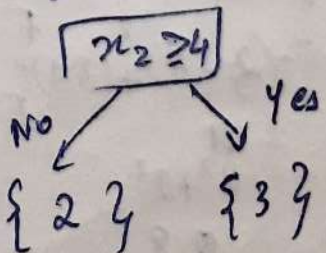
$S(1.5,4)$

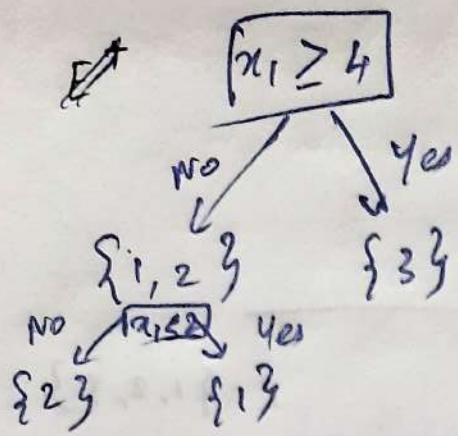




indices



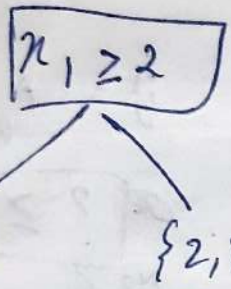




avg $\hat{y}_{No} = 1+5 = \frac{6}{2} = 3$

avg $\hat{y}_{Yes} = 1$

$E^+_{x_1,2} = 5 -$

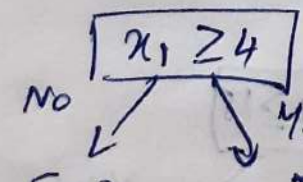


$\hat{y} = 1$ $\hat{y} = 3.5$ average.

$E^- = (1-1)^2$ $E^+ = (3.5-5)^2 + (3.5-2)^2$

$= 0.$ $E^+ = (1.5)^2 + (1.5)^2$

$\therefore E^+ = 2.25 + 2.25 = 4.5$ $\boxed{E_{x_1,2} = 4.5}$



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

$E^- = (2-5)^2 + (3-1)^2$ $E^+ = (2-2)^2$

$E^- = 2^2 + 2^2$ $E^+ = 0$

$E^- = 8$ $E^+ + E^- = 8$ $\boxed{x_{1,4} = 8}$

$x_2 \geq 1.5$

No → $\{1\}$ → $\hat{y} = 1$

Yes → $\{2, 3\}$ → $\hat{y} = 3.5$

$5+2 = \frac{7}{2} = 3.5$ avg

$E^- = \{(1-1)^2\}$

$E^- = 0$

$E^-_{x_2, 1.5} = 0$

$E^+_{x_2, 1.5} = \{(5-3.5)^2 + (2-3.5)^2\}$

$E^+_{x_2, 1.5} = \{2.25 + 2.25\}$

$E^+_{x_2, 1.5} = 4.5$ $= 4.5$

$x_2 \geq 4$

No → $\{2\}$ → $\hat{y} = 5$

$(5-5)^2 = 0$

Yes → $\{3\}$ → $\hat{y} = 2$

$E^+_{x_2, 4} = \{(2-2)^2\} = 0$

$x_2 \geq 1.5$

No → $\{3\}$ → $\hat{y} = 3$

$(3-3)^2$

$E^-_{x_2, 1.5} = 0$

Yes → $\{1, 2\}$ → $\hat{y} = \frac{1+5}{2} = 3$

$E^+_{x_2, 1.5} = \{(1-3)^2 + (5-3)^2\}$

$E^+_{x_2, 1.5} = \{(-2)^2 + (2)^2\}$

$E^+_{x_2, 1.5} = 8$

$E^+_{x_2, 1.5} + E^-_{x_2, 1.5} = 8$

$\hat{y} = 1.5$

$x_2 \geq 4$

No → $\{1, 3\}$ → $\hat{y} = 1$

Yes → $\{2\}$ → $\hat{y} = 5$

$E x_2^+, 4 = 0$ (0)

$E x_2^-, 4 = (1 - 1.5)^2 + (2 - 1.5)^2$

$= 0.25 + 0.25$

$= 0.5$

$E x_2^+, 4 + E x_2^-, 4 = 0.5$

j^*, s^*

$\min \{ E x_{1,2}, E x_{1,4}, E x_{2,1.5}, E x_{2,4} \}$

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

$\min = 0.5$

$j^* s^* \quad [x_2, 4]$

Build Tree

$x_2 \geq 4$

No → $\{1, 3\}$: $\vec{y} = 1.5$

$E^- y - \vec{y}$

$(1 - 1.5)^2 + (2 - 1.5)^2$

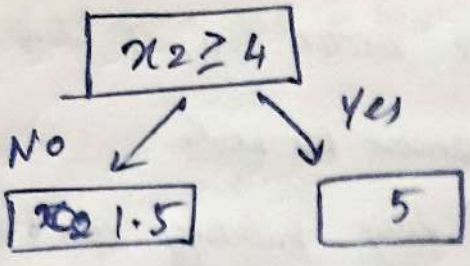
$0.25 + 0.25$

$= 0.5$

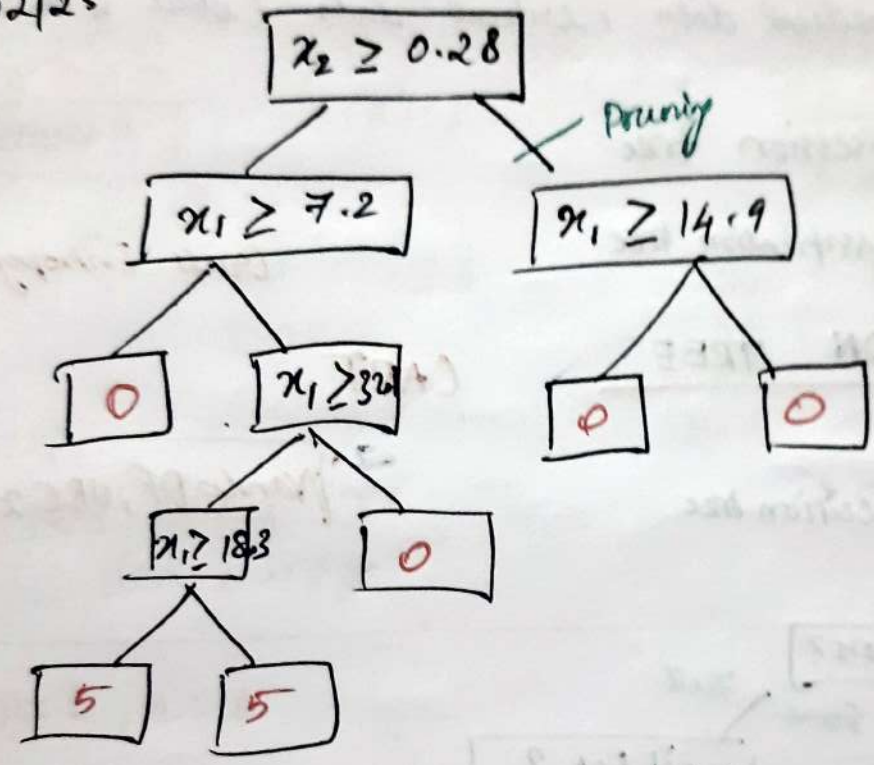
Yes → $\{2\}$: $\vec{y} = 5$

$E^+ y - \vec{y}$

$E^+ = 0$



07/02/25



for very high depth tree
↓
overfitting

If total no of leaves high, it is overfitting.

"Regularizing" to avoid overfitting

we need to add penalty term for the cost function.

(i) $C_\alpha(T) = \sum_{i=1}^{n} L(T(x^{(i)}, y^{(i)}) + \alpha|T|$ → No. of leaves

cost function of tree

Training loss

6 leaves in the previous

Now we reconstruct the tree by adding the penalty term.

α - controls how much the tree grows / shrinks.

(ii) Pruning → let the tree grow, and then remove the leaves. (eliminate the node) → remove 1 node → check wal accuracy

compute val. error → error (0,0). ideal.

if error is 0,0 → remove the node.

if val. error ↑ drop knockoff · don't touch

Both i) & ii) are complementary.

