Date __/__/__

Ex-1 Time series Data (forecasting the stock index)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X
	Price of Petrol	Price of Gold	Conflict b/w neighbour country	Trap (Y/No)	Previous Day Year value	Mid copy	Fixed asset	$\bar{x}$
x_1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_1
x_2								
x_3								
x_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}	x_{i7}	y_i
1								
x_N								
x_{new}	$x_{new,1}$	$x_{new,2}$	$x_{new,3}$				x_{new7}	y_{new} ?

$y \in \mathbb{R}$

Ex-2: (Time Series) : Predicting Rainfall in a particular locality. (Regression)

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	y_{mm} (amount of Rainfall in 3rd week of Jan)
2001								y_1
2002								y_2
2003								y_3
⋮								⋮
2024								y_{24}

$y_{25} = ?$

$y_{20} = ?$

Mathematical formulation of this Problem

Problem Statement:

Given $D = \{x_i, y_i\}_{i=1}^{N=24}$, $x_i \in \mathbb{R}^{P=8}$, $y_i \in \mathbb{R}$

Predict y_{25} for x_{25} where $x_{25} \neq x_i \; \forall i = 1$ to 24

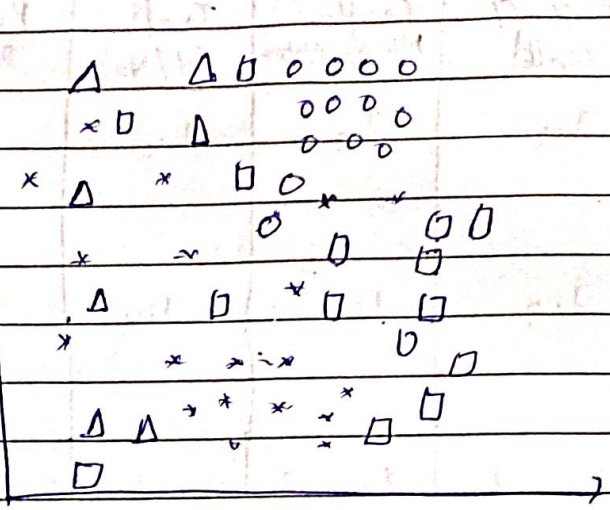
W Date __/__/__

VVI

KNN Classification

↳ K Nearest Neighbour

ex:



Δ : class 1
○ : class 2
* : class 3
□ : class 4

$$d_1 = \sqrt{\sum_{i=1}^{N} (x_1 - x_{ij})^2}$$

$$d_2 = \sqrt{\sum_{i=1}^{N} (x_2 - x_{ij})^2}$$

!

$$d_N = \sqrt{\sum_{i=1}^{N} (x_N - x_{ij})^2}$$

K odd for proper distribution i.e. K = 5 (3 & 9).

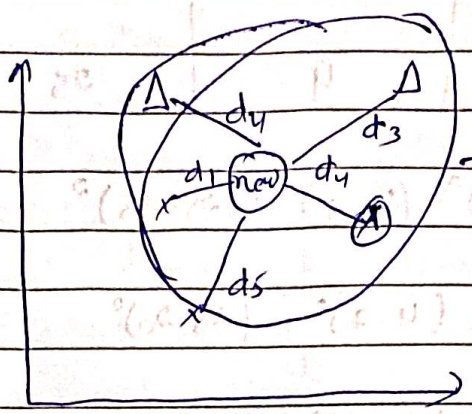
	X_1	X_2	X_3	...	X_{10}	Y
x_1	x_{11}	x_{12}	x_{13}		$x_{1,10}$	y_1
x_2						y_2
x_3						y_3
⋮						⋮
x_{100}	$x_{100,1}$	$x_{100,2}$	$x_{100,3}$		$x_{100,10}$	y_{100}
new sample x_{new}	$x_{new,1}$	$x_{new,2}$	$x_{new,3}$	- -	$x_{new,10}$	$y_{new} = ?$

$$d_1 = \sqrt{(x_{new,1} - x_{11})^2 + (x_{new,2} - x_{12})^2 + (x_{new,3} - x_{13})^2 + \cdots + (x_{new,10} - x_{10})^2}$$

N = 100

P = 10

Date _/_/_



a

The following table represents the training samples. a factory produces a new paper il that pass laboratory with acid durability. Shor.. Strength is 4 kg per sq m & cost is 25 Rs per kg. apply KNM - classification of with K = 5 to predict the quality of the tissue papper from the given data set.

x_1 Acid durability (hrs)	x_2 Strength (kg/m^2)	x_3 cost ₹ Pure/kg	y opinion from People.
x_1 → 8	6	15	Good
x_2 → 9	7	20	bad
x_3 → 2	5	18	bad
x_4 → 1	3	9	bad
x_5 → 6	7	8	Good
x_6 → 4	5	10	bad
x_7 → 5	5	6	Good
x_8 → 3	6	10	Good

Date __/__/__

X_{new}	5	4	25	?

