

FEBRUARY'22						
SU	MO	TU	WE	TH	FR	SA
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28					

019-846 • Week 04

2022 January 19
Wednesday

M.L

Linear Regression

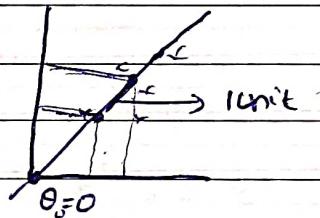
Equation of straight line

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

θ_1 = slope or coefficient

θ_0 = intercept

→ when θ_0 (intercept) is 0 that means the line is intersecting the y-axis at 0.



→ θ_1 = Slope, i.e. one unit movement towards the x-axis what is the unit movement in y-axis

$$\text{Cost-function} = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 \rightarrow \text{Square Error function}$$

→ $\frac{1}{m}$ gives average of all the values.

The note of the perfect personality is not rebellion, but peace. - Oscar Wilde,

Notes:

Appointment:

Phones:

FEBRUARY

MARCH

APRIL

FEBRUARY '22						
SU	MO	TU	WE	TH	FR	SA
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28					

081-344 • Week 04

2022 January 21
Friday

chart we need to solve

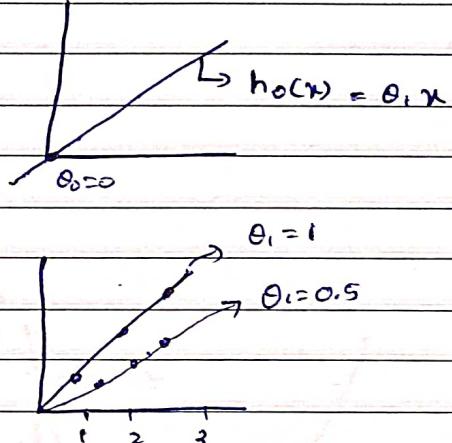
minimize cost function by adjusting θ_0, θ_1

① $h_\theta(x) = \theta_0 + \theta_1 x$ if $\theta_0 = 0$

$h_\theta(x) = \theta_1 x$

Ex: our data points are

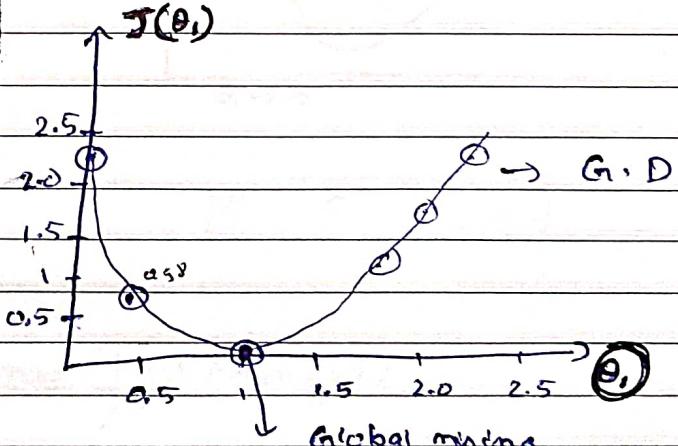
② $S_{\text{sum}} = (1,1) \quad \boxed{\theta_0=0} \quad \boxed{\theta_1=1}$
 $(2,2)$
 $(3,3)$



③ $J(\theta_1) = \frac{1}{2m} \sum_{i=1}^3 (h_\theta(x) - y)^2$
 $= \frac{1}{2m} [(1-1)^2 + (2-2)^2 + (3-3)^2]$
 $= 0$

④ $\theta_0=0, \theta_1=0.5$

$J(\theta_1) = \frac{1}{2m} [(0.5-1)^2 + (1-2)^2 + (1.5-3)^2]$
 $= 0.58$



Do we not realize that self respect comes with self reliance? - A. P. J. Abdul Kalam

Notes:

Appointment:

Phone:

FEBRUARY

MARCH

APRIL

FEBRUARY'22						
SU	MO	TU	WE	TH	FR	SA
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28					

023-342 • Week 04

2022 January 23
Sunday

Convergence

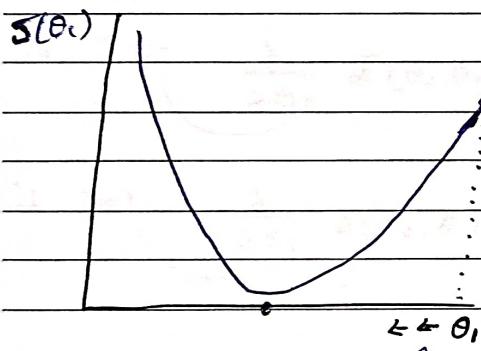
Repeat until convergence

{

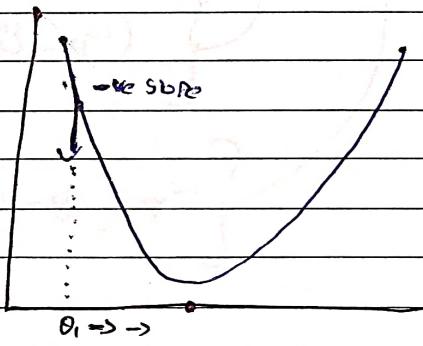
$$\theta_j = \theta_j - \alpha \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_j} \rightarrow \text{derivative of slope}$$

}

Explanation for derivative of slope.



the slope
(forward)



$\theta_i \rightarrow -$

$$\theta_i = \theta_i - \alpha (+ve)$$

$$\theta_i = \theta_i - \alpha (-ve)$$

$\alpha \rightarrow$ (speed) model must converge (hyper parameter)

Trees are the earth's endless effort to speak to the listening heaven. - Rabindranath Tagore

Notes:	Appointment:	Phones:

FEBRUARY

MARCH

APRIL

FEBRUARY'22

SU	MO	TU	WE	TH	FR	SA
	1	2	3	4	5	
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28					

025-340 • Week 05

2022 January 25
TuesdayGradient descent algorithm

repeat until convergence

{

$$\theta_j = \theta_j - \alpha \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_j}$$

}

$$\frac{\partial J(\theta_0, \theta_1)}{\partial \theta_0} = \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$$

if $J=0$ $\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})$

if $J=1$ $\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) x^{(i)}$

You cannot believe in god until you believe in yourself. - Swami Vivekananda

Notes:

Appointment:

Phones:

FEBRUARY

MARCH

APRIL

FEBRUARY'22						
SU	MO	TU	WE	TH	FR	SA
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28					

027-338 • Week 05

2022 January 27
Thursday

Performance metrics

R^2 and Adjusted R^2

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$

SS_{res} = Sum of residuals

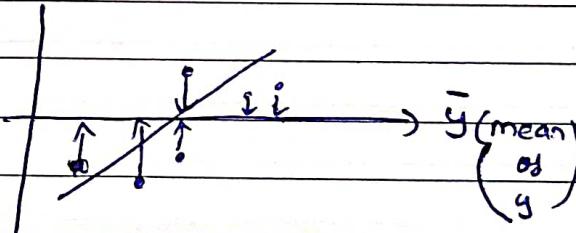
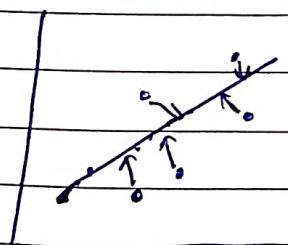
SS_{tot} = Sum of total.

$$SS_{res} = \sum (y_i - \hat{y}_i)^2$$

$$SS_{tot} = \sum (y_i - \bar{y})^2$$

Sum of residuals

Sum of total



$$\therefore 1 - \frac{\text{low}}{\text{high}} \rightarrow \text{small number} \Rightarrow \rightarrow \text{high}$$

We should not give up and we should not allow the problem to defeat us. - A. P. J. Abdul Kalam

Notes:

Appointment:

Phones:

FEBRUARY

MARCH

APRIL

FEBRUARY'22						
SU	MO	TU	WE	TH	FR	SA
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28					

0.29-886 • Week 0.5

2022 January 29
Saturday

→ Drawback with R^2 is

if we keep on adding new features the R^2 value will keep on increasing irrespective of correlation b/w that new independent & dependent feature.

→ To solve that Adjusted R^2

$$R^2 \text{ adjusted} = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

P = no. of Features

n = no. of data points

Don't cry because it's over, smile because it happened. - Dr. Seuss

Notes:

Appointment:

Phones:

FEBRUARY

MARCH

APRIL

FEBRUARY'22						
SU	MO	TU	WE	TH	FR	SA
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28					

031-334 • Week 06

2022 January 31
Monday

Ridge & Lasso regressions

Overshifting

- Low Bias
- High Variance

Ridge → L2 regularization.

$$J(\theta, \theta_0) = \frac{1}{2m} \sum_{i=1}^m (\hat{y}_i - y^{(i)})^2$$

$$(\hat{y}_i - y^{(i)}) + \lambda \underbrace{(\text{slope})^2}_{\text{Penalty term}}$$

→ Reduce overfitting.

↗ → how fast we want to change the

→ making low Bias &
high Variance

Steeper

I'm torn between the desire to create and the desire to destroy. - Charles M. Schulz

Notes:

Appointment:

Phones:

FEBRUARY

MARCH

APRIL

MARCH'22						
SU	MO	TU	WE	TH	FR	SA
	1	2	3	4	5	
5	7	8	9	10	11	12
3	14	15	16	17	18	19
10	21	22	23	24	25	26
27	28	29	30	31		

0.82-3.33 • Week 06

2022 February 01
Tuesday

Lasso Regression

$$(\hat{y} - y)^2 + \lambda |\text{slope}|$$

$$h_{\theta}(x) = \hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \dots + \theta_n x_n$$

Slope $= |\theta_0 + \theta_1 + \theta_2 + \dots + \theta_n| \rightarrow$ Feature selection

Assumptions of linear regression.

- ① Error terms are normally distributed
- ② Normal / Gaussian Distribution \rightarrow Model will get Trained well
- ③ Standardization {scaling data} \rightarrow Z-score $\mu=0, \sigma=1$
- ④ Model based on Linearity
- ⑤ Multicollinearity, Variation Inflation factor can solve multicollinearity.
- ⑥ Residual must be uncorrelated \rightarrow Exogeneity
- ⑦ Error term must show case constant \rightarrow Homoscedasticity

I'm told I'm very charming when people do what I want. - Steven Brust,

Notes:

Appointment:

Phone:

FEBRUARY

MARCH

APRIL

MARCH'22						
SU	MO	TU	WE	TH	FR	SA
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

034-331 • Week 06

2022 February 03
Thursday

Logistic Regression (Classification)

→ Why not use linear regression?

① Linear regression line can be affected by outliers. This leads to wrong predictions.

② Sometimes our prediction can get greater than 1 and less than 0 if we use linear regression.