⇒ useful for medical data (critical data (start w this)

Tree → Regression tree
→ classification tree

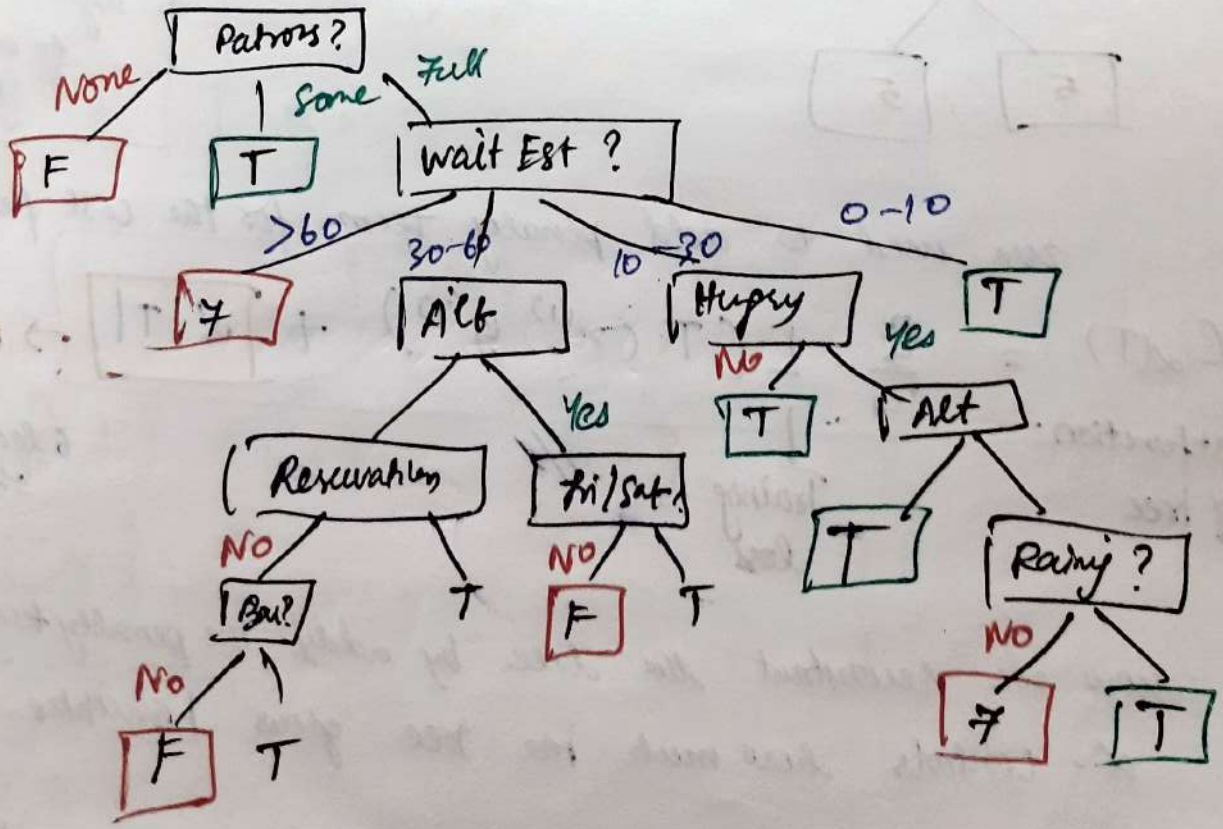
Cross Entropy Loss

CLASSIFICATION TREE

CART

A learned decision tree

Wando DF, UBC 2012

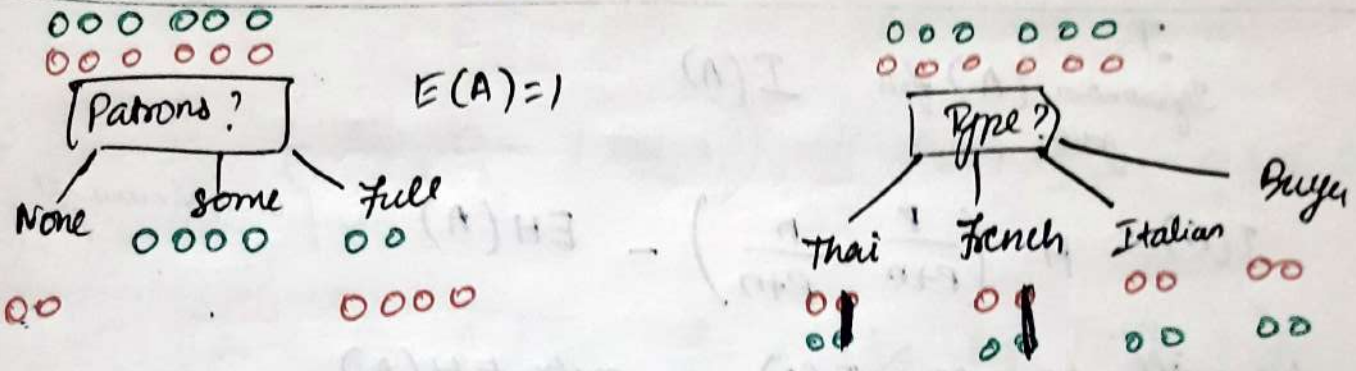


Why didn't we have SSE in logistic → $\hat{y}$ → label

Shannon's Info Theory

ENTROPY - randomness / impurity / chaos in data.

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



More homogenous, less entropy.

Eg: $p=6, n=6$

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

H ← [entropy rep by]

$\log_2$ → bits 0/1

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

(K-labels)

$$= -\frac{6}{12}\log_2\frac{6}{12} - \frac{6}{12}\log_2\frac{6}{12}$$

$$= -\tfrac{1}{2}\log_2\left(\tfrac{1}{2}\right) - \tfrac{1}{2}\log_2\left(\tfrac{1}{2}\right)$$

$$= -\tfrac{1}{2}(-1) - \tfrac{1}{2}(-1)$$

$$= \tfrac{1}{2} + \tfrac{1}{2} = 1$$

How do we find homogenity?

$E(A) = 1$

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

Expectation

I
Information gain (A) for A $I(A)$

$$I(A) = H\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - EH(A)$$ [entropy across all terms]

We will be max $I(A)$, make $EH(A)$ homogenous, entropy ↓, info gain ↑

11/02/25

1. Entropy
2. Information gain

[Expected entropy EH]

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

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

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

$$= 1$$

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

$$= \sum \left[\frac{p_1+n_1}{12} H\left(\frac{0}{2}, \frac{2}{2}\right) + \frac{p_2+n_2}{12} H\left(\frac{4}{4}, \frac{0}{4}\right) + \right.$$

for Patrons none, some & full

$$\left. \frac{p_3+n_3}{12} H\left(\frac{2}{6}, \frac{4}{6}\right) \right]$$

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

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

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

$$= 1/6 \left(\times \infty (0) \right) + 1/3 \left(-1 (0) \cdot (-\infty) \right) + 1/2$$

$$+ 1/2 \left(-1/3 \log_2 1/3 \left(-\frac{2}{3} \log_2 \left(-2/3 \right) \right) \right.$$

$$\rightarrow 1/2 \left(-1/3 \left(\log 1 - \log 3 \right) \right) \times \left(-2/3 \left(\log 2 - \log 3 \right) \right.$$

$$= -1/6 \left(0 - 1.585 \right) \times \left(-2/3 \left(1 - 1.585 \right) \right)$$

$$= 0.2641 \times \frac{1.17}{3}$$

$$= 0.2641 \times 0.39$$

expected entropy
0.455

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

$$= 1 - 0.455$$

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

Random forest (numerical example)

Age	Salary	Spend pattern	Buy Product
16	0	High	Yes
30	90,000	Low	No
31	90,000	High	No
55	5,00,000	Medium	No

B = 2

1. Create 2 bags B_1 & B_2

B_1 =

Age	Salary	Spend pattern	Buy Pdt
55	5,00,000	Med	No
30	90,000	Low	
31	90,0000	High	
55	5,00,000	Med	No

b_2 =

Age	Salary	Spend Pattern	Buy Pdt
16	0	High	Yes
16	0	High	Yes
31	90,000	High	No
16	0	High	Yes

Tree

T_1

if we pick 2 feature

① for features in (Age, Salary)

for feature in (Age, Salary)

{
  for each value in features

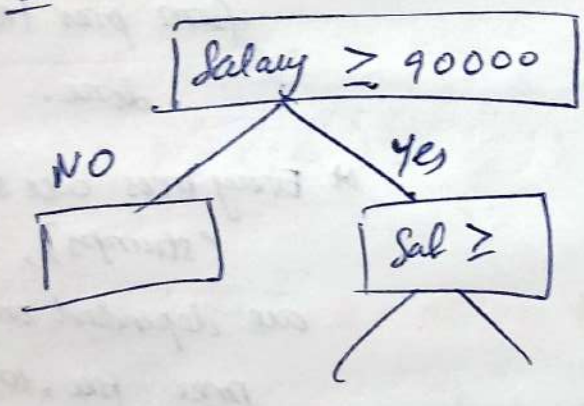
    IG:(IG.value) . append

IG for age as feat

// assume IG = 1.2 IG of Age = 1.2

IG of Salary = 1.8 → IG of Salary is high

T1



→ select another 2 random features m=2

Salary, Spend pattern

IG = 1 assume

IG = −0.2

Then construct another Tree T_2,

then aggregate the features

pass test val → 60.