$d_1 = \sqrt{(5-8)^2 + (4-6)^2 + (25-15)^2} = \sqrt{9+4+100} = \sqrt{113}$

$d_2 = \sqrt{(5-9)^2 + (4-7)^2 + (25-20)^2} = \sqrt{16+9+25} = \sqrt{50}$

$d_3 = \sqrt{(5-2)^2 + (4-5)^2 + (25-18)^2} = \sqrt{9+1+49} = \sqrt{59}$

$d_4 = \sqrt{(5-1)^2 + (4-3)^2 + (25-9)^2} = \sqrt{16+1+256} = \sqrt{273}$

$d_5 = \sqrt{(5-6) + (4-7)^2 + (25-8)^2} = \sqrt{1+} = \sqrt{299}$

$d_6 = \sqrt{(5-4)^2 + (4-5)^2 + (25-10)^2} = = \sqrt{227}$

$d_7 = \sqrt{(5-5)^2 + (4-5)^2 + (25-6)^2} = = \sqrt{362}$

$d_8 = \sqrt{(5-3)^2 + (4-6)^2 + (25-10)^2} = = \sqrt{233}$

$d_9 = \sqrt{(5-2)^2 + (4-7)^2 + (25-8)^2} = = \sqrt{307}$.

step-2 : after arrange the distance in non-decreasing order

$d_2 < d_3 < d_1 < d_6 < d_8$

∴ The neighbour are = x_2, x_3, x_1, x_6, x_8

bad bad Good bad Good

3 → bad

2 → Good

as no. of bad quality = 3 > No. of Good Quality

∴ New item is of Bad Quality □

Date __/__/__

Q. 2. A Mobile manufacturing Company has manufactured a new mobile asset. the Company want to push adds to potential usser to buy new mobile. the Company has data set that Contains multiple through the social network profile the data set Contains lots of information about the ussers but the approx. Sallary, age, and amount of time spent with the mobile for internet are the 3 key predictors to determine the potential Candidate. the target is the purposes item ~~buyer~~ earlier of similar type.
apply KNN to predict whether a new Usser is a potential Candidate for the Add or not.

Q. Write a program in python to cleanse the data set, determine the priciple components, plot the graph & Predict the class for any new items.

	X_1 User ID	X_2 MC ID	X_3 age	X_4 Gender	X_5 Sabry	X_6 PIN code	X_7 time stay log with mobile with internet	Y Purchase item of similar type earlier
x_1	1509784	204.192.1.3	32	M	50K	751024	2.5	Yes.
x_2	1508788	1.1.1.32	45	F	60K	751021	1	No
x_3	201569	20.20.20.3	19	F	5K	751001	5	No
x_4	192036	192.168.1.3	21	M	30K	621569	3	No
x_5	2016167	11.20.32.69	27	M	45K	321992	2	Yes.
x_6	172962	192.163.15	25	F	30K	552693	5	Yes
x_7	281969	10.6.42.10	17	F	50K	456921	4	No
X_{New}	242592	92.10.3.41	22	M	70K	681024	3.4	?

K=3

Date __/__/__

- determination of a potential candidate to purchase a new mobile depends upon, age, salary & time spend

$d_1 = \sqrt{(20-32)^2 + (70-50)^2 + (3.4-2.5)^2} =$

$d_2 = \sqrt{(20-45)^2 + (70-60)^2 + (3.4-1)^2} =$

$d_3 = \sqrt{(20-19)^2 + (70-5)^2 + (3.4-5)^2} =$

$d_4 = \sqrt{(20-24) + (70-30)^2 + (3.4-3)^2} =$

$d_5 = \sqrt{(20-27)^2 + (70-45)^2 + (3.4-2)^2} =$

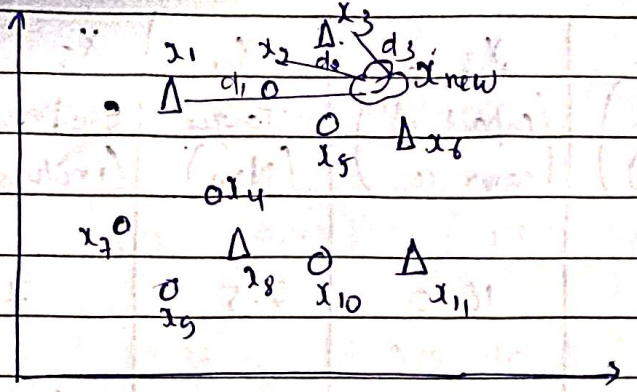
$d_6 = \sqrt{(20-25)^2 + (70-30)^2 + (3.4-5)^2} =$