→ For 2nd reason we have to make a limit to the line without crossing 0.

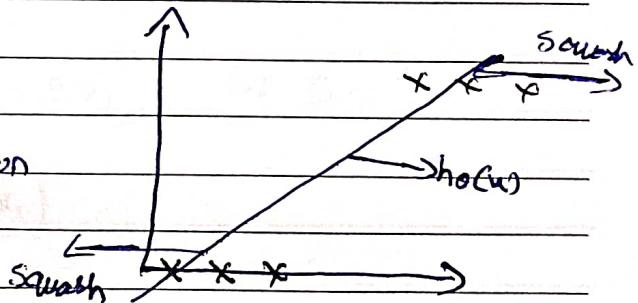
For that we are using Sigmoid

Decision Boundary Logistic Regression.

$$h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \dots + \theta_n x_n$$

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

→ Apply some function on linear regression to squash this line.



You don't have to be great to start, but you have to start to be great. - Zig Ziglar

Notes:

Appointment:

Phones:

MARCH

APRIL

MARCH'22

SU	MO	TU	WE	TH	FR	SA
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

036-329 • Week 06

2022 February 05
Saturday

$$h_{\theta}(x) = \theta_0 + \theta_1 x_1$$

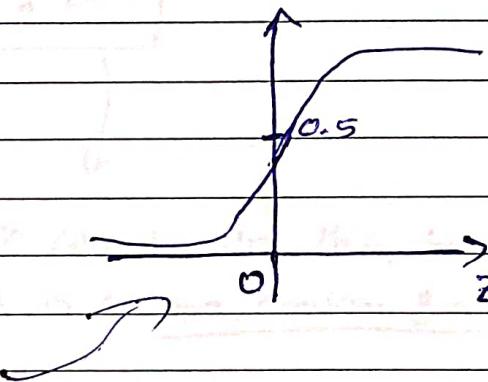
$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1)$$

let $Z = \theta_0 + \theta_1 x_1$

$$\rightarrow \therefore h_{\theta}(x) = g(Z)$$

g = Sigmoid or logistic

$$h_{\theta}(x) = \frac{1}{1 + e^{-Z}}$$



→ Based on the graph

$$\begin{cases} g(z) \geq 0.5 \\ \text{when } z \geq 0 \end{cases} \quad \begin{cases} g(z) \leq 0.5 \\ \text{when } z \leq 0 \end{cases}$$

Training Set

$$\{(x^1, y^1), (x^2, y^2), (x^3, y^3), \dots, (x^n, y^n)\}$$

$$y \in \{0, 1\} \rightarrow 2 \text{ class}$$

$$Z = \theta_0 + \theta_1 x_1$$

if $\theta_0 = 0$

$$h_{\theta}(z) = \frac{1}{1 + e^{-Z}}$$

$$Z = \theta_1 x_1$$

Let your life lightly dance on the edges of time like dew on the tip of a leaf. - Rabindranath Tagore

Notes:

Appointment:

Phones:

MARCH

APRIL

MARCH'22						
SU	MO	TU	WE	TH	FR	SA
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

038-327 • Week 07

2022 February 07
Monday

Change Parameter θ_i to get best fit curve.

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

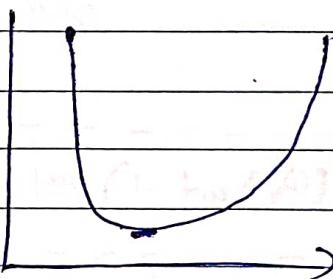
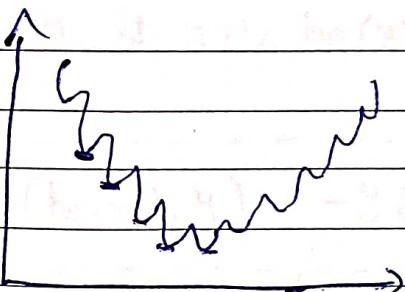
here
$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

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

We cannot use this, because it is a non convex function.

Non Convex Function

Convex Function



The revolution introduced me to art, and in turn, art introduced me to the revolution! - Albert Einstein

Notes:

Appointment:

Phones:

MARCH

APRIL

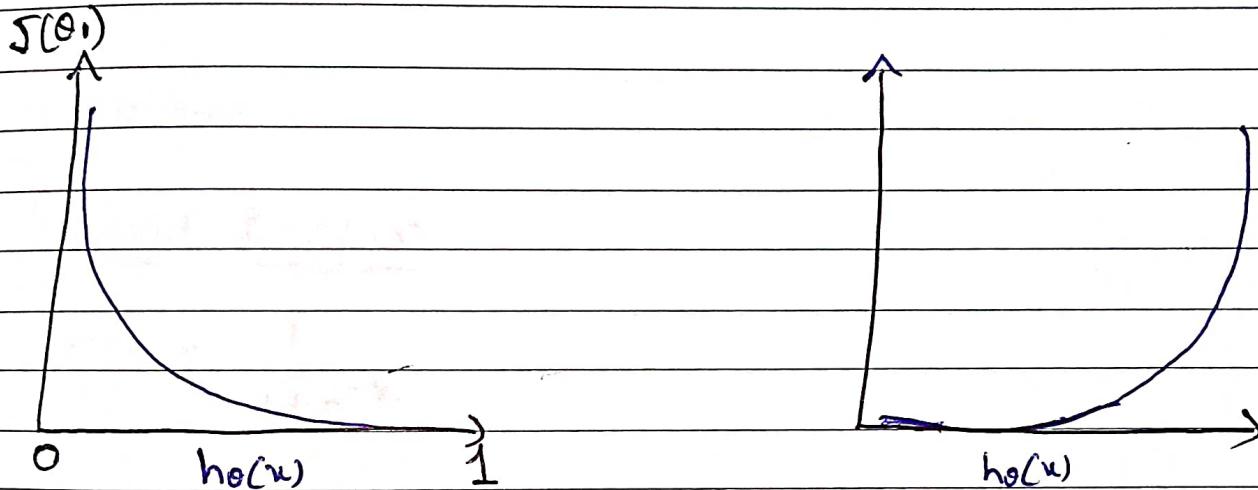
MARCH'22						
SU	MO	TU	WE	TH	FR	SA
	1	2	3	4	5	
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

040-325 • Week 07

2022 February 09
Wednesday

Logistic regression cost function.

$$J(\theta_0) = \begin{cases} -\log(h_{\theta}(x)) & y=1 \\ -\log(1-h_{\theta}(x)) & y=0 \end{cases}$$



$$\text{Cost} = 0 \text{ if } y=1, h_{\theta}(x)=1 \quad \text{if } y=0$$

MARCH

$$\text{Cost}(h_{\theta}(x^i), y) = -y \log(h_{\theta}(x^i)) - (1-y) \log(1-h_{\theta}(x^i))$$

if $y=0$, $\uparrow 0$ if $y=1$, $\uparrow 0$

$$\frac{1}{m} \sum_{i=1}^m$$

Beauty is truth's smile when she beholds her own face in a perfect mirror. - Rabindranath Tagore

Notes:

Appointment:

Phone:

APRIL

MARCH '22						
SU	MO	TU	WE	TH	FR	SA
	1	2	3	4	5	
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

042-323-0707

2022 February 11
Friday

Just like LR

Repeat until convergence

{

$$\theta_j = \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j}$$

f

→ log likelihood.

⇒ Sigmoid function

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

~~g(x) = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots)~~

$$P(x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots)}}$$

or simply

$$P(x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}} = \frac{e^{(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}}{1 + e^{(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}} - ①$$

Do not take life too seriously. You will never get out of it alive. - Elbert Hubbard

Notes:

Appointment:

Phones:

MARCH

APRIL

Sigmoid = Probability

logit = the or -ve.

044-321 • Week 07

MARCH'22						
SU	MO	TU	WE	TH	FR	SA
1	2	3	4	5		
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

2022 February 13
Sunday

$$1 - P(x) = 1 - \frac{e^{(\beta_1 x + \beta_0)}}{1 + e^{\beta_1 x + \beta_0}}$$

$$1 - P(x) = \frac{1}{1 + e^{(\beta_1 x + \beta_0)}} \quad \text{--- (2)}$$

$$\beta_1 x + \beta_0 = z$$

$$(1) / (2) = \frac{P(x)}{1 - P(x)} = \frac{\frac{e^z}{1 + e^z}}{\frac{1}{1 + e^z}} = e^z$$

$$\frac{P(x)}{1 - P(x)} = e^z$$

$$\log \left(\frac{P(x)}{1 - P(x)} \right) = z \quad \therefore \quad \boxed{\log \left(\frac{P(x)}{1 - P(x)} \right) = \beta_1 x + \beta_0}$$

$$\begin{cases} z = -ve & \text{if } P(x) < 0.5 \\ z = +ve & \text{if } P(x) > 0.5 \end{cases}$$

I'm a conundrum. Or an enigma. I forget which. - James A. Owen

Notes:

Appointment:

Phones:

MARCH

APRIL

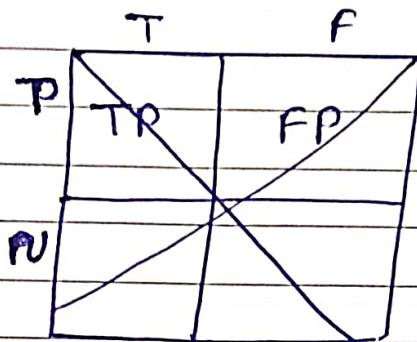
MARCH'22						
SU	MO	TU	WE	TH	FR	SA
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

046-319 • Week 08

2022 February 15
Tuesday

Performance metrics { classification }

Actual		TP	FP	T	
Predicted	TP			TN	F
	FN				



TP = Positive values Predicted +ve (correct)

TN = -ve values Predicted -ve (correct)

FP = -ve Predicted as +ve

FN = +ve Predicted as -ve

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Each day of our lives we make deposits in the memory banks of our children. -Charles R. Swindoll

Notes:

Appointment:

Phone:

MARCH

APRIL

MARCH'22						
SU	MO	TU	WE	TH	FR	SA
	1	2	3	4	5	
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

048-317 • Week 08

2022 February 17
Thursday

0 → 900 } Imbalanced dataset
1 → 100 }

0 → 600 } Balanced dataset
1 → 400 }

→ So consider Imbalanced dataset 0 : 900
 1 : 100
                                                ~~~~~  
model gives → 0  
that gives 90% accuracy.

→ So we use Precision, Recall, F-score

Precision =  $\frac{TP}{TP + FP}$       ex: Spam classification

Recall =  $\frac{TP}{TP + FN}$       ex: Cancer or not

$F\beta = \frac{(1+\beta^2)(\text{Precision} * \text{Recall})}{\beta^2 * \text{Precision} * \text{Recall}}$       ex: Stock market crash or not  
(Company, People)

I'm not stubborn. My way is just better. - Maya Banks.