At each step instead of considering all features, we select a random subset of features usually $m = \sqrt{p}$ features or $m = \sqrt{d}$.

x_1	x_2	y	r	$\hat{y}$	new r	...
1	2	7	7	7	0	...
2	3	4	4	3.5	0.5	...
5	8	1	1	10	-10	...

example:

Dataset

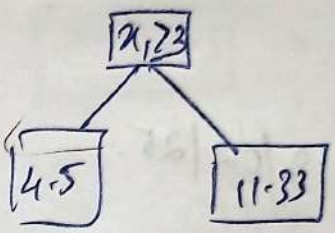
x_1	y	r
1	3	3
2	6	6
3	8	8
4	11	11
5	15	15

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

i)

① $r = y$

② a) fit a model



#1,2 #3,4,5

b; c) Compute r

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

assume $\lambda = 0.1$

$3 + 0.1 \times 4.5$

$= 3.45$

$F(X,0) + \lambda_b (X;0)$

updating

$6 + 0.1 \times 4.5$

$8 * 0.1$

x_1	r	$\hat{y}$	new_r
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

$y - \hat{y}_{pred} = new\_r$

~~8~~ [

$3 - (3 + 0.1 \times$

ii) Next

$x_1 \geq 2.5$

0.858

all samples on left side

Inference

$x_1 = 5$ $\quad \sum_{i=1}^{B} \lambda f_b(x)$

$\lambda_{01} f_1(5) + \lambda f_2(5) + \lambda f_3(5) + \cdots \lambda f_b(5)$

$= 0.8$

https://www.kaopeich.com/data414/notes/17_ensemble_methods_notes.pdf

18/02/25

Boosted Regression Tree (~~Ada Boost Regression~~ Gradient Boosting)

Boosted Classification Tree (Ada Boost classification)

Analogy of Boosting - Internet

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

$f_0(x) = 0$

f_1 → weak learner, f_2 → weak learner

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

2. For iteration 1 to B

{ a) fit a model $f_b(X; \theta) \rightarrow r$

b) update the final model by adding shrinkage version

$F(X; \theta) = f(X, \theta) + \lambda_1 f_b(X; \theta)$

c) update the residuals

$r^{(i)} = r^{(i)} - \lambda f_b(X; \theta)$

3. Final model

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

Numerical of:-

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 — $r = y$

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

Step a:

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

$x_1 \geq 2.5$? (split)

(4+5)/2 = 4.5 (#1, 2)

(6+7)/2 = 6.5 (#3, 4)

Step b: Update the model ($\lambda = 0.1$ assume)

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

Step c: Update residuals

$$r_i^{(1)} = r^{(1)} - \lambda f_1(x^{(1)})$$
$$r^{(2)} = r^{(2)} - \lambda f_1(x^{(2)})$$
$$r^{(3)} = r^{(3)} - \lambda f_1(x^{(3)})$$
$$r^{(4)} = r^{(4)} - \lambda f_1(x^{(4)})$$

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

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

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

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

residuals are reduced

Iteration 2

Step a:- assume best split

Fit a model $f_2(x) \rightarrow r$

$x_1 \geq 2.5$?

(3.55+4.55)/2 = 4.05

(5.35+6.35)/2 = 5.85

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

Step b:- Update the model

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

Step c:- Update residuals

$$r_0^{(1)} = r^{(1)} - \lambda f_2(x^{(1)})$$
$$r^{(2)} = r^{(2)} - \lambda f_2(x^{(2)})$$
$$r^{(3)} = r^{(3)} - \lambda f_2(x^{(3)})$$
$$r^{(4)} = r^{(4)} - \lambda f_2(x^{(4)})$$

$r_1 = 3.55 - (0.1 \times 4.05) = 3.145$

$r_2 = 4.55 - (0.1 \times 4.05) = 4.145$

$r_3 = 5.35 - (0.1 \times 5.85) = 4.765$

$r_4 = 6.35 - (0. \times 5.85) = 5.765$

} residuals have reduced even more.

$$\mathcal{F}(X;\theta) = f_0(X) + \lambda f_1(X) + \lambda[f_2(X) + \lambda f_3(X) \ldots + \lambda f_b(X)$$

Whenever cont data only → to measure effectiveness → MSE, SSE residuals

↓

but in class^n it doesn't work well (residual update approach)

↓

ADABOOST class^n.

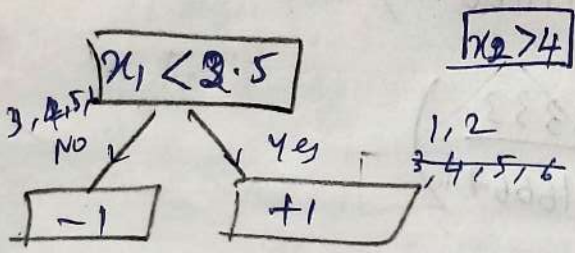
ADABOOST Classif^n :.

Setup ⇒ Binary class^n setting

⇒ Multiple weak learners / models with target $\in \{-1, +1\}$

Adaboost classⁿ

example: $w(n) = 1$



If 1

#	x_1	x_2	y	w
1	1	2	+1	1
2	2	3	+1	1
3	3	3	+1	1
4	4	5	−1	1
5	5	5	−1	1
6	6	6	−1	1

$$\epsilon^{(i)} = \frac{\sum_{i=1}^{N} w^{(i)} I\{y^{(i)} \neq f_b(x^{(i)}; \theta_b)\}}{\sum_{i=1}^{N} w^{(i)}}$$

$$\epsilon_1^{(i)} = \frac{w^{(i)} \times 0}{} = \frac{0}{6} = 0$$

→ Indicator fn

$$\epsilon^{(2)} = \frac{w^{(2)} \{I\{y^{(i)} \neq f_b(x^{(i)}; \theta_b)\}}{\sum_{i=1}^{N} w^{(2)}}$$

$$= 1 \times 0 = 0/6 = 0 \quad 2..$$

$$\epsilon^{(3)} = \frac{w^{(3)} I\{y^{(3)} \neq f_b(x^{(3)}; \theta_b)\}}{\sum_{i=1}^{N} (w^{(i)})}$$

$$= w^{(3)} \downarrow 1$$

$$= 1 \times 1 = 1/6 = 0.16667$$

$$\epsilon^{(4)} = \frac{w^{(4)} I\{y^{(4)} = f_b(x^4; \theta_b)\}}{6}$$

$$= 0/6 = 0$$

$$\epsilon^{(5)} = 0 \qquad \epsilon^{(6)} = 0$$

$$\lambda_b = \frac{1}{2} \log\left(\frac{1 - 0.16667}{0.16667}\right)$$

$$= \frac{1}{2} \log\left(\frac{0.8333}{0.16667}\right)$$

$$\lambda_b = \frac{1}{2} \log(0.8333) - \log(0.16667)$$

$$\lambda_b = 0.8047$$

Obtaining weights

$$w^{(1)} = w^{(1)} \times e^{-0.8047} = 0.4472$$

$$w^{(2)} = w^{(2)} \times e^{-0.8047} = 0.4472$$

$$w^{(3)} = w^{(3)} \times e^{0.8047} = 2.2360$$

$$w^{(4)} = w^{(4)} \times e^{-0.8047} = 0.4472$$

$$w^{(5)} = w^{(5)} \times e^{-0.8047} = 0.4472$$

$$w^{(6)} = w^{(6)} \times e^{-0.8047} = 0.4472$$

x_1	x_2	y	w

21/02/25 absent

Gradient Boosting Algorithm (Regression)

$$h = \frac{1}{2}(y - \hat{y})^2$$

$$\frac{\partial L}{\partial \hat{y}} = \hat{y} - y$$

$$\boxed{\frac{-\partial h}{\partial \hat{y}} = y - \hat{y}}$$

→ residual / Pseudo-residuals

To minimize the loss we need to move in the dir of -ve grad

or $h(y, p) = -[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$

↓ Cross-Entropy Loss fn.

$$\frac{\partial L}{\partial p} = -\left[\frac{y}{p} + \left(\frac{1-y}{1-p}\right)(-1)\right]$$

$$\frac{\partial L}{\partial p} = \frac{p - y}{p(1-p)}$$

→ 1st derivative

→ it's weighing factor we can remove it in 2nd derivative

{ p – sigmoid func^n $p = \frac{1}{1+e^{-z}}$ }

$$\boxed{\frac{-\partial L}{\partial p} = \frac{y - p}{p(1-p)} \simeq y - p}$$

$$\frac{\partial^2 L}{\partial p^2} = \frac{\partial}{\partial p^2}(p - y)$$

$$\boxed{\frac{\partial^2 L}{\partial p^2} = p(1-p)}$$

→ since 2nd derivative of sigmoid fun^n = $g(z)(1 - g(z))$

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 +ve class prob

$P(+ve) = 3/6 = 0.5$

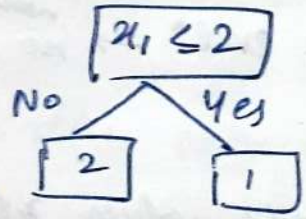
$$F_0 = \log\left(\frac{p}{1-p}\right)$$

$$F_0 = \log\left(\frac{0.5}{1-0.5}\right) = \log(1) = 0$$ → log odds

Step 2: Compute residuals $\quad r_i = y_i - p_i$

Sample	True - y	initial $p = 0.5$	Residual $r = y - p$
1	1	0.5	0.5
2	0	:	-0.5
3	1	:	0.5
4	1	:	0.5
5	0	:	-0.5
6	0	:	-0.5

Step 3:- Fit a decision stump (model - classifier) ⇒ Apply D. Tree Algo



Samples $\{3,4\}$ $\quad \{1,2,5,6\}$

* We can't take avg of 'y' val, because prob values are not additive in nature

$$W = \frac{\Sigma(y-p)}{\Sigma p(1-p)+\lambda} \rightarrow \text{Log(odds) adjustment}$$

$$W_{(wt)} = \frac{\text{Sum of residuals}}{\text{Sum [Prob} \times (1-\text{Prob})]} \rightarrow 2^{nd} \text{ der of } L$$

where λ = regularizn constant.