$d_7 = \sqrt{(20-17)^2 + (70-50)^2 + (3.4-4)^2} =$

Date __/__/__

15-Jan

KNN Regression:



	X_1	X_2	X_3	X_4	X_5
y_1					
y_2					
y_3					
⋮					
y_{11}					
x_{new}	?	$x_{new,2}$	$x_{new,3}$	$x_{new,4}$	$x_{new,5}$
z_{new}	$z_{new,1}$	$z_{new,2}$	?	$z_{new,4}$	$z_{new,5}$
t_{new}	$t_{new,1}$	$t_{new,2}$	$t_{new,3}$	$t_{new,4}$	?

- i) Compute the distances from new-sample to the given samples.
  Let the distances are: $d_1, d_2, d_3 \dots, d_{11}$

ii) Arrange the distances in non-decreasing order for N-neighbours.

Let $k = 3$ (given)

$d_5 < d_6 < d_3$

iii) The 3 nearest neighbours are: x_5, x_6 & x_3

iv) ∴ target of $x_{new} = \frac{x_{51} + x_{61} + x_{31}}{3}$

z_{new}

Date __/__/__

Problem:

K = 3.

	X_1 (Mileage in km/l)	X_2 (Mean eco-speed Km/h)	X_3 (capacity of engine cc)	X_4 Torque (Nm)	X_5 Suspension (inches)
x_1 →	15.2	80	1600	2.5	7
x_2 →	18.4	95	1520	3.4	10
x_3 →	20.4	98	1400	3.0	8
x_4 →	25.2	84	1360	2.8	6.5
x_5 →	22.8	76	1440	4.8	9
Q1.	18.4		1480	7.2	8
Q2.	19.5	85	1600	?	9.2.

$$d_1 = \sqrt{(18.4-15.2)^2 + (1480-1600)^2 + (7.2-2.5)^2 + (8-7)^2}$$

$$= \sqrt{10.24 + 14,400 + 22.09 + 1} \Rightarrow \sqrt{14433.33}$$

$$d_2 = \sqrt{(18.4-18.4)^2 + (1480-1520)^2 + (7.2-3.4)^2 + (10-8)^2}$$

$$= \sqrt{0 + 1600 + 14.44 + 4} = \sqrt{1618.44}$$

$$d_3 = \sqrt{(18.4-20.4)^2 + (1480-1400)^2 + (7.2-3.0)^2 + (8-8)^2}$$

$$= \sqrt{4 + 6400 + 17.64 + 0} \Rightarrow \sqrt{6421.64}$$

Date __/__/__

$d_4 = \sqrt{(18.4-25.2)^2 + (1480-1360)^2 + (7.2-2.8)^2 + (8-6.5)^2}$

$= \sqrt{46.24 + 14,400 + 19.36 + 2.25} = \sqrt{14467.85}$

$d_5 = \sqrt{(18.4-22.8)^2 + (1480-1440) + (7.2-4.8)^2 + (8-9)^2}$

$= \sqrt{19.36 + 1600 + 5.76 + 1} = \sqrt{1626.12}$

arranging the distance in non-decreasing order.

$d_2 < d_5 < d_3$

∴ The nearest neighbour are: x_2, x_5, x_3.