Notes:

Appointment:

Phone:

MARCH

APRIL

MARCH'22						
SU	MO	TU	WE	TH	FR	SA
	1	2	3	4	5	
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

050-315 • Week 08

2022 February 19  
Saturday

## Naive Bayes Intuition

### - Baye's Theorem

X X X OO	$P(X) = \frac{3}{5}$	$X = \text{Red}$
	$P(O) = \frac{2}{4} = \frac{1}{2}$	$O = \text{Green}$

$$P(R \text{ and } G) = P(R) \times P(G|R) \rightarrow \text{Conditional Probability}$$

$$P(G \text{ and } R) = P(G) \times P(R|G)$$

$$\Rightarrow P(A \text{ and } B) = P(B \text{ and } A)$$

$$P(A) \times P(B|A) = P(B) \times P(A|B)$$

$$P(B|A) = \frac{P(B) \times P(A|B)}{P(A)} \rightarrow \text{Bayes theorem}$$

A brother may not be a friend, but a friend will always be a brother.- Benjamin Franklin

Notes:

Appointment:

Phones:

MARCH

APRIL

MARCH'22						
SU	MO	TU	WE	TH	FR	SA
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

052-813 • Week 09

2022 February 21  
Monday

Input

Output

$[x_1 \ x_2 \dots x_n]$

$y$

$$P(y/x_1, x_2, \dots, x_n) = \frac{P(y) * P(x_1|y) * P(x_2|y), P(x_3|y) \dots P(x_n|y)}{P(x_1) * P(x_2) * P(x_3) \dots P(x_n)}$$

ex.  $x_1 \ x_2 \ x_3 \ x_4 \ y$

Yes

No

$$P(y=\text{yes}/x_i) = \frac{P(\text{yes}) * P(x_1|\text{yes}) * P(x_2|\text{yes}) * P(x_3|\text{yes}) * P(x_4|\text{yes})}{P(x_1) * P(x_2) * P(x_3) * P(x_4) \rightarrow \text{const}}$$

$$P(y=\text{no}/x_i) = \frac{P(\text{no}) * P(x_1|\text{no}) * P(x_2|\text{no}) * P(x_3|\text{no}) * P(x_4|\text{no})}{P(x_1) * P(x_2) \rightarrow \text{const}}$$

if  $P(\text{yes}/x_i) = 0.13$

$P(\text{no}/x_i) = 0.05$

Normalizing

$$P(\text{yes}/x_i) = \frac{0.13}{0.13 + 0.05} = 0.72$$

$$P(\text{no}/x_i) = \frac{0.05}{0.05 + 0.13} = 0.28$$

The important thing is not to stop questioning. Curiosity has its own reason for existing. - Albert Einstein

Notes:

Appointment:

Phone:

MARCH

APRIL

24 February 2022  
Thursday

FEBRUARY '22						
SU	MO	TU	WE	TH	FR	SA
6	7	8	9	10	11	
13	14	15	16	17	18	
20	21	22	23	24	25	
27	28					

Week 09 • 055-310

<u>Day</u>	<u>outlook</u>	<u>Temperature</u>	<u>Humidity</u>	<u>wind</u>	<u>Play Tennis</u>
D1	sunny	Hot	High	Weak	NO
D2	sunny	Hot	High	Strong	NO
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	NO
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	NO
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	mild	Normal	Strong	Yes
D12	Overcast	mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	mild	High	Strong	NO

There is only one happiness in this life, to love and be loved. - George Sand

Phones:

Appointment:

Notes:

MARCH '22						
SU	MO	TU	WE	TH	FR	SA
	1	2	3	4	5	
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

056-309 • Week 09

2022 February 25  
Friday

Consider only outlook & temperature.

### outlook (column)

	Yes	No	$P(Y)$	$P(N)$
Sunny	2	3	$P(Y S) = 2/5$	$P(N S) = 3/5$
overcast	4	0	$P(Y O) = 4/4 \approx 1$	$P(N O) = 0/4 = 0$
Rain	3	2	$P(Y R) = 3/5 \approx 0.6$	$P(N R) = 2/5 \approx 0.4$
Total =	9	5		

### Temperature

	Yes	No	$P(Y)$	$P(N)$	
Hot	2	2	2/4	2/5	Play
mild	4	2	4/6	2/5	Yes $P(\text{yes}) = 9/14$
Cold	3	1	3/4	4/5	No $P(\text{no}) = 5/14$
Total	9	5			

The last capitalist we hang shall be the one who sold us the rope. - Karl Marx

Notes:

Appointment:

Phone:

MARCH

APRIL

MARCH'22						
SU	MO	TU	WE	TH	FR	SA
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

058-307 • Week 0.9

2022 February 27  
Sunday

Test data (Sunny, Hot) → OIP ??

$$P(\text{Yes} | \text{Sunny, Hot}) = P(\text{Yes}) \times P(\text{Sunny} | \text{Yes}) \times P(\text{Hot} | \text{Yes})$$

$$\frac{P(\text{Sunny})}{P(\text{Hot})}$$

$$= \frac{9}{14} \times \frac{2}{9} \times \frac{2}{9}$$

$$= \frac{2}{63} = 0.031$$

$$P(\text{No} | \text{Sunny, Hot}) = P(\text{No}) \times P(\text{Sunny} | \text{No}) \times P(\text{Hot} | \text{No})$$

$$\frac{P(\text{Sunny})}{P(\text{Hot})}$$

$$= \frac{5}{14} \times \frac{3}{5} \times \frac{2}{5}$$

$$= \frac{3}{35} = 0.085$$

Prediction ↗

$$P(\text{No} | \text{Sunny, Hot}) = 0.085 = \frac{0.085}{0.031 + 0.085} = 0.73 = 73\% (\text{No})$$

$$P(\text{Yes} | \text{Sunny, Hot}) = 1 - 0.73 = 0.27 = 27\% (\text{Yes})$$

Good friends, good books, and a sleepy conscience: this is the ideal life. - Mark Twain

Notes:

Appointment:

Phones:

MARCH

APRIL

APRIL'22						
SU	MO	TU	WE	TH	FR	SA
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

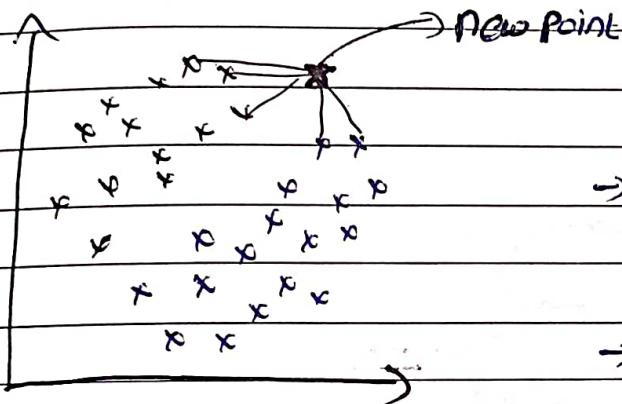
060-305 • Week 10

2022 March 01  
Tuesday

## K-Nearest Neighbour

(X) Outliers

(X) Imbalanced



if  $[K=5]$

→ Calculate distance b/w new point and nearest points.

(Euclidean distance)

→ Vote

→ 3 Black

2 Blue

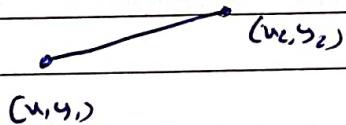
→ So New Point goes into Black.

→ How can we calculate distance?

→ In case regression

→ Get the average of nearest 5 points.

Euclidean distance -



Manhattan distance



$$\sqrt{(u_1 - u_2)^2 + (y_1 - y_2)^2}$$

$$|(u_2 - u_1) + y_2 - y_1|$$

Show me a family of readers, and I will show you the people who move the world. - Napoléon Bonaparte

Notes:

Appointment:

Phones:

MARCH

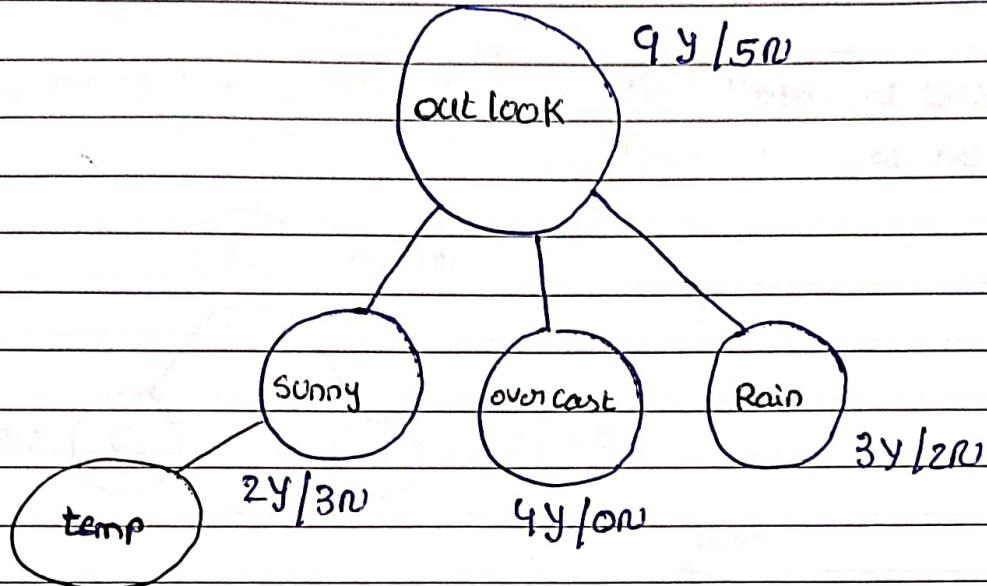
APRIL

APRIL'22						
SU	MO	TU	WE	TH	FR	SA
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

062-303 • Week 10

2022 March 03  
Thursday

## Decision Tree



Pure split - Overcast - 4y/0n → Always Yes (Pure) - No split.

→ How can we say pure split?