∴ We use 'W' to fill the leaves

$$W(\text{right side}) = \frac{-0.5-0.5+0.5-0.5}{(0.5\times0.5)+(0.5\times0.5)+(0.5\times0.5)+(0.5\times0.5)}$$

$$W(\text{right}) = \frac{-1}{1} = -1$$

Step 4:

$W(\text{left}) = 2 \longrightarrow$ Leaves

Summation of log odds → convert it into prob. value → if its > 0.5 then +ve class

$F = F_0 + \lambda F_1 + \lambda F_2 + \ldots$

$0.1 \rightarrow$ log(odds) value.

We have constructed a linear boundary in high dim space.

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

monomial of x with degree 3.

we are capturing the relation b/w x_1, x_2 ...

if $d = 1000$

$\downarrow$

$P \simeq O(d^3)$

$\hookrightarrow$ 1 billion features

$\downarrow$

1 billion O^T

$\Downarrow$

Problem:

meaning 1M times slower than what we do in 1000 dim space.

for every step update in grad desnt, we need to compute IB times.

Kernel

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

$\downarrow$

Kernel fun^n

transform $x \rightarrow \phi x$ into high dim space

$y \rightarrow \phi x$

$x = (x_1, x_2, x_3)$

$y = (y_1, y_2, y_3)$

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

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

$y = [4, 5, 6]$

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

$\downarrow$

dot pt b/w ϕx & ϕy

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

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

$= 1024$

$[1\ 2\ 3\ 2\ 4\ 6\ 3\ 6\ 9] \begin{bmatrix} 16 \\ 20 \\ 24 \\ 20 \\ 25 \\ 30 \\ 24 \\ 30 \\ 36 \end{bmatrix}$

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

$[1\ 2\ 3] \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$ $4 + 10 + 18 = (32)^2$

$= 1024$

Types of Kernels

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

$$= \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} K(x^{(j)}, x^{(i)}) \right)$$

Testing Phase

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

One test instance

$$= \sum_{i=1}^{n} \beta_i \phi(x^{(i)}) \phi(x^{(t)})$$ - dot pdt in transformed space.

$$y\_pred = \sum_{i=1}^{n} \beta_i K(x^{(i)}, x^{(t)})$$ → lower dim space.

$x^{(i)}$ ↓ training $x^{(t)}$ ↓ test sample

Example

x_1	x_2	y
1	2	0
3	4	0

$n = 2$

$$\phi(x) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_1 x_2 \end{bmatrix}$$ 3-D

$X = [x_1 \; x_2]^T$ $X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ 2-D

$X \in \mathbb{R}$

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

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

$\theta = [\theta_1, \theta_2, \theta_3]^T$

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

3dim 2din 3 dim

$$\theta = \begin{bmatrix} ? \\ ? \\ ? \end{bmatrix}_{3 \times 1}$$

Try by me $\beta_i = [\beta_1 \ \beta_2]_{1\times 2}$ $\phi(x) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_1 x_2 \end{bmatrix}_{3\times 1}$

~~$\theta = \beta_2 x_1^2$~~ $\phi(x) * \beta_i$

$(\beta_i)^T \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}_{2\times 1}$ $(\phi(x))^T = [x_1^2 \ x_2^2 \ x_1 x_2]$

$\begin{bmatrix} \beta_1 x_1^2 + \beta_1 x_2^2 + \beta_1 (x_1 x_2) \\ \beta_2 x_1^2 + \beta_2 x_2^2 + \beta_2 (x_1 x_2) \end{bmatrix}$ dim is wrong

---

$\phi(x^{(1)}) = \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix}_{3\times 1}$ kernel trick $\theta = \beta_1 \phi(x^{(1)}) + \beta_2 \phi(x^{(2)})$

$\phi(x^{(2)}) = \begin{bmatrix} 9 \\ 16 \\ 12 \end{bmatrix}_{3\times 1}$ $\theta = \begin{bmatrix} \beta_1 + 9\beta_2 \\ 4\beta_1 + 16\beta_2 \\ 2\beta_1 + 12\beta_2 \end{bmatrix}$

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

$y\_pred = \sum_{i=1}^{n} \beta_i K(x^{(i)}, x^{(t)})$ → here we need all the training data, unlike we need only θs for LRs. (Computationally expensive ~ 1 billion samples)

Three lines / hyperplanes

$2x_1 + 3x_2 - 5 = 0$

$-x_1 + 4x_2 + 7 = 0$

$5x_1 - 12x_2 + 10 = 0$

Dataset

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

1st line, 1st point

$$d_1 = \frac{|\beta_0 + \beta_1 x_1 + \beta_2 x_2|}{\sqrt{\beta_1^2 + \beta_2^2}}$$

$$= \frac{|-5 + 2\times 3 + 3\times 4|}{\sqrt{2^2+3^2}} = \frac{6+12-5}{\sqrt{4+9}} = \frac{13}{\sqrt{13}} = 3.6$$

2nd line 2nd point

$$d = \frac{|7 + (-1)\times 2 + 4\times 3|}{\sqrt{(-1)^2+(4)^2}} = \frac{|7-2+12|}{\sqrt{1^2+4^2}} = \frac{17}{\sqrt{17}} = \sqrt{17} = 4.123$$

3rd line

$$d = \frac{|10 + (5)(1) + (-12)(1)|}{\sqrt{5^2+(-12)^2}} = \frac{10+5-12}{\sqrt{25+144}} = \frac{3}{\sqrt{169}} = 0.2307$$

3rd line - min value

1st line 2nd point

(1) d_2 $= \frac{|\beta_0 + \beta_1 x_1 + \beta_2 x_2|}{\sqrt{\beta_1^2 + \beta_2^2}} = \frac{-5 + 2\times 2 + 3\times 2}{\sqrt{2^2 + 3^2}}$

$= \frac{|5 + 4 + 6|}{\sqrt{13}} = \frac{5}{\sqrt{4}}$

$= 1.25$ 2.21

(1) d_3 $\frac{-5 + 1\times(2) + (-1)\times 3}{\sqrt{13}} = \frac{-5+2-3}{\sqrt{4}}$

$= \frac{-8+2}{4} = \frac{|-6|}{4}$

$= 1.86$

(1) d_4

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

$= \frac{-6}{\sqrt{13}} = 1.86$

1st hyperplane margin
Min = 1.664

min (3.25, 1.25, -1.5, -1.5) ⇒ 1.25 d_2

2nd line

d_1 = $-x_1 + 4x_2 + 7 = 0$

$\frac{|7 + (-1)\times 3 + 4(4)|}{\sqrt{(-1)^2 + (4^2)}} = \frac{7 - 3 + 16}{\sqrt{16+1}}$ 4.8

$= \frac{20}{\sqrt{17}}$

$$d_2 = \frac{|7 + (-1) \times (2) + 4 \times (3)|}{\sqrt{15}} = \frac{7 - 2 + 12}{\sqrt{17}}$$

$$= \frac{17}{\sqrt{17}}$$

$$= 4.12$$

$$d_3 = \frac{|7 + (-1) \times (1) + (4) \times (-1)|}{\sqrt{15}} = \frac{7 - 4 - 1}{\sqrt{17}}$$

$$\frac{7 - 5}{\sqrt{17}}$$

$$= 0.48$$

$$d_4 = \frac{|7 + (-1) \times -2 + 4 \times 1|}{\sqrt{15}} = \frac{7 + 2 + 4}{\sqrt{17}} = \frac{13}{\sqrt{17}}$$

$$= 3.15$$

$$\min(d_1 d_2 d_3 d_4) = 0.48$$

∴ 2nd hyperplane margin = 0.48

Revise:

1. Projection
2. Norm of a vector
3. Vector addn
4. Orthogonality
5. Subspaces

Line 3:

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

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

$$d_1 = \frac{|10 + (5 \times 3) + (-12 \times 4)|}{\sqrt{5^2 + 12^2}}$$

$$= \frac{|10 + 15 - 48|}{\sqrt{169}} = \frac{23}{13} = 1.76$$

$$d_2 = \frac{|10 + (5 \times 2) + (-12 \times 3)|}{\sqrt{169}} = \frac{|10 + 10 - 36|}{13} = \frac{16}{13}$$

$$= 1.23$$

$$d_3 = \frac{|10 + 5 \times 1 + (-12 \times -1)|}{\sqrt{169}} = 2.07$$

$$d_4 = \frac{|10 + 5 \times (-2) + -12 \times 1|}{\sqrt{169}} = 0.92$$

3rd hyperplane margin = 0.92

Of all 3 margins M_1, M_2, M_3 M-1 is max, so the 1st hyperplane is optimum.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{MSE}$$

$$MAE \rightarrow \text{absol } \epsilon = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

21/03/25

Threshold assumed = 0.5

Observation	Actual label	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

Calculate 1) Accuracy
2) Precision
3) Recall / sensitivity / True +ve Rate
4) Specificity
5) False +ve rate (1 - specificity)
6) F1
7) ROC curve

Build confusion matrix

Threshold
0.7
0.5
0.4
0.2

	Orig Pos	Orig Neg
good +ve	4	3
pred -ve	1	2

orig pos = 5

orig neg = 5

1) Accuracy

(TP+TN)/Total

= (4+2)/10 = 6/10 = 0.60.