$y_{new} = \frac{95 + 98 + 76}{3} \simeq 89.6.$

2. Q. $d_1 = \sqrt{(19.5-15.2)^2 + (85-80)^2 + (1600-1600)^2 + (9.2-7)^2}$

$= \sqrt{48.33}$

$d_2 = \sqrt{(19.5-18.4)^2 + (85-95)^2 + (1600+1520) + (10-9.2)^2}$

$= \sqrt{1.21 + 100 + 6400 + 0.64} = \sqrt{6501.85}$

$d_3 = \sqrt{(19.5-20.4)^2 + (85-98)^2 + (1600-1400)^2 + (9.2-8)^2}$

$= \sqrt{1.21 + 169 + 40,000 + 1.44} = \sqrt{40,171.85}$

$d_4 = \sqrt{(19.5-25.2)^2 + (85-84)^2 + (1600-1360)^2 + (9.2-6.5)^2}$

$= \sqrt{32.49 + 1 + 57,600 + 7.29} = \sqrt{57,640.78}$

Date __/__/__

$$d_5 = \sqrt{(19.5 - 22.8)^2 + (85 - 76)^2 + (1600 - 1440)^2 + (9.2 - 9)}$$

$$\sqrt{10.89 + 81 + 25600 + 0.04} \Rightarrow \sqrt{25691.93}$$

after array the distance in non-decreasing order.

$$\Rightarrow d_1 < d_2 < d_5$$

The three nearest neighbour are x_1, x_2, x_5

$$\Rightarrow Y = \frac{2.5 + 3.4 + 4.8}{3} \qquad \Rightarrow \frac{x_{14} + x_{24} + x_{54}}{3}$$

$$\Rightarrow \frac{10.7}{3} = 3.566$$

Date: _/_/_

18-Jan

Simple Linear Regression

Problem statement:

Given a dataset $D = \{x_i, y_i\}_{i=1}^{N}$ $\quad x_i \in \mathbb{R}$, $y_i \in \mathbb{R}$

Then for any new sample $x_{new} \neq x_i \ \forall i = 1$ to N predict the target $y_{new} = ?$

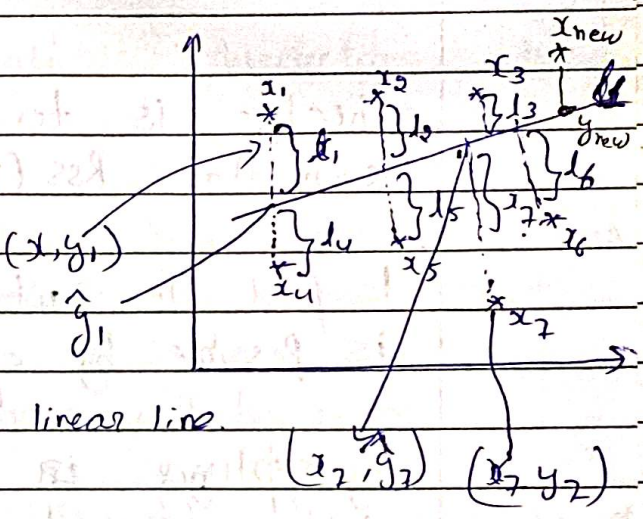
Explanation:

Now as it is a linear regression model, we can represent it as:

$$y_i = \beta_0 + \beta_1 x_i \longrightarrow (1)$$

where: β_1 → slop of the linear line.

β_0 → intercept



As we have to predict:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \longrightarrow (2)$$

where $\hat{\beta}_0, \hat{\beta}_1$ are the estimate of β_0 & β_1 & $\hat{y}_i$ is the estimate of y_i.

- i.e we have to estimate the two unknown $\hat{\beta}_0$ & $\hat{\beta}_1$. ie to determine the ~~co~~ suitable $\hat{\beta}_0$ & $\hat{\beta}_1$, there are several methods. However, we may use RSS (Residual sum square) error method to estimate $\hat{\beta}_0$ & $\hat{\beta}_1$.

i.e $RSS(\beta) = e_1^2 + e_2^2 + e_3^2 + \cdots + e_N^2$

$$e_1^2 = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_i - \hat{y}_i)^2 + \cdots + (y_N - \hat{y}_N)^2$$

Date __/__/__

$$= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$+ \cdots + (y_N - \hat{\beta}_0 - \hat{\beta}_1 x_N)^2$$

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \quad \text{---------- (3)}$$

objective is to determine $\hat{\beta}_0$ & $\hat{\beta}_1$ such that RSS (β) will be minimum

to find the suitable value RSS (β) will be minimum. is possible by appling ~~application~~ of derivatives.

by appling ~~in~~ the calculus of derivative w.r.t to β we have ~~$\hat{\beta}_1 =$~~

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{j=1}^{N} (x_i - \bar{x})^2} \quad \text{---------- (4a)}$$

and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ ---------- (4b)

Now, by using the data set we can estimate $\hat{\beta}_1$

D:

	Predictor	Target
	x_1	X
observation x_1	x_{11}	y_1
observation x_2	x_{21}	y_2
observation x_3	x_{31}	y_3
⋮	⋮	⋮
observation x_N	x_{N1}	y_N
x_{new}	x_{new1}	?