→ Entropy

→ Gini Coefficient (g) Gini Impurity

→ How the features are selected

→ Information Gain

Dost thou love life? Then do not squander time, for that is the stuff life is made of. -Benjamin Franklin

Notes:

Appointment:

Phone:

APRIL'22						
SU	MO	TU	WE	TH	FR	SA
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

064-301 • Week 10

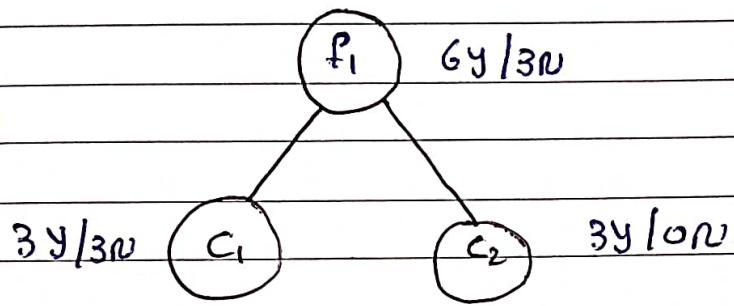
2022 March 05  
Saturday

## ① Entropy:

$$H(S) = -P_{(+)}\log_2(P_{(+)}) - P_{(-)}\log_2(P_{(-)})$$

$P_{(+)} = \text{Probab. of Yes}$

$P_{(-)} = \text{" " of No}$



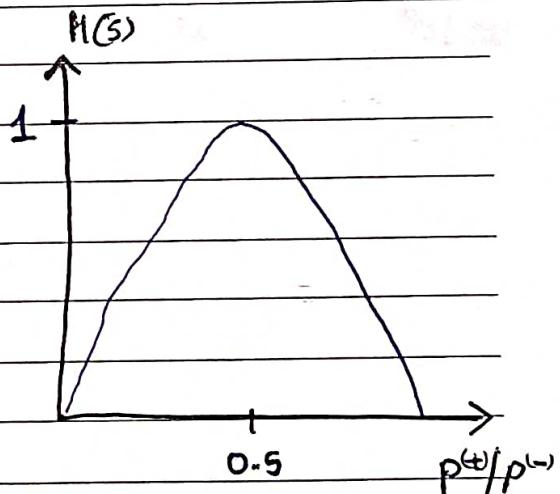
for  $C_2$

$$H(S) = -\frac{3}{3}\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right)$$

$$= -1\log_2(1)$$

$$= [0]$$

→ completely pure split



for  $C_1$

$$H(S) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$$

$$= [1] \rightarrow \text{completely impure split}$$

APRIL'22

SU	MO	TU	WE	TH	FR	SA
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

066-299-Week 11

2022 March 07  
Monday

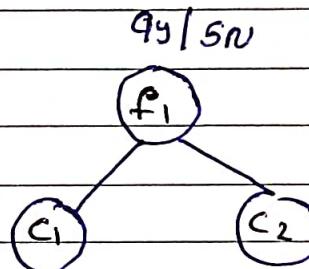
② Which feature to take to split?

→ Which feature should I take and make the split.

### Information Gain

$$\text{Grain}(S, f_1) = H(S) - \sum_{\text{Value}} \frac{|S_v|}{|S|} H(S_v)$$

Root node



$$H(S) = -P_{C_1} \log_2(P_{C_1}) - P_{C_2} \log_2(P_{C_2})$$

$$= -\frac{9}{14} \log_2(9/14) - \frac{5}{14} \log_2(5/14)$$

$$H(S) = 0.94$$

$$H(C_1) = -\frac{6}{8} \log_2(\frac{6}{8}) - \frac{2}{8} \log_2(\frac{2}{8})$$

$$H(C_2) = 0.81 \quad \text{Similarly} \quad H(C_2) = 1$$

$$\text{Grain}(S, f_1) = 0.94 - [\frac{8}{14} (0.81) + \frac{6}{14} (1)]$$

$$\text{Grain}(S, f_1) = 0.049$$

$f_1$  = Feature 1

Today's accomplishments were yesterday's impossibilities. — Robert H. Schuller

Notes:

Appointment:

Phones:

APRIL'22						
SU	MO	TU	WE	TH	FR	SA
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

068-297 • Week 11

2022 March 09

Wednesday

## Gini Impurity

$$G.I = 1 - \sum_{i=1}^n (P_i)^2$$

$$G.I = 1 - [(P_{+})^2 + (P_{-})^2]$$

lets take a node

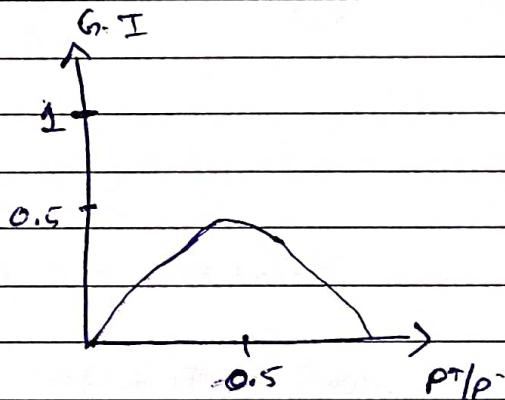
2y/bw

C<sub>1</sub>

$$G.I = 1 - \left[ \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right]$$

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

$$= 0.5$$



→ Entropy takes more time

→ Gini impurity have less time

Go to heaven for the climate and hell for the company.- Benjamin Franklin Wade

Notes:

Appointment:

Phones:

APRIL

APRIL'22

SU	MO	TU	WE	TH	FR	SA
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

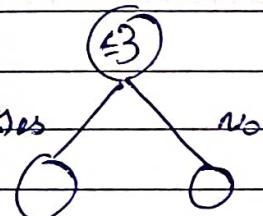
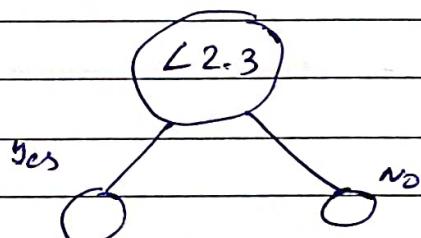
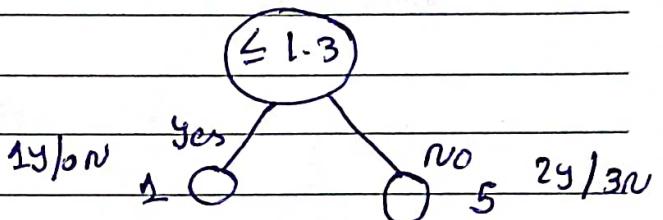
070-295 • Week 11

2022 March 11 Friday

What if Continuous Variables

$$f_1 \text{ O.P} \Rightarrow \text{Sort } f_1$$

2.3	y	1.3
1.3	y	2.3
4	n	3
5	y	4
7	n	5
3	n	1



→ like that for every now it will create and calculate

Information Gain. Then it will take best Information Gain.

Count your age by friends, not years. Count your life by smiles, not tears. - John Lennon

Notes:

Appointment:

Phones:

APRIL'22

SU	MO	TU	WE	TH	FR	SA
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

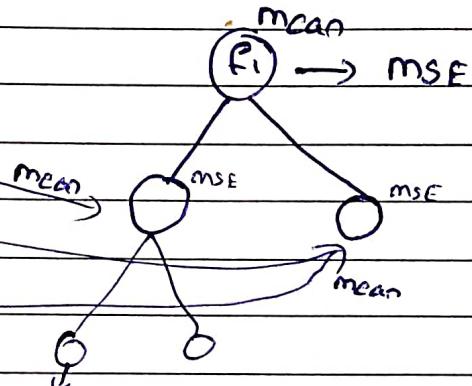
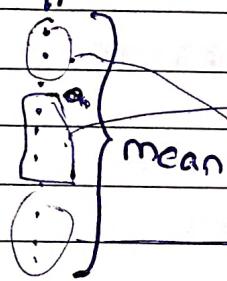
072-293 • Week 11

2022 March 13

Sunday

## Decision tree regression

$f_1 \ f_2 \ O/P$



Final mean is answer.

→ Check Decision tree by normalized nerd YouTube channel

One must wait until the evening to see how splendid the day has been. - Sophocles

Notes:

Appointment:

Phone:

APRIL

APRIL'22

SU	MO	TU	WE	TH	FR	SA
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

074-291 • Week 12

2022 March 15

Tuesday

⇒ What if there are very large no. of nodes?

→ Overshooting

Prevent overshooting

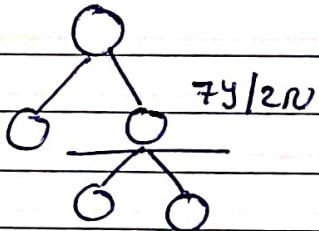
- Post Pruning
- Pre Pruning

→ Post Pruning

So in that node we can say that "70% accy"

Yes

So, no need of splitting.



→ This dividing is called Post Pruning.

→ Pre Pruning

→ hyperparameters will decide, like 'max\_depth', 'max\_leaf'

Happiness is not something ready made. It comes from your own actions. - Dalai Lama XIV

Notes:

Appointment:

Phones:

APRIL'22						
SU	MO	TU	WE	TH	FR	SA
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

076-289 • Week 12

2022 March 17  
Thursday

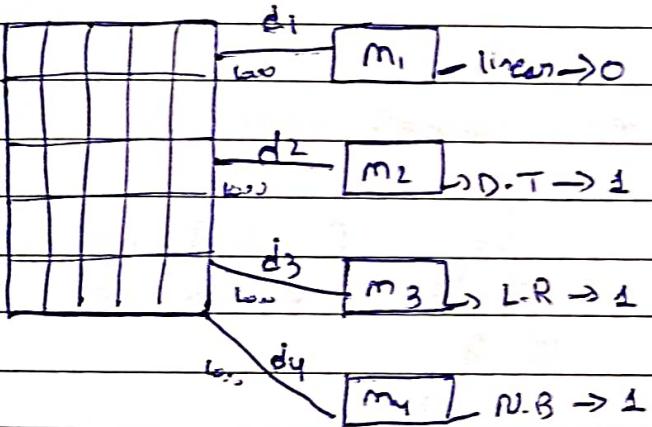
## Ensemble Techniques

### Ensemble Techniques



Bagging

Boosting



Sample data =  $d \quad [d < D]$

→ Some data can repeat,  $d_1$  in  $d_2, d_3, \dots$

→ Majority Voting so 1 }

Bootstrap Aggregation

→ If regression Average

I am not a product of my circumstances. I am a product of my decisions. Stephen Covey

Notes:

Appointment:

Phone:

APRIL'22

SU	MO	TU	WE	TH	FR	SA
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

07.8-287 • Week 12

2022 March 19

Saturday

Boosting → Sequential set of all models combined together. These models are weak learners. After that they become strong learners.