2) Precision

TP/(TP+FP)

= 4/(4+3) = 4/7 = 0.57

3) Recall

TP/(FP+FN)

= 4/(4+1) = 4/5 = 0.80

4) Specificity

TN/(TN+FP)

2/(2+3) = 2/5 = 0.40

5) False positive rate

= 1 − Specificity = 1 − 0.40

= 0.60

6) F2 score

$$F_1 = 2 \cdot \frac{0.57 \times 0.80}{0.57 + 0.80} = \frac{0.912}{1.37}$$

= 0.665

Obs	Act	Pred score	Pred Label (0.7)	0.5	0.4	0.2
	1	0.85	1	1	1	1
	0	0.60	0	1 FP	1 FP	1 FP
	1	0.70	1	1	1	1
	1	0.40	0 FN	0 FN	1	1
	0	0.55	0	1 FP	1 FP	1 FP
	1	0.50	0 FN	1	1	1
	0	0.65	0	1 FP	1 FP	1 FP
	0	0.35	0	0	0	1 FP
	1	0.60	0 FN	1	1	1
	0	0.20	0	0	0	1 FP

Confusion matrix for 0.7

	OP	ON
PP	2	0
PN	3	5

Confusion matrix for 0.4

	OP	ON
PP	5	3
PN	0	2

Confusion matrix for 0.2

	OP	ON
PP	5	5
PN	0	0

Conf matrix for 0.5

4	3
1	2

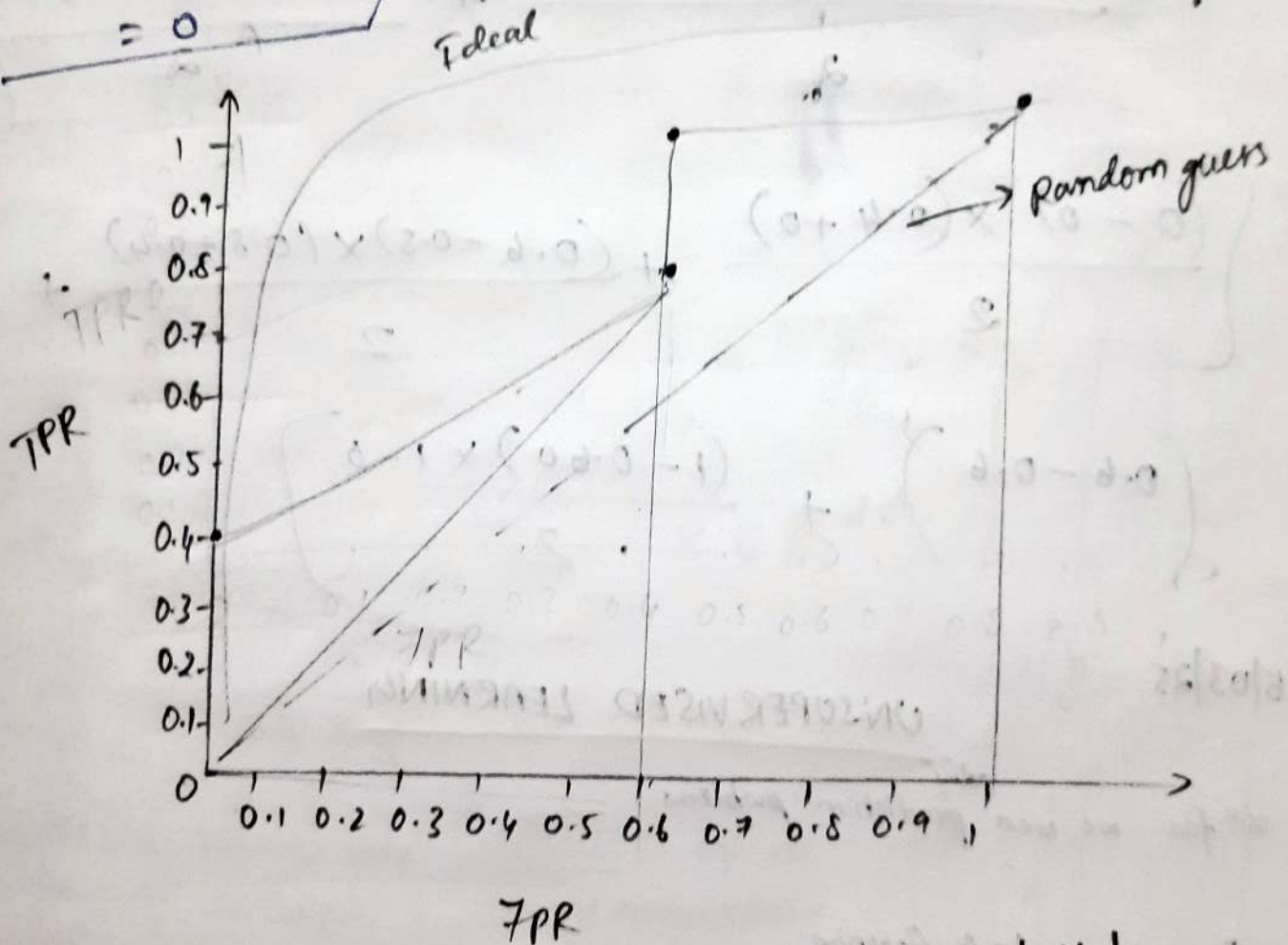
TPR = Sensitivity

FPR = 1 − Specificity

$TPR = TP / (TP + FN)$

$FPR = 1 - \left(\frac{TN}{TN + FP} \right)$

for 0.7	for 0.5	for 0.4	for 0.2
TPR = $\frac{2}{2+3}$ = $\frac{2}{5}$ = 0.4	TPR = 0.8	TPR = $\frac{5}{5+0}$ = 1	TPR = $\frac{5}{5+0}$ = 1
TPR = $1-\left(\frac{5}{5+0}\right)$ = 1 − 1 = 0	1 − 0.40 = 0.60	$1-\frac{2}{2+3}$ = $1-\frac{2}{5}$ = 0.60	TPR = $1-\frac{0}{0+5}$ = 1



$\frac{1}{2}(a+b)h$

$\frac{1}{2}(0.4+0.8)\times 0.6$

$1-0.6 = 0.4 \times 1$

$= 0.4$ sq units

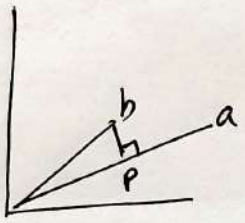
$$AUC = \sum_{i=1}^{4} \frac{(FPR_i - FPR_{i-1}) \times (TPR_i + TPR_{i-1})}{2}$$

$$= \left[\frac{(FPR_1 - FPR_0) \times (TPR_1 + TPR_0)}{2} + \frac{(FPR_2 - FPR_1) \times (TPR_2 + TPR_1)}{2} + \right.$$

$$\left. \frac{(FPR_3 - FPR_2) \times (TPR_3 + TPR_2)}{2} + \frac{(FPR_4 - FPR_3) \times (TPR_4 + TPR_3)}{2} \right]$$

$$= \left\{ \frac{(0-0) \times (0.4+0)}{2} + \frac{(0.6-0.5) \times (0.8+0.4)}{2} + \right.$$

$$\left. (0.6-0.6) + \frac{(1-0.60) \times 1-0}{2} \right]$$



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

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

$$a^T = \begin{bmatrix} 1 & 2 \end{bmatrix}_{1\times 2} \cdot \begin{bmatrix} 3 \\ 4 \end{bmatrix}_{2\times 1} = \begin{bmatrix} 3+8 \end{bmatrix} = 24$$

a $\quad$ a^T

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

b $\quad$ $1+4$

$$X = \frac{24}{5}$$

$$p = Xa$$

$$= \frac{24}{5} \times \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad p = \begin{bmatrix} 24/5 \\ 24/5 \times 2 \end{bmatrix}$$

$[A^T A]^{-1} \rightarrow (24)^{-1}$

$p =$

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

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

$$= \frac{[1+4]}{[5]} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \frac{5 \begin{bmatrix} 3 \\ 4 \end{bmatrix}}{5}$$

$$= \begin{bmatrix} 1+2 \\ 2+4 \end{bmatrix} \frac{}{5}$$

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

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

PCA

u to be a subspace of unit length (a)