Date __/__/__

Then substituting $\hat{\beta}_1$ in Eq - (4b) we can estimate $\hat{\beta}_0$.

⇒ finally we can estimate $\hat{y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$ Ans

- Limitation of Simple linear equation module:

1) for any sample with multiple predictors this method is not suitable.

2) this method assumes that there exist a linear relationship b/w the predictor & the target.

→ However if Not then error will be more

* Solution 1): to over come ~~the difficulty of~~ the limitation 1 multiple linear equation module may be use

→ to over come limitation 2. any Non linear regration module is ussed. for ex: SVM (support vector Machine).

6) Estimate the sales amount of an item from the given data.

x	y
Annual spend on advertisement (₹)	Amount of sale / Turnover (₹)
16850	2050658
30783	6088372
12682	5500678
5675	8232824
27305	4660419
42095	?

Date __/__/__

$$RSS(\beta) = (Y - X\beta)^T (Y - X\beta) \quad \text{---- (1)} \qquad RSS(\beta) \in \mathbb{R}$$

We have to find 1st order derivative
- we have to differentiate eq-(1) w.r.t. β

Let $z = Y - X\beta$

$z^T \Rightarrow (Y - X\beta)^T$

$\therefore RSS(\beta) = z^T z$

$$\Rightarrow \frac{\partial}{\partial \beta}(RSS(\beta)) = \frac{\partial}{\partial \beta}(z^T z) \quad \text{---- (1A)}$$

for Refences:

Case:1

$$\frac{\partial y}{\partial \beta} = \begin{bmatrix} \frac{\partial y}{\partial \beta_0} \\ \frac{\partial y}{\partial \beta_1} \\ \vdots \\ \frac{\partial y}{\partial \beta_p} \end{bmatrix}$$

y: Scalar
β: vector

Case:2

$$\frac{\partial Y}{} = \begin{bmatrix} \frac{\partial Y_1}{\partial \beta_0} & \frac{\partial Y_2}{\partial \beta_0} & \frac{\partial Y_3}{\partial \beta_0} & \cdots & \frac{\partial Y_N}{\partial \beta_0} \\ \frac{\partial Y_1}{\partial \beta_1} & \frac{\partial Y_2}{\partial \beta_1} & \frac{\partial Y_3}{\partial \beta_1} & \cdots & \frac{\partial Y_N}{\partial \beta_1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \frac{\partial Y_1}{\partial \beta_i} & \frac{\partial Y_2}{\partial \beta_i} & \frac{\partial Y_3}{\partial \beta_i} & \cdots & \frac{\partial Y_N}{\partial \beta_i} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \frac{\partial Y_1}{\partial \beta_p} & \frac{\partial Y_2}{\partial \beta_p} & \frac{\partial Y_3}{\partial \beta_p} & \cdots & \frac{\partial Y_N}{\partial \beta_p} \end{bmatrix}$$

Y: vector $\in \mathbb{R}^N$
β: vector $\in \mathbb{R}^p$

Date __/__/__

Case: 3

if A : matrix
x : vector.

Y	$\frac{\partial Y}{\partial x}$	
Ax	A^T	(5A)
x^TA	A	(5B)
x^Tx	$2x$	(5C)
x^TAx	$Ax + A^Tx$	(5D)

Further if $(y = f(x) ; u = g(x))$ - $\frac{dy}{dx} = \frac{\partial y}{\partial u} \cdot \frac{du}{\partial x}$

Chain Rule

if Case of matrix / vector, we have

$z = f(\beta)$, $RSS = \phi(z)$

$\therefore \frac{\partial}{\partial \beta}(RSS) = \frac{\partial z}{\partial \beta} \cdot \frac{\partial}{\partial z}(RSS)$

Now from eq (4A) we have

$$\frac{\partial}{\partial \beta} RSS(\beta) = \frac{\partial}{\partial \beta}(z^Tz)$$

$$= \frac{\partial z}{\partial \beta} \cdot \frac{\partial}{\partial z}(z^Tz)$$

$$= \frac{\partial}{\partial \beta}(Y - X\beta) \cdot 2z \quad (\because \text{by eq 5C})$$