→  $m_1 \rightarrow m_2 \rightarrow m_3 \rightarrow m_4 \rightarrow \text{opt}$

weak  
learner

w.L

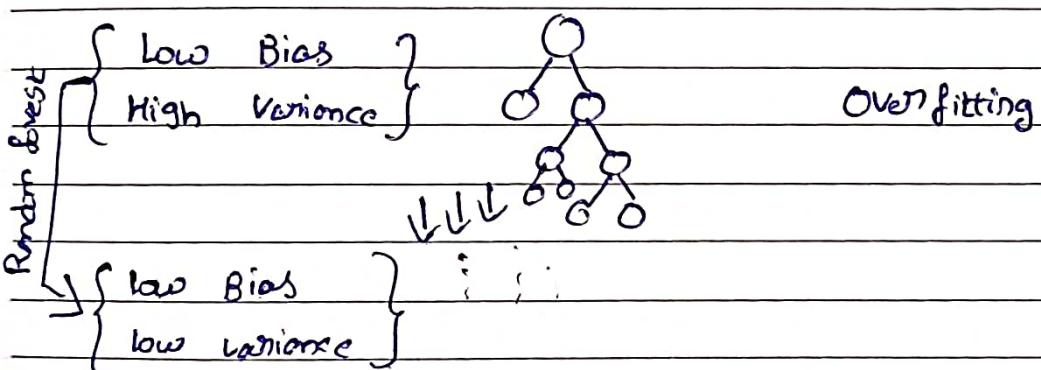
w.L

w.L

Strong  
learning

## Random forest classifier and Regressor

What is the main problem with D.T?



→ Like Bagging but all are decision trees.

→ Row Sampling, Feature Sampling

→ Selecting some features f

Efficiency is doing the thing right. Effectiveness is doing the right thing. -Peter F. Drucker

Notes:

Appointment:

Phone:

APRIL'22						
SU	MO	TU	WE	TH	FR	SA
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

080-285 • Week 13

2022 March 21  
Monday

- b) Is normalization required in Random forest.
- i) No, because whatever the value is, the decision will split the same.
- ii) Is standardization or normalization required?
- A) Yes, because if the values are high distance evaluation become complex, if standardized then easy to find distance.
- iii) Is random forest impacted by outliers?
- A) No, outlier impact is very less on random forest.  
because random forest handles outliers by essentially binning them.
- c) Why Naive Bayes is called as 'naive'?
- Consider apple → Red, round, 4 diameter  
These all features are related and dependent on existence of other feature.
- But Naive Bayes consider every feature as independent.
- It is Naive because it make assumptions may or may not correct.

I'd rather be a little weird than all boring.- Rebecca McKinsey

Notes:

Appointment:

Phone:

APRIL'22						
SU	MO	TU	WE	TH	FR	SA
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

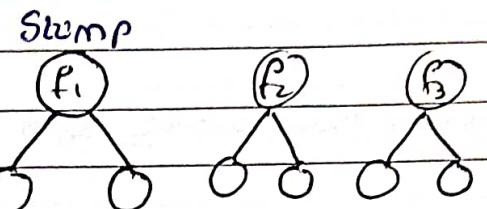
082-288-Wk 18

2022 March 23

Wednesday

Adaboost:

$f_1$	$f_2$	$f_3$	$f_4$	O/P	weight
-	-	-	-	Yes	$1/7$
-	-	-	-	No	$1/7$
-	-	-	-	Yes	$1/7$
-	-	-	-	No	$1/7$ wrong
-	-	-	-	Yes	$1/7$
-	-	-	-	Yes	$1/7$
-	-	-	-	No	$1/7$



① Total error =  $1/7$

② Performance of Stump =  $\frac{1}{2} \log_e \left( \frac{1 - TE}{TE} \right) = \frac{1}{2} \log_e \left( \frac{1 - 1/7}{1/7} \right) = 0.895$

③ update all these weights

New sample weight = weight \*  $e^{-PS}$  =  $\frac{1}{7} \times e^{-0.895} = 0.05$

Incorrect record = weight \*  $e^{+PS}$  =  $\frac{1}{7} \times e^{0.895} = 0.349$

Weight	Normalized Wt.	Buckets
0.05	0.07	[0 - 0.07]
0.05	0.07	[0.07 - 0.14]
0.05	0.07	[0.14 - 0.21]
0.349	0.537	[0.21 - 0.747] → Big Bucket size.
0.05	0.07	[0.747 - 0.75]
0.05	0.07	[0.75 - 0.82]
0.05	0.07	[0.82 - 0.895]
0.05	0.07	[0.895 - 0.97]
0.05	0.07	[0.97 - 1.047]
0.05	0.07	[1.047 - 1.12]
0.05	0.07	[1.12 - 1.197]
0.05	0.07	[1.197 - 1.27]
0.05	0.07	[1.27 - 1.347]
0.05	0.07	[1.347 - 1.42]
0.05	0.07	[1.42 - 1.497]
0.05	0.07	[1.497 - 1.57]
0.05	0.07	[1.57 - 1.647]
0.05	0.07	[1.647 - 1.72]
0.05	0.07	[1.72 - 1.797]
0.05	0.07	[1.797 - 1.87]
0.05	0.07	[1.87 - 1.947]
0.05	0.07	[1.947 - 2.02]
0.05	0.07	[2.02 - 2.097]
0.05	0.07	[2.097 - 2.17]
0.05	0.07	[2.17 - 2.247]
0.05	0.07	[2.247 - 2.32]
0.05	0.07	[2.32 - 2.397]
0.05	0.07	[2.397 - 2.47]
0.05	0.07	[2.47 - 2.547]
0.05	0.07	[2.547 - 2.62]
0.05	0.07	[2.62 - 2.697]
0.05	0.07	[2.697 - 2.77]
0.05	0.07	[2.77 - 2.847]
0.05	0.07	[2.847 - 2.92]
0.05	0.07	[2.92 - 2.997]
0.05	0.07	[2.997 - 3.07]
0.05	0.07	[3.07 - 3.147]
0.05	0.07	[3.147 - 3.22]
0.05	0.07	[3.22 - 3.297]
0.05	0.07	[3.297 - 3.37]
0.05	0.07	[3.37 - 3.447]
0.05	0.07	[3.447 - 3.52]
0.05	0.07	[3.52 - 3.597]
0.05	0.07	[3.597 - 3.67]
0.05	0.07	[3.67 - 3.747]
0.05	0.07	[3.747 - 3.82]
0.05	0.07	[3.82 - 3.897]
0.05	0.07	[3.897 - 3.97]
0.05	0.07	[3.97 - 4.047]
0.05	0.07	[4.047 - 4.12]
0.05	0.07	[4.12 - 4.197]
0.05	0.07	[4.197 - 4.27]
0.05	0.07	[4.27 - 4.347]
0.05	0.07	[4.347 - 4.42]
0.05	0.07	[4.42 - 4.497]
0.05	0.07	[4.497 - 4.57]
0.05	0.07	[4.57 - 4.647]
0.05	0.07	[4.647 - 4.72]
0.05	0.07	[4.72 - 4.797]
0.05	0.07	[4.797 - 4.87]
0.05	0.07	[4.87 - 4.947]
0.05	0.07	[4.947 - 5.02]
0.05	0.07	[5.02 - 5.097]
0.05	0.07	[5.097 - 5.17]
0.05	0.07	[5.17 - 5.247]
0.05	0.07	[5.247 - 5.32]
0.05	0.07	[5.32 - 5.397]
0.05	0.07	[5.397 - 5.47]
0.05	0.07	[5.47 - 5.547]
0.05	0.07	[5.547 - 5.62]
0.05	0.07	[5.62 - 5.697]
0.05	0.07	[5.697 - 5.77]
0.05	0.07	[5.77 - 5.847]
0.05	0.07	[5.847 - 5.92]
0.05	0.07	[5.92 - 5.997]
0.05	0.07	[5.997 - 6.07]
0.05	0.07	[6.07 - 6.147]
0.05	0.07	[6.147 - 6.22]
0.05	0.07	[6.22 - 6.297]
0.05	0.07	[6.297 - 6.37]
0.05	0.07	[6.37 - 6.447]
0.05	0.07	[6.447 - 6.52]
0.05	0.07	[6.52 - 6.597]
0.05	0.07	[6.597 - 6.67]
0.05	0.07	[6.67 - 6.747]
0.05	0.07	[6.747 - 6.82]
0.05	0.07	[6.82 - 6.897]
0.05	0.07	[6.897 - 6.97]
0.05	0.07	[6.97 - 7.047]
0.05	0.07	[7.047 - 7.12]
0.05	0.07	[7.12 - 7.197]
0.05	0.07	[7.197 - 7.27]
0.05	0.07	[7.27 - 7.347]
0.05	0.07	[7.347 - 7.42]
0.05	0.07	[7.42 - 7.497]
0.05	0.07	[7.497 - 7.57]
0.05	0.07	[7.57 - 7.647]
0.05	0.07	[7.647 - 7.72]
0.05	0.07	[7.72 - 7.797]
0.05	0.07	[7.797 - 7.87]
0.05	0.07	[7.87 - 7.947]
0.05	0.07	[7.947 - 8.02]
0.05	0.07	[8.02 - 8.097]
0.05	0.07	[8.097 - 8.17]
0.05	0.07	[8.17 - 8.247]
0.05	0.07	[8.247 - 8.32]
0.05	0.07	[8.32 - 8.397]
0.05	0.07	[8.397 - 8.47]
0.05	0.07	[8.47 - 8.547]
0.05	0.07	[8.547 - 8.62]
0.05	0.07	[8.62 - 8.697]
0.05	0.07	[8.697 - 8.77]
0.05	0.07	[8.77 - 8.847]
0.05	0.07	[8.847 - 8.92]
0.05	0.07	[8.92 - 8.997]
0.05	0.07	[8.997 - 9.07]
0.05	0.07	[9.07 - 9.147]
0.05	0.07	[9.147 - 9.22]
0.05	0.07	[9.22 - 9.297]
0.05	0.07	[9.297 - 9.37]
0.05	0.07	[9.37 - 9.447]
0.05	0.07	[9.447 - 9.52]
0.05	0.07	[9.52 - 9.597]
0.05	0.07	[9.597 - 9.67]
0.05	0.07	[9.67 - 9.747]
0.05	0.07	[9.747 - 9.82]
0.05	0.07	[9.82 - 9.897]
0.05	0.07	[9.897 - 9.97]
0.05	0.07	[9.97 - 10.047]
0.05	0.07	[10.047 - 10.12]
0.05	0.07	[10.12 - 10.197]
0.05	0.07	[10.197 - 10.27]
0.05	0.07	[10.27 - 10.347]
0.05	0.07	[10.347 - 10.42]
0.05	0.07	[10.42 - 10.497]
0.05	0.07	[10.497 - 10.57]
0.05	0.07	[10.57 - 10.647]
0.05	0.07	[10.647 - 10.72]
0.05	0.07	[10.72 - 10.797]
0.05	0.07	[10.797 - 10.87]
0.05	0.07	[10.87 - 10.947]
0.05	0.07	[10.947 - 11.02]
0.05	0.07	[11.02 - 11.097]
0.05	0.07	[11.097 - 11.17]
0.05	0.07	[11.17 - 11.247]
0.05	0.07	[11.247 - 11.32]
0.05	0.07	[11.32 - 11.397]
0.05	0.07	[11.397 - 11.47]
0.05	0.07	[11.47 - 11.547]
0.05	0.07	[11.547 - 11.62]
0.05	0.07	[11.62 - 11.697]
0.05	0.07	[11.697 - 11.77]
0.05	0.07	[11.77 - 11.847]
0.05	0.07	[11.847 - 11.92]
0.05	0.07	[11.92 - 11.997]
0.05	0.07	[11.997 - 12.07]
0.05	0.07	[12.07 - 12.147]
0.05	0.07	[12.147 - 12.22]
0.05	0.07	[12.22 - 12.297]
0.05	0.07	[12.297 - 12.37]
0.05	0.07	[12.37 - 12.447]
0.05	0.07	[12.447 - 12.52]
0.05	0.07	[12.52 - 12.597]
0.05	0.07	[12.597 - 12.67]
0.05	0.07	[12.67 - 12.747]
0.05	0.07	[12.747 - 12.82]
0.05	0.07	[12.82 - 12.897]
0.05	0.07	[12.897 - 12.97]
0.05	0.07	[12.97 - 13.047]
0.05	0.07	[13.047 - 13.12]
0.05	0.07	[13.12 - 13.197]
0.05	0.07	[13.197 - 13.27]
0.05	0.07	[13.27 - 13.347]
0.05	0.07	[13.347 - 13.42]
0.05	0.07	[13.42 - 13.497]
0.05	0.07	[13.497 - 13.57]
0.05	0.07	[13.57 - 13.647]
0.05	0.07	[13.647 - 13.72]
0.05	0.07	[13.72 - 13.797]
0.05	0.07	[13.797 - 13.87]
0.05	0.07	[13.87 - 13.947]
0.05	0.07	[13.947 - 14.02]
0.05	0.07	[14.02 - 14.097]
0.05	0.07	[14.097 - 14.17]
0.05	0.07	[14.17 - 14.247]
0.05	0.07	[14.247 - 14.32]
0.05	0.07	[14.32 - 14.397]
0.05	0.07	[14.397 - 14.47]
0.05	0.07	[14.47 - 14.547]
0.05	0.07	[14.547 - 14.62]
0.05	0.07	[14.62 - 14.697]
0.05	0.07	[14.697 - 14.77]
0.05	0.07	[14.77 - 14.847]
0.05	0.07	[14.847 - 14.92]
0.05	0.07	[14.92 - 14.997]
0.05	0.07	[14.997 - 15.07]
0.05	0.07	[15.07 - 15.147]
0.05	0.07	[15.147 - 15.22]
0.05	0.07	[15.22 - 15.297]
0.05	0.07	[15.297 - 15.37]
0.05	0.07	[15.37 - 15.447]
0.05	0.07	[15.447 - 15.52]
0.05	0.07	[15.52 - 15.597]
0.05	0.07	[15.597 - 15.67]
0.05	0.07	[15.67 - 15.747]
0.05	0.07	[15.747 - 15.82]
0.05	0.07	[15.82 - 15.897]
0.05	0.07	[15.897 - 15.97]
0.05	0.07	[15.97 - 16.047]
0.05	0.07	[16.047 - 16.12]
0.05	0.07	[16.12 - 16.197]
0.05	0.07	[16.197 - 16.27]
0.05	0.07	[16.27 - 16.347]
0.05	0.07	[16.347 - 16.42]
0.05	0.07	[16.42 - 16.497]
0.05	0.07	[16.497 - 16.57]
0.05	0.07	[16.57 - 16.647]
0.05	0.07	[16.647 - 16.72]
0.05	0.07	[16.72 - 16.797]
0.05	0.07	[16.797 - 16.87]
0.05	0.07	[16.87 - 16.947]
0.05	0.07	[16.947 - 17.02]
0.05	0.07	[17.02 - 17.097]
0.05	0.07	[17.097 - 17.17]
0.05	0.07	[17.17 - 17.247]
0.05	0.07	[17.247 - 17.32]
0.05	0.07	[17.32 - 17.397]</