$\vec{x}$

projected point of $\vec{x}$ = proj matrix of $(u) \cdot \vec{x}$

$$= \frac{u u^T}{u^T u} \cdot \vec{x}$$

($u^T u$: unit vector)

$$= (u u^T)\vec{x}$$

$$= (\vec{x} \cdot u)\, u^T$$

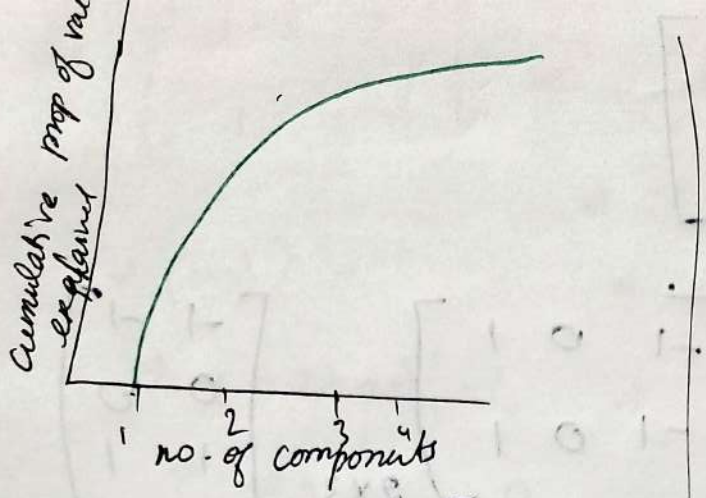
Find u such that

$$= \underset{u}{\text{argmax}} \frac{1}{n} \sum_{i=1}^{n} \left\| \text{proj}(u) \cdot x^{(i)} \right\|_2^2 = (\vec{x} \cdot u)^2$$

Find u which maximizes the variance (variance)

$$= \boxed{\underset{u}{\text{argmax}} \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} u)^2}$$

The norm of a scalar multiplied by a unit vector is equal to the square of



Last step:

$$[n \times d] \times PC = [n \times k]$$

this dataset is used to train ML model.

eg:

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

$(d=2)$ $(k=1)$

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

① Std of each column

$$\frac{x-\bar{x}}{\sigma} \text{ or } \frac{x-\mu}{\sigma} \qquad X_{std} =$$

mean of 1st col = 4 μ_1

mean of 2nd col = 200 μ_2

$$\sigma^2 = \frac{1}{n}\sum(x_i-\mu)^2$$

$$= \frac{1}{3-1}\sum(2-4)^2 + (4-4)^2 + (6-4)^2$$

$$\frac{1}{3}[2^2+0+2^2] \Rightarrow 4+4 = \frac{8}{2} = 4$$

$$\sigma^2 = 2.67 \qquad \boxed{\sigma = 2}$$

$\sigma_2 \Rightarrow (100-100)^2 + (200-100)^2 + (300-100)^2$

$$= \frac{1}{2}\left(0 + (100)^2 + (200)^2\right)$$

$\sigma^2 = 2$

$\sigma_2 = 100$

1st column std

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

$$\frac{4-4}{2} = 0$$

$$\frac{6-4}{2} = 1$$

2nd col std

$$\Rightarrow \frac{100-200}{100} = \frac{-100}{100}$$

$$\Rightarrow -1$$

$$\Rightarrow \frac{200-200}{100} \Rightarrow 0$$

$$\Rightarrow \frac{300-200}{100} \Rightarrow 1.$$

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

Compute Covariance

② $\frac{1}{n-1} X_{std}^T X_{std} = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}_{2\times 3} \begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}_{3\times 2}$

$$\Rightarrow \begin{bmatrix} (1+0+1) & (1+0+1) \\ (1+0+1) & (1+0+1) \end{bmatrix}_{2\times 2}$$

$$\Rightarrow \frac{1}{n-1} \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}_{2\times 2} \Rightarrow \frac{1}{2} \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

③ Compute eigen values & eigenvectors of the covariance matrix

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

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

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

$$\Rightarrow \begin{bmatrix} 1-\lambda & 1 \\ 1 & 1-\lambda \end{bmatrix} \rightarrow ①$$

④ Decide the needed % of variance explained & choose 'k' principal components.

⑤ Transform X → Z = into new space of principal components

$A \times P_c = Z$

⑥ Use this Z to train a model.

$\det \begin{bmatrix} 1-\lambda & 1 \\ 1 & 1-\lambda \end{bmatrix} \Rightarrow (1-\lambda)(1-\lambda) - (1 \times 1)$

$\Rightarrow (1-\lambda)^2 - 1$

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

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

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

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

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

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

$\Rightarrow$

To find eigenvectors for $\lambda_2 = 2$

$(A - \lambda I)V = 0$ subs λ in ①

[if the λ is higher eg: 4 & 2 choose 4 ↓ the correspondig eigen vector will capture ↑ variance]

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

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

$-1x + y = 0$ (or) $1x - 1y = 0$

$\boxed{x = y}$ $\boxed{x = y}$

$V_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

x, y

For $\lambda_1 = 0$.

$(A - 0I) \Rightarrow \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

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

$x + y = 0$

$x = -y$

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

28/03/25

normalize:

$$\frac{1}{\sqrt{1^2+1^2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \rightarrow \text{vector}$$

↓ magnitude

$$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \rightarrow \text{to make it as unit vector}$$

$$\begin{bmatrix} \frac{1}{\sqrt{1^2+1^2}} & \frac{1}{\sqrt{2^2+2^2}} \\ \left(\begin{matrix}1\\1\end{matrix}\right) & \left(\begin{matrix}1\\1\end{matrix}\right) \end{bmatrix}$$

(4) We only pick $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ (v_1).

(5) $Z_{pca} = Z_{(std)} \cdot v_1$

$$Z_{pca} = \underset{3\times 2}{\begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}} \times \underset{2\times 1}{\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}}$$

$$= \begin{bmatrix} -1 \times \frac{1}{\sqrt{2}} + (-1 \times \frac{1}{\sqrt{2}}) \\ 0 + 0 \\ \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{2}{\sqrt{2}} \\ 0 \\ \frac{2}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} -\sqrt{2} \\ 0 \\ \sqrt{2} \end{bmatrix}_{3\times 1}$$

example K=2

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

1) $C_1 = \{2, 3, 4, 6\} \rightarrow 1+2+8+9 = 20/4 = 5$

$C_2 = \{1, 5\} \rightarrow \begin{bmatrix} 5 \\ 5 \end{bmatrix}$

Centroid of C1 = 3.75

Centroid of C2 = 3

$C_1 \rightarrow \begin{bmatrix} 4.5 \\ 5 \end{bmatrix}$

(2) - 3.75 = -1.75

(2) - 3 = 1 (2) belongs to C_2.

for (3) - 3.75 → 0.75

3 - 3 = 0. ✓ (3) belongs to C2

for (4)

4 - 3.75 = 0.25 ✓

4 - 3 = 1 (4) belongs to C_1

for (6)

6 - 3.75 = 2.25 ✓

6 - 3 = 3 (6) belongs to C_1

Centroid

$C_1 = \begin{bmatrix} 5 \\ 5 \end{bmatrix}$ $C_2 = \begin{bmatrix} 4.5 \\ 5 \end{bmatrix}$

$d_{11} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $\begin{matrix} 5-1 \\ 5-1 \end{matrix}$ $\begin{matrix} 4.5-1 \\ \end{matrix}$ $d_{11} \rightarrow C_2 = \begin{bmatrix} 3.5 \\ 4 \end{bmatrix} =$

$d_{11} \rightarrow C_1 = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$ $\sqrt{32} = 5.65$ 12.25

$d_{11} \rightarrow c_1 = 5.65$ $\quad d_{11} \rightarrow c_2 = 5.31$ ✓ min

d_{11} shd be in cluster (2)

$d_{12} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$d_{12} \rightarrow c_2 \begin{bmatrix} 4 \\ 3 \end{bmatrix}$ $\quad d_{12} \rightarrow c_2 \begin{bmatrix} 3.5 \\ 3 \end{bmatrix}$

$\sqrt{4^2+3^2}$ $\quad \sqrt{3.5^2+3^2}$

$= 5$ $\quad = 4.609$ min

d_{12} shd bein cluster (2)

$d_{13} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$

$d_{13} \rightarrow c_1 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$ $\quad d_{13} \rightarrow c_2 = \begin{bmatrix} 2.5 \\ 3 \end{bmatrix}$

$= \sqrt{3^2+3^2}$ $\quad = \sqrt{(2.5)^2+3^2}$

$= \sqrt{18}$ $\quad = \sqrt{15.25}$

$= 4.24$ $\quad = 3.905$ ✓ min

d_{13} shd be in cluster (2)

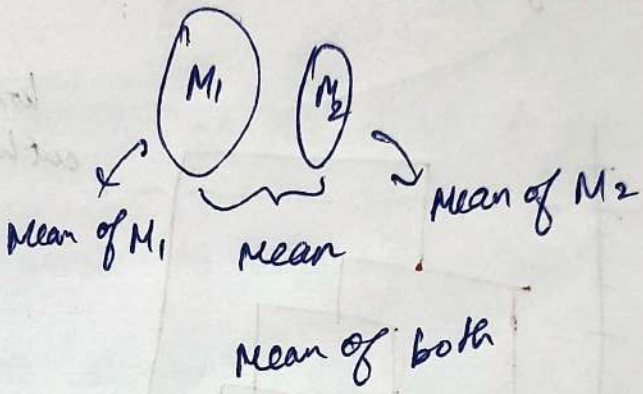
$d_{14} = \begin{bmatrix} 8 \\ 8 \end{bmatrix}$ $\quad d_{14} \rightarrow c_1 = \begin{bmatrix} -3 \\ -3 \end{bmatrix}$ $\quad d_{14} \rightarrow c_2 \Big[$

Complete linkage

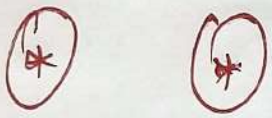
min → max

choosing min / max → comes from the data.