APRIL'22

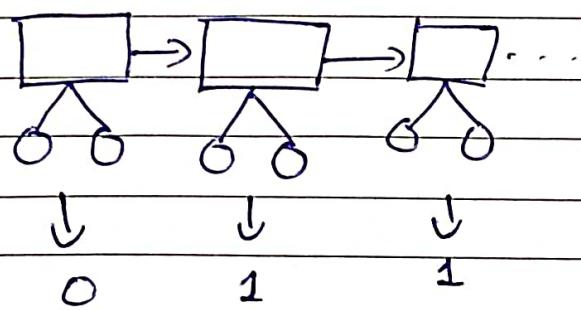
SU	MO	TU	WE	TH	FR	SA
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

084-281 • Week 13

2022 March Friday

25

- So, now random numbers b/w 0 to 1 are generated then according to the random numbers the records are selected based on bucket.
- The voting are got more bucket size, so that record will select multiple times.



- Voting - classification
- Mean - regression

The best way to find yourself is to lose yourself in the service of others. - Mahatma Gandhi

Notes:

Appointment:

Phones:

APRIL

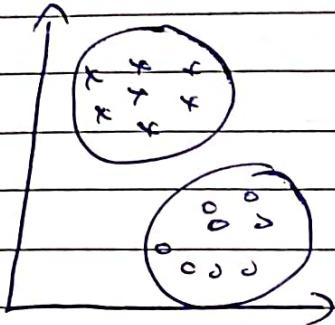
APRIL'22						
SU	MO	TU	WE	TH	FR	SA
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

086-279 • Week 13

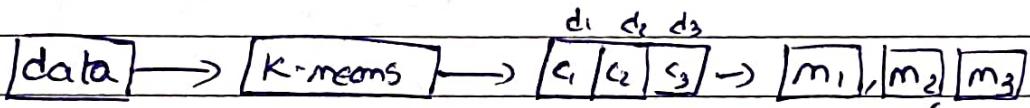
2022 March 27

Sunday

## K-means clustering (unsupervised)



Custom Ensemble technique → used in industry



→ Kmeans

K = Centroids

- ① We try different K-values →  $K=2, 3, 4, \dots$  (WCSS)
- ② Initialize K no. of centroids
- ③ Compute average to update centroid

Dreaming about being an actress, is more exciting than being one. - Marilyn Monroe

Notes:

Appointment:

Phones:

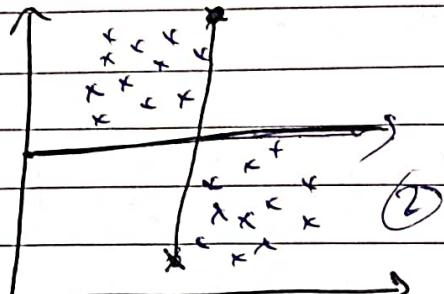
APRIL'22

SU	MO	TU	WE	TH	FR	SA
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

088-277 • Week 14

2022 March 29

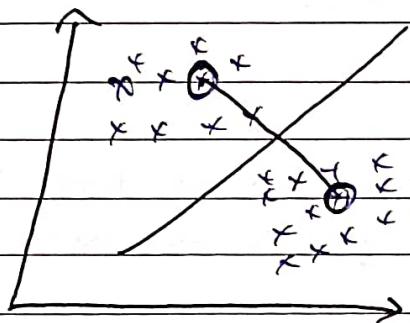
Tuesday



\* = centroid

\* Kmeans \* make sure that  
the centroids initialized at 1st  
are far away from data.

(1)

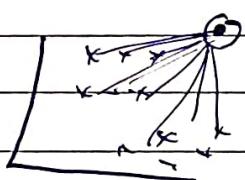
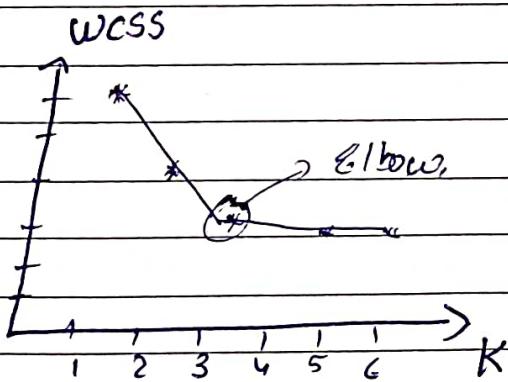


## Elbow method

Nonpoints

Distance bw

WCSS → Within cluster sum of square = [Centroid to every point]



To believe in something, and not to live it, is dishonest. - Mahatma Gandhi

Notes:

Appointment:

Phones:

APRIL 22

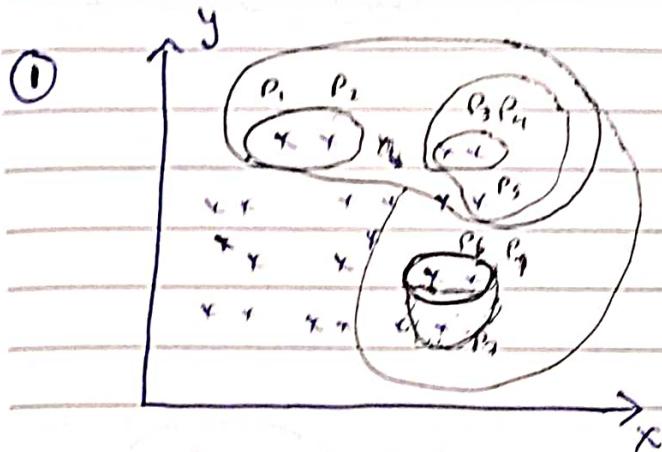
SU	MO	TU	WE	TH	FR	SA
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

090-276 • Week 14

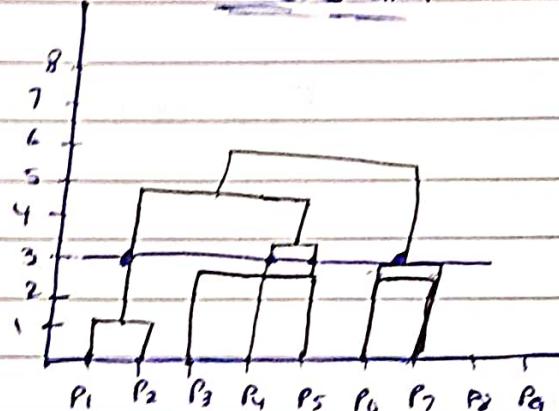
2022 March 31

Thursday

## ② Hierarchical clustering



Dendrogram



- We need to find the longest vertical line that has no horizontal line passing through it.
- Hierarchical clustering takes more time than k-means

The weak can never forgive. Forgiveness is the attribute of the strong. - Mahatma Gandhi,

Notes:

Appointment:

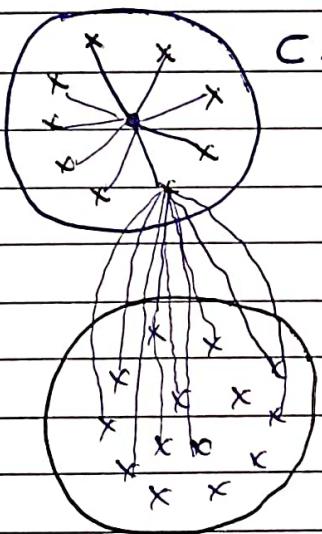
Phones:

MAY'22						
SU	MO	TU	WE	TH	FR	SA
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

091-274 • Week 14

2022 April 01  
FridayValidate clustering

→ Silhouette →  $[-1 \text{ to } +1]$   
 ↗ K-means  
 ↗ Hierarchical

 $\circ = \text{centroid}$  $a(i) = \text{distance b/w centroid f each f every point.}$  $b(i) = (\text{distance b/w a single point in } C_1 \text{ to all points in } C_2) / \text{their mean.}$ → Good model  $b(i) \gg a(i)$ 

$$a(i) = \frac{1}{|C_1| - 1} \sum_{j \in C_1} d(i, j)$$

$$b(i) = \min_{C_2} \frac{1}{|C_2|} \sum_{j \in C_2} d(i, j)$$

You are never too old to set another goal or to dream a new dream. — C. S. Lewis

Notes:

Appointment:

Phones:

MAY'22						
SU	MO	TU	WE	TH	FR	SA
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

093-272 • Week 14

2022 April 03  
Sunday

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \text{ is } |C_i| > 1$$

$$S(i) = 0 \text{ is } |C_i| = 1$$

MAY'22						
SU	MO	TU	WE	TH	FR	SA
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

0.95-270 • Week 15

2022 April 05

Tuesday

## DB Scan

→ Density Based Spatial clustering of Applications with noise

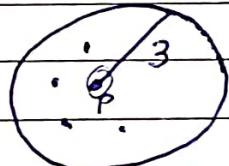
- ①  $\epsilon$ psilon
- ② min Points
- ③ Core Points
- ④ Border Points
- ⑤ noise Point

ie

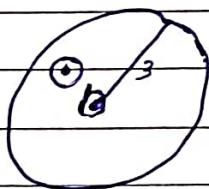
min Points = 4

$\epsilon$ psilon = radius = 3

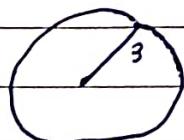
○ = core point



⇒ Within this cluster radius if there are minimum points mentioned or more than them it is considered as core point



→ If the core point condition didn't satisfy when drawn another circle then if another core point is present than it is considered as border point (b)



→ If none of conditions are satisfied than it is noise.

I think insomnia is a sign that a person is interesting. - Avery Sawyer.

Notes:

Appointment:

Phones:

MAY'22						
SU	MO	TU	WE	TH	FR	SA
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

097-268 • Week 15

2022 April 07

Thursday

### Advantages of DBScan

- Is great at separating clusters of high density vs clusters of low density within a given dataset.
- Is great with handling outliers within the dataset.

### Disadvantages of DBScan

- Does not work well when dealing with clusters of varying densities. It separates low vs high densities but finds difficult to cluster with similar density.
- Struggles with high dimensional data.

MAY'22						
SU	MO	TU	WE	TH	FR	SA
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

0.9-266 • Week 15

2022 April 09

Saturday

## XG Boost

Salary      Credit      Approval      Residual

Extreme Gradient Boosting

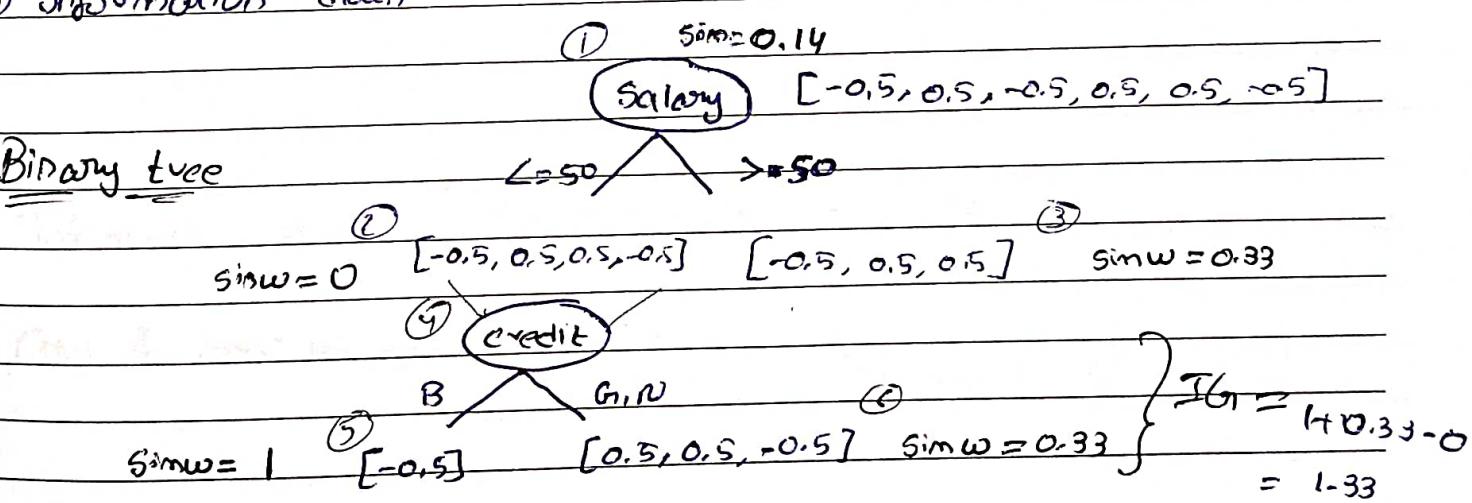
L=50	B	0	-0.5
L=50	G	1	0.5
L=50	G	1	0.5
>50	B	0	-0.5
>50	G	1	0.5
>50	N	1	-0.5

Bare model  $\rightarrow P_r = 0.5$

① Create a Binary decision tree using the feature

② Calculate Similarity weight =  $\frac{\sum (\text{Residual})^2}{\sum (P_r(1-P_r)) + \lambda}$

③ Information Gain



Never bend your head. Hold it high. Look the world straight in the eye. - Helen Keller

Notes:

Appointment:

Phones:

MAY'22						
SU	MO	TU	WE	TH	FR	SA
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

101-264 • Week 16

2022 April 11

Monday

Sum of residual square for ② (Similarity weight)

$$= \frac{[-0.5 + 0.5 + 0.5 - 0.5]^2}{0.5(1-0.5) + 0.5(1-0.5) + 0.5(1-0.5) + 0.5(1-0.5)} \\ = 0$$

Sum of similarity weight for ③

$$= \frac{(-0.5 + 0.5 + 0.5)^2}{0.5(1-0.5) + 0.5(1-0.5) + 0.5(1-0.5)} \\ = \frac{0.25}{0.75} = \frac{1}{3} = 0.33$$

Similarity weight for ①

$$= \frac{0.25}{1.75} = \frac{1}{7} = 0.14$$

$$\text{Information gain} = 0 + 0.33 - 0.14 = 0.19$$

Only for Base model.  $\log\left(\frac{P}{1-P}\right) = \log\left(\frac{0.5}{0.5}\right) = 0$

One can never consent to creep when one feels an impulse to soar. - Helen Keller.

Notes:

Appointment:

Phones:

22 20 - | - | - | 103-262 • week 20  
29 30 31 | | |

## Inferencing

4

$$[0 + \alpha(1)]$$

$$= \tilde{c} [0 + \alpha(1)]$$

$$M \in [a + \alpha_1(DT_1) + \alpha_2(DT_2) + \alpha_3(DT_3) + \dots + \alpha_n(DT_n)]$$

MAY'22						
SU	MO	TU	WE	TH	FR	SA
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

105-260 • Week 16

2022 April 15  
Friday

## ② XG Boost Regressor $\rightarrow$ 1

Exp	Grp	Salary	R <sub>i</sub>	Stdev
2	Yes	40K	-11K	
2.5	Yes	41K	-9K	
3	No	52K	1K	
4	No	60K	9K	
4.5	Yes	62K	911K	
Avg = <u>51K</u>		#		

Base model = 51K

$$[-11, -9, 1, 9, 11] \quad \frac{1}{5}$$

[cut]

$\leq 2$        $> 2$

[-11]

$$\frac{121}{1+1} = 60.5$$

[-9, 1, 9, 11]

$$\frac{(-9+1+9+11)^2}{4+1} = \frac{144}{5} = 28.5$$

It is not that I'm so smart. But I stay with the questions much longer.- Albert Einstein

Notes:

Appointment:

Phones:

MAY'22

SU	MO	TU	WE	TH	FR	SA
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

107-258 • Week 16

2022 April 17  
Sunday

$$\text{Similarity weight} = \frac{\sum (\text{residuals})^2}{\text{Nb. of Res} + x}$$

$$\text{Information Gain} = 60.5 + 28.8 - \frac{1}{6} = 89.13$$

- Just like that create for each and every row then using Information Gain see best split & use it.
- Inference

$$\text{OLP} = S_1 + \lambda_1 (-10) + \lambda_2 (DT_B) + \lambda_3 (DT_T) + \dots + \lambda_n (DT_H)$$

Don't gain the world and lose your soul, wisdom is better than silver or gold. -Bob Marley

Notes:

Appointment:

Phones:

MAY'22						
SU	MO	TU	WE	TH	FR	SA
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

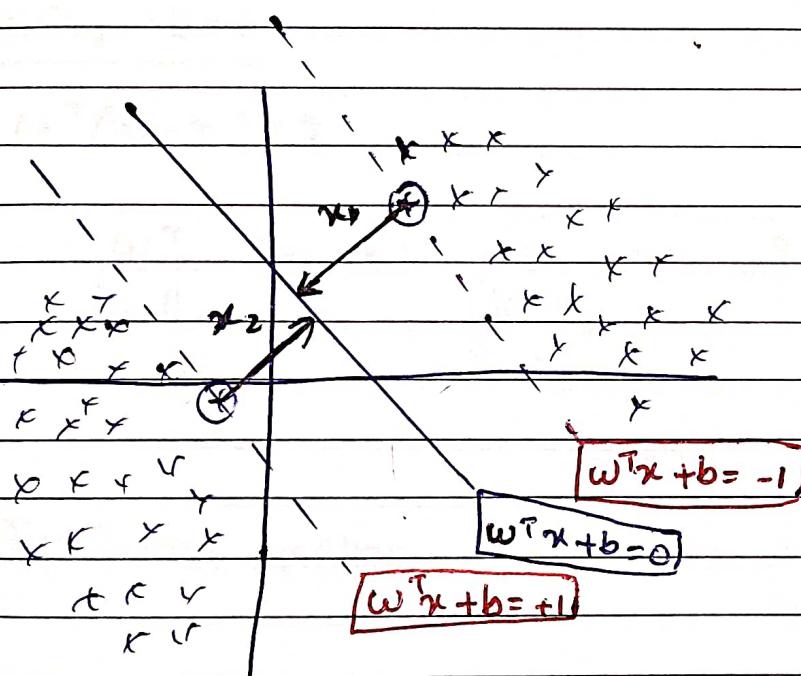
10.9-256 • Week 17

2022 April 19

Tuesday

## Support Vector machine

- ① Support vectors
- ② Hyperplanes
- ③ Marginal Distance
- ④ Linear Separable
- ⑤ Non-linear separable.