Average linkage



Centroid linkage



Cluster centroid c_1

Cluster centroid c_2

Dissimilarity b/w dot pdt

dot pdt is high → similar

1 − dot pdt → will give dissimilarity

Mostly single & complete link

Hierarchial

Agglomerative → building branches from branches

Example:

Data

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

Single linkage

Euclidian distance

5-observations $\Rightarrow {}^5C_2 \Rightarrow \frac{5!}{3!\,2!} \Rightarrow \frac{20}{2} = 10.$

	A	B	C	D	E
A	0	1	3.6	5	6.4
B	1	0	2.8	4.24	5.65
C	3.6	2.8	0	1.41	2.82
D	5	4.24	1.41	0	1.41
E	6.4	5.65	2.82	1.41	0

Fill up the values.

1st row:

$A = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $B = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Rightarrow \sqrt{1^2+0^2} = 1$ (DBA)

$A = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $B = \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \Rightarrow \sqrt{9+4} \Rightarrow \sqrt{13} = 3.605$

$A = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $D = \begin{bmatrix} 5 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \end{bmatrix} \Rightarrow \sqrt{16+9} \Rightarrow 5.$

$A = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $E = \begin{bmatrix} 6 \\ 5 \end{bmatrix} \Rightarrow \begin{bmatrix} 5 \\ 4 \end{bmatrix} \Rightarrow \sqrt{25+16} \Rightarrow \sqrt{41} \Rightarrow 6.4$

2nd row:

$B = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ $C = \begin{bmatrix} 4 \\ 3 \end{bmatrix} \Rightarrow \begin{bmatrix} 2 \\ 2 \end{bmatrix} \Rightarrow \sqrt{4+4} \Rightarrow 2.82$

$B = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ $D = \begin{bmatrix} 5 \\ 4 \end{bmatrix} \Rightarrow \begin{bmatrix} 3 \\ 3 \end{bmatrix} \Rightarrow \sqrt{9+9} = 4.24$

$B = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ $E = \begin{bmatrix} 6 \\ 5 \end{bmatrix} \Rightarrow \begin{bmatrix} 4 \\ 4 \end{bmatrix} \Rightarrow 5.65$

3rd row:

$$C = \begin{bmatrix} 4 \\ 3 \end{bmatrix} \quad D = \begin{bmatrix} 5 \\ 4 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ 1 \end{bmatrix} \rightarrow \sqrt{2} = 1.41$$

$$C = \begin{bmatrix} 4 \\ 3 \end{bmatrix} \quad E = \begin{bmatrix} 6 \\ 5 \end{bmatrix} \rightarrow \begin{bmatrix} 2 \\ 2 \end{bmatrix} \rightarrow \sqrt{4+4} \rightarrow 2.82$$

4th row:

$$D = \begin{bmatrix} 5 \\ 4 \end{bmatrix} \quad E = \begin{bmatrix} 6 \\ 5 \end{bmatrix} \rightarrow \begin{matrix} 1 \\ 1 \end{matrix} \rightarrow \sqrt{2} = 1.41$$

Min dist $A \rightarrow B$

Merge A & B $\rightarrow$ Now clusters are : (AB) C D E

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

1st row

$$C = \begin{bmatrix} 4 \\ 3 \end{bmatrix} \quad B = \begin{bmatrix} A \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 2 \\ 2 \end{bmatrix} \rightarrow \sqrt{4+4} = \sqrt{8} = 2.82$$

min

$$C = \begin{bmatrix} 4 \\ 3 \end{bmatrix} \quad A = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 3 \\ 2 \end{bmatrix} \rightarrow \sqrt{9+4} \rightarrow \sqrt{13} = 3.60$$

$D = \begin{bmatrix} 5 \\ 4 \end{bmatrix}$ $A = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 4 \\ 3 \end{bmatrix} \Rightarrow \sqrt{16+9} \Rightarrow 5$

$D = \begin{bmatrix} 5 \\ 4 \end{bmatrix}$ $B = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 3 \\ 3 \end{bmatrix} \Rightarrow \sqrt{9+9} \Rightarrow 4.24$ min

$E = \begin{bmatrix} 6 \\ 5 \end{bmatrix}$ $A \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 5 \\ 4 \end{bmatrix} \sqrt{25+16} = 6.4$

$E \begin{bmatrix} 6 \\ 5 \end{bmatrix}$ $B \begin{bmatrix} 2 \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 4 \\ 4 \end{bmatrix} \Rightarrow \sqrt{32} = 5.65$ min

---

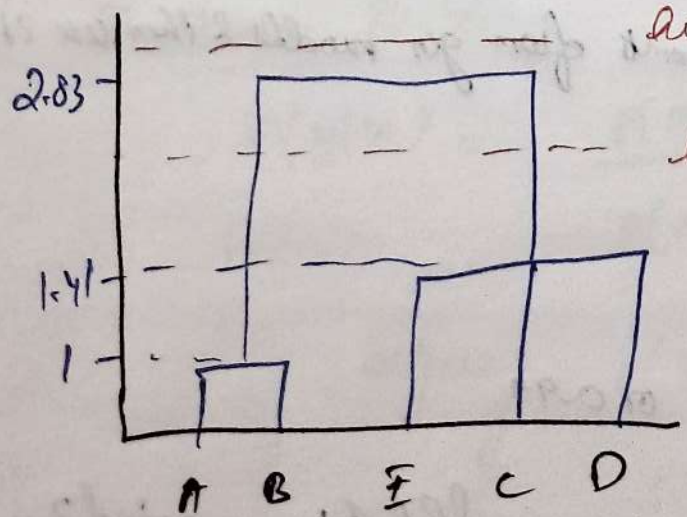
2nd row

C & D ⇒ 1.41

C & E = 2.82

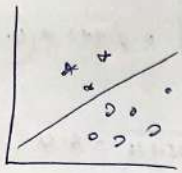
3rd row

D & E ⇒ 1.41



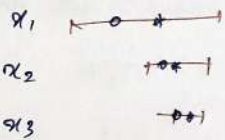
03/04/25

Discriminative Learning



Discriminative model

Generative learning - Generate models



o - 'computer-like'

+ - computer.

Learns only 1 cluster's pattern, properties, distribution

After learning → probability

When dataset is small, we can generate datapoints from gen models & then use it for discriminative models.

Puzzle:-

1) Accuracy of test = 99% or 0.99

2) Tested +ve

3) Rare disease → 1 in 10000

Shd I be worried?

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

	D	D'
T		
T'		

$P(D) = 0.0001$

$P(D') = 0.9999$

$$= \frac{0.0001 \times 0.99}{0.0001 \times 0.99 + 0.01 \times 0.9999} = \frac{0.000099}{0.000099 + 0.009999}$$

$$= \frac{0.000099}{0.010098} = 0.009803921\overline{6}$$

Discriminative model

$P(y|x)$

↓ conditional

Generative model

$P(x, y) = P(x|y) \cdot P(y)$

↓ Joint probability ↓ cond^n ↓ marginal

Prediction Problems

$$P(y|x) = \frac{P(x|y) \cdot P(y)}{P(x|y=0)P(y=0) + P(x|y=1)P(y=1)}$$

(likelihood, class prior, Posterior, Prior)

$$\underset{y}{\text{argmax}}\ P(x|y) \cdot P(y)$$

04/04/25

GENERATIVE MODELS

Naive Bayes Classifier

Applic: IRL – spam, nospam

email with these words

↓

"To buy this product"

↓

[ ] . T or [ ] (OR)

Convert to vector containing numbers

Complexity:

⇒ context transfer is difficult.

⇒ Semantic meanings need to be captured

"-kiy ---- empire---"

↓

this rep shd be closer to empire

relative distance b/w kiy & empire shd be closer than that of kiy & animal.