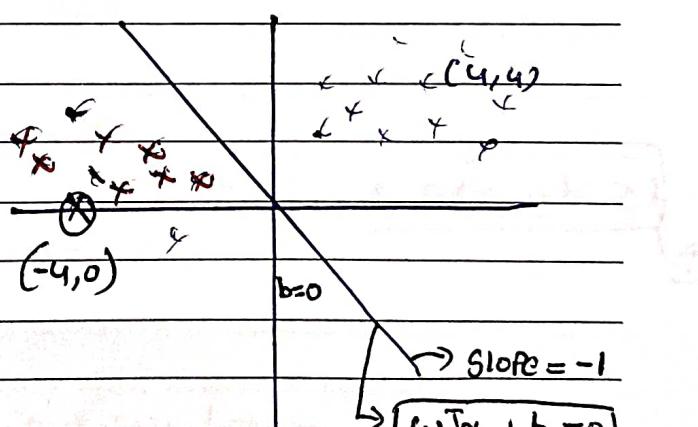
$$y = w^T x + b \rightarrow \text{eq of a line}$$

$$m = -1, b = 0$$

$$y = w^T x + 0$$

$$\begin{bmatrix} -1 \\ 0 \end{bmatrix} \begin{bmatrix} -4 & 0 \end{bmatrix} = 4 \Rightarrow \text{+ve value} \Rightarrow \text{going to be always +ve.}$$

If at first you don't succeed, try, try again. Then quit. No use being a damn fool about it. -W.C. Fields



Notes:

Appointment:

Phones:

MAY'22

SU	MO	TU	WE	FR	SA
1	2	3	4	5	6
8	9	10	11	12	13
15	16	17	18	19	20
22	23	24	25	26	27
29	30	31			

111-254 • Week 17

2022 April 21  
Thursday

$$y = w^T x$$

$$= \begin{bmatrix} -1 \\ 0 \end{bmatrix} \begin{bmatrix} 4 \\ 4 \end{bmatrix} = -4 \Rightarrow -ve \Rightarrow \text{These side points are going to be always -ve.}$$

$$\rightarrow w^T x_1 + b = -1$$

$$\rightarrow w^T x_2 + b = 1$$

$$w^T(x_2 - x_1) = 2$$

$$\|w\| = \text{norm of } w$$

$$\frac{w^T}{\|w\|} (x_2 - x_1) = \frac{2}{\|w\|} \text{ maximize} \quad \text{by dividing } w^T \text{ by } \|w\| \\ \text{whole magnitude will go off.} \\ \text{by direction don't care.}$$

$$\text{Optimization function } (w, b) \max \frac{2}{\|w\|}$$

Such that

$$y_i \left\{ \begin{array}{l} 1 \quad w^T x_i + b \geq 1 \\ -1 \quad w^T x_i + b \leq -1 \end{array} \right\} \Rightarrow y_i * w^T x_i + b_i \geq 1 \quad (1)$$

 $\rightarrow \text{If (1) is not } \geq 1 \text{ then it is misclassification.}$ 

Do what you can with all you have, wherever you are. - Theodore Roosevelt

Notes:

Appointment:

Phones:

MAY'22						
SU	MO	TU	WE	TH	FR	SA
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

113-252 • Week 17

2022 April 23  
Saturday

$$(w, b) = \min \frac{\|w\|}{2} + C_i \sum_{i=1}^n \xi_i$$

$\xi_i \rightarrow$  How many errors?

$\{ \rightarrow$  Value of error

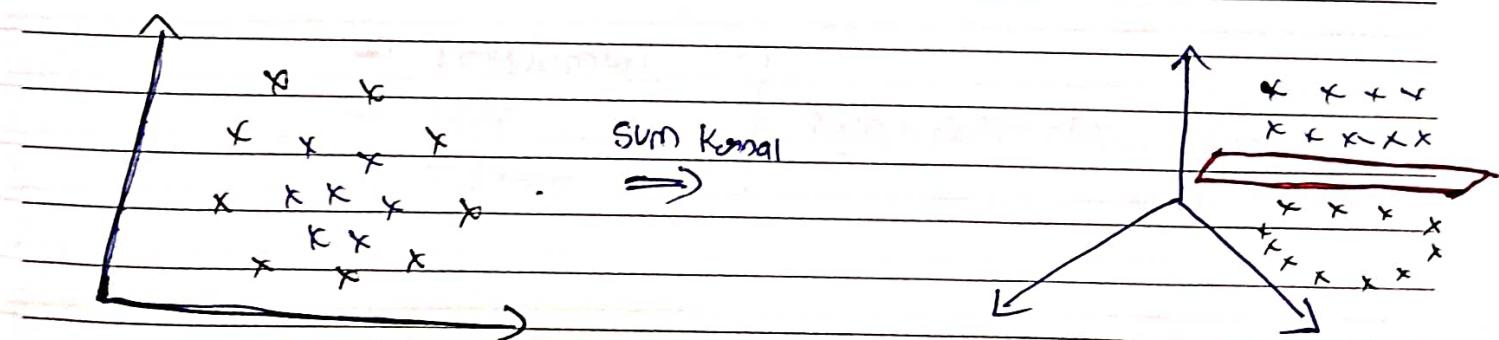
→ If some points are present in bw margin & line on Hypothesis So those are errors.

→  $C_i \rightarrow$  How many errors we can have so that the margin doesn't change  
(Regularization)  $\uparrow C_i$

## SUM KERNELS

① Soft margin  $\rightarrow$  contains errors

② Hard margin  $\rightarrow$  contains no points in bw.



Reading is to the mind, as exercise is to the body. - Brian Tracy

Notes:

Appointment:

Phones:

MAY'22

SU	MO	TU	WE	TH	FR	SA
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

115-250 • Week 18

2022 April 25  
Monday

## Polynomial Kernel

$$d(x_1, x_2) = (x_1^T \cdot x_2 + 1)^d$$

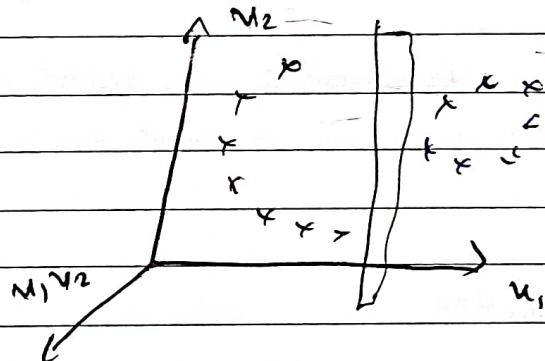
d=dimension

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \cdot \begin{bmatrix} x_1 & x_2 \end{bmatrix} = \begin{bmatrix} x_1^2 & x_1 \cdot x_2 \\ x_1 \cdot x_2 & x_2^2 \end{bmatrix}$$

Outline

So,

$$\begin{array}{c|c|c|c|c|c|c|c} x_1 & x_2 & y_0 & x_1^2 & x_2^2 & x_1 \cdot x_2 & \dots & \dots \end{array}$$



→ 3 types of kernels

- Polynomial
  - RBF
  - Sigmoid
- } hyper parameters tuning

Old Man's Advice to Youth: 'Never Lose a Holy Curiosity.' - Albert Einstein

Notes:

Appointment:

Phones:

MAY'22						
SU	MO	TU	WE	TH	FR	SA
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

117-248 • Week 18

2022 April 27

Wednesday

## Feature Engineering

→ Different types of encoding.

Nominal encoding

- Nominal → one hot
- one hot (KDD orange)
- Mean encoding

Ordinal → Label Encoding

→ Target Guided ordinal encoding

Target Guided ordinal → (1) mean based on categories.  
 → (2) give rank based on highest mean.

Mean encoding → (1) same as 'target guided' but instead of rank we replace mean values.

The fault, dear Brutus, is not in our stars, but in ourselves. - William Shakespeare

Notes:

Appointment:

Phones:

MAY'22						
SU	MO	TU	WE	TH	FR	SA
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

119-246 • Week 18

Ozby

2022 April 29  
Friday

→ why feature scaling?

Height    weight    BMI

Features (have)

→ magnitude  
→ units

→ low computation

→ Faster convergence

→ Handling missing values in categorical variable

① Delete Rows

② Replace with most frequent values

③ Apply classifier algorithm and Predict

④ Apply unsupervised ML

⑤ mean / median

⑥ Regression imputation, interpolation, extrapolation

⑦ Arbitrary imputation

⑧ Frequent category imputation

⑨ End of distribution imputation

⑩ Capturing N/A values with new features

} Only for

} Categorical

Happiness does not come from without, it comes from within. - Helen Keller

Notes:

Appointment:

Phones:

JUNE '22						
SU	MO	TU	WE	TH	FR	SA
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

121-244 • Week 19

2022 May 01

Sunday

MAY

→ How to handle categorical features many categories.

→ Count / Frequency encoding.

→ The Problem is also called as High Cardinality.

⇒ Count / Frequency encoding

→ Replace the categorical variables by its count.

### Advantages :-

- ① Very simple to implement
- ② Does not increase the feature dimensional space

### Disadvantages :-

- ① If some of the labels have same count
  - ② Adds some arbitrary numbers, and ∵ weights to be different labels, that may not be related to their predictive power.
- ⇒ consider example of Sonar dataset.

JUNE

JULY

AUGUST

I have no special talents. I am only passionately curious. - Albert Einstein SMART

Notes:

Appointment:

Phones:

JUNE'22						
SU	MO	TU	WE	TH	FR	SA
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

188-242 • Week 10

2022 May 03  
Tuesday

→ Life cycle of a data science projects.

## ① Data collection:-

- From company side
- 3rd party API's
- Surveys

## ② Feature Engineering ... //

- Handling the missing values.

Different types of missing Data?

→ missing Completely at random.

→ If the Probability of being missing is same for all observations.  
→ when data is MCAR, there is no relationship between the data missing and any other values, observed or missing.

→ missing Data not at random. - Systematic missing values.

→ When the missing value have no relation with any other column but have relation with the column in which the value is missing.

Male men - hide their salary  
woman - hide their age.

Imagine your life is perfect in every respect; what would it look like? - Brian Tracy

Notes:	Appointment:	Phones:



JUNE

JULY

AUGUST