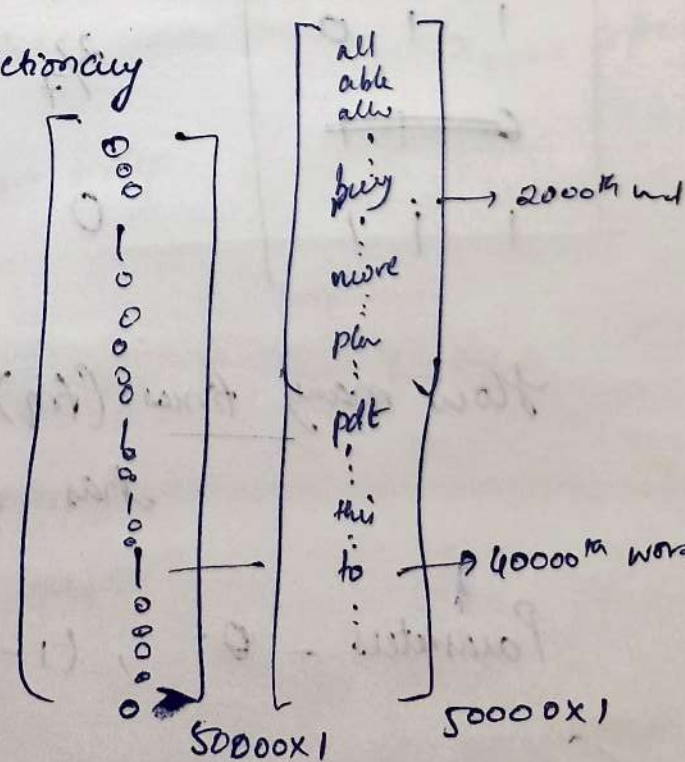
SEMANTIC SIMILARITY

Assume that we have 50,000 words like in a dictionary

"To buy this product"

Text modality → Vector =

	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



Assign prob for every row/

#				
8	1	0	0	Span
9	0	1	0	Span
10	1	0	0	Span
11	1	1	0	No span

$2^3 = 8$ combinations

All possible comb^n — once we fill all the prob, for the new sample

x_1	x_2	x_3	prob $y=1$ (spam)	Prob $y=0$ (No spam)
0	0	0	1	0
0	0	1	0	1/4 1/4
0	1	0	2/7	0
0	1	1	0	0
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

joint prob

0.28
7 | 2.0
1.0d
20

How many times (frq) that a combination -100 has led to spam?

Parameters $- \theta$, $(1-\theta)$

(Multinomial distb^n we have vector of possibilities $\theta_1, \theta_2 \ldots \theta_{20}$ .. etc.)

※ if our test sample is 1,1,0. → we classify it as spam

$2/7 = 0.28$

$1/4 = 0.25$

⇒ if our feature explodes (like 50,000 or more)
we can't explore all possibilities,
(exponential) computationally expensive.
Though it's a working model, it is not practical.

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

[Fraction of samples in a class]

Naive Bayes assumes – Boolean

↓

follows binomial distb"

eg:

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

Text msg: Free win

Features:

	word "free" x_1	word "win" x_2	Label
#1	Yes/1	Yes/1	Spam
#2	No/0	Yes/1	Spam
#3	No/0	No/0	Not spam
#4	No/0	Yes/1	Not spam
#5	Yes/1	No/0	Not spam

Step 1:

Calculate prior probabilities

Prior prob y=spam = $2/5$, y=not spam = $3/5$

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

y=1 ↓ spam, X ↓ free win

$= P(X|y)$

① Prior

$P(spam = 1) = 2/5$ or 0.4

$P(Not\ spam) = 3/5 = 0.6$

② Likelihood

For spam class, how many times free appear? 1 of 2

Free appears 1 of 2 spam class

from train data

$P(Free = yes \mid spam = 1) = 1/2$

$P(win = yes \mid spam = 1) = 2/2 = 1$

For not spam class;

$P(Free = yes \mid spam = 0 \text{ or not spam}) = 1/3$

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

③ Bayes rule - Inference / Test phase

$P(Free = yes, win = yes)$ → Text

$$P(Spam \mid X) = \frac{P(X \mid spam)\ P(spam)}{}$$

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

$$= P(X_1 \mid spam),\ P(X_2 \mid spam)$$

$$= 1/2 \times 1$$

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

$$= \frac{1}{5} = 0.2$$

$$\boxed{P(\text{spam} \mid X) = 0.2}$$

Similarly

$$P(\text{Not spam} \mid X) = P(X \mid \text{Not spam})\, P(\text{Not spam})$$

$$P(X \mid \text{Not spam}) \times \frac{3}{5}$$

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

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

$$= \frac{1}{3} \times \frac{1}{3}$$

$$P(\text{Not spam} \mid X) = \frac{1}{3} \times \frac{1}{3} \times \frac{3}{5}$$

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

$$\boxed{P(\text{Not spam} \mid X) = 0.066}$$

Free Discount $P(X_1 = 1)$ ↓ Free Discount → 0 $(X_2 = 0)$

09/04/25

Bag of words

Text to vector

Daphne Co-- Stanford

PROBABILISTIC GRAPHICAL MODELS (PGM)

(or)

BAYESIAN NETWORKS

HMMs → is PGM

practice tables to represent probabilities.

{ r, r, r, b }

$P(B_1)$ = (Marginal)

B_1	r	b
	3/4	1/4

$P(B_2 | B_1)$ = (Conditional)

B_1 \ B_2	r	b
r	2/3	1/3
b	1	0

= 1

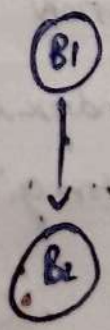
CPT - cond. prob table

$P(B_1, B_2) = P(B_2, B_1)$ (Joint)

B_1 \ B_2	r	b
r	$3/4 \times 2/3 = 1/2$	$3/4 \times 1/3 = 1/4$
b	$1/4 \times 1 = 1/4$	0

Joint table - prob (overall sums to 1)

Graphical represent^n - SIMPLE GRAPHICAL MODEL



Judea Pearl - Stanford

↓

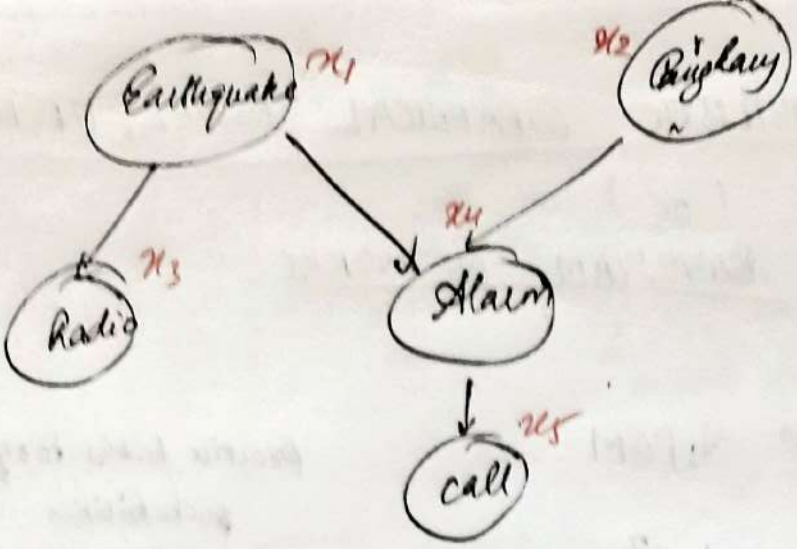
Turing Award

Nodes - random variable

Edges - Direct Influence ("causation")

Causation → difficult to compute

Directed Acyclic



Assume Boolean

call - observed

Earthquake & Burglary hidden

Joint probability

$P(B, E, A, R, C)$

$= P(B) \cdot P(E|B)\, P(A|B,E)$

$P(R|B,E,A)\, P(C|B,E,A,R)$

order doesn't matter

↓

This eqn is equivalent to

Constructing the joint prob table $2^5 = 32$

using assumptions

$P(C|B,E,A,R)$

knock off others

only depends on the immediate parent

$P(R|B,E,A) = P(R|E)$

We can ignore others.

$P(A|BE) =$

Conditional independence

Constructing Joint prob

E	B	R	A	C
0	0	0	0	0
-	-	-	-	1
-	-	-	-	-
-	-	-	-	-

Constructing prob table is diff $2^5 = 32$.

[These Naive Bayes - we assume independence to reduce space]

↓

Here we are assuming differently

↓

ASSUMPTION

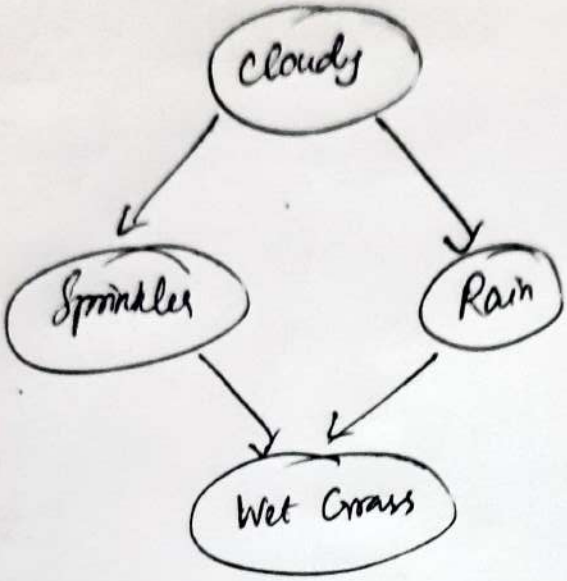
X_i's are independent of $X_{ancestors}$ given X parents

Don't confuse w HMM time series - these features are not sequence

$$P(x_1, x_2 \dots x_d) = \prod_{i=1}^{d} P(x_i \mid \text{parents}(x_i))$$

Other eg:

$P(C = F) \to 0.5 \quad P(C = T) = 0.5$



$P(S = 1)$?

$P(R|C)$

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

$P(W|R,S)$

R	P(W=F)	P(W=T)
F	1	0
F	0.1	0.9
T	0.1	0.9
T	0.01	0.99