Date __/__/__

$$= \frac{\partial Y}{\partial \beta} \cdot 2z - \frac{\partial}{\partial \beta}(X\beta) \cdot 2z$$

$$= 0 - X^T \cdot 2(Y - X\beta) \quad [\because \text{by eq-5A.}]$$

$$\Rightarrow \frac{\partial}{\partial \beta}(RSS(\beta)) = -2X^TY + 2X^TX\beta = 0 \quad [\because \text{for minimization}]$$

$$\Rightarrow 2X^TX\beta = 2X^TY$$

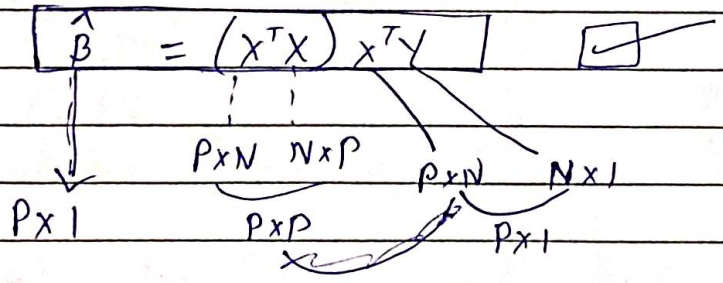
$$\Rightarrow X^TX\beta = X^TY$$

$$\Rightarrow (X^TX)^{-1}(X^TX)\beta = (X^TX)^{-1}X^TY$$

[$\because$ by Multiplying $(X^TX)^T$ on B.S]

$$\Rightarrow I\beta = (X^TX)^{-1}X^TY$$

$$\Rightarrow \boxed{\hat{\beta} = (X^TX)X^TY}$$



Algorithm (Multiple linear Regression)

1. Input: Dataset $= \{x_i, y_i\}_{i=1}^{N}$ $x_i \in \mathbb{R}^P$, $y_i \in \mathbb{R}$

and new sample $x_{new} \neq x_i$, $\forall i = 1$ to N

2. Compute $\hat{\beta} = (X^TX)^{-1}X^TY$

3. formulate $\tilde{x} = \begin{bmatrix} 1 \\ x_{new} \end{bmatrix} = \begin{bmatrix} 1 \\ x_{new,1} \\ \vdots \\ x_{new,P} \end{bmatrix}$

Date __/__/__

4) estimate $y_{new} = \hat{\beta}^T \tilde{x}$

$1 \times (p+1)$ $\quad$ $(p+1) \times 1$

1×1 = Real No.

y_{new} is a no.

$y_{new} \in \mathbb{R}$

25/01/24 Thu

PN: 136 ex-3

Q.

X_1 = GPA

X_2 = IQ

X_3 = level

X_4 = interaction b/w GPA & IQ

X_5 = interaction b/w GPA & level

$\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$

$\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$

(a) i)

Ans =>

X_1 (GPA)	X_2 IQ	X_3 (level)	X_4 1: College 0: H.S	X_5 GPA × IQ	X_5 GPA × level	Y Starting salary in thousands
$\beta_0 = 50$						
$\beta_1 = 20$						
$\beta_2 = 0.07$						
$\beta_3 = 35$						
$\beta_4 = 0.01$						
$\beta_5 = -10$						

a) $y_{\text{new college}} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5$

GPA: G = 50 + 20G + 0.07·IQ + 35×1 + 0.01·GIQ + (-10) G×1

IQ: Q =

$\Rightarrow 50 + 20G + 0.07Q + 35 + 0.01GQ - 10G$

$\Rightarrow 85 + 10G + 0.07Q + 0.01GQ$

$y_{\text{new high school}} \Rightarrow 50 + 20G + 0.07Q + 35 \times 0 + 0.01 \times GQ - 10G \times 0$

$= 50 + 20G + 0.07Q + 0 + 0.01GQ - 0$

$= 50 + 20G + 0.07Q + 0.01GQ$

$\Rightarrow$ College value is more.

Date __/__/__

College: $85 + 10G$

High School: $50 + 20G$

$\Rightarrow$ $85 + 10G = 50 + 20G$

$85 - 50 = 20G - 10G$

$35 = 10G$

$G = 3.5$

b) $50 + 20G + 0.07Q + 35 + 0.01GQ$

$50 + 20 \times 4 + 0.07 \times 110 + 35 + 0.01 \times 4 \times 110$

$50 + 80 + 7.7 + 35 + 0.0$

$\Rightarrow G = 4$ & $Q = 110$

$85 + 10G + 0.07Q + 0.01GQ$

$85 + 10 \times 4 + 0.07 \times 110 + 0.01 \times 110 \times 4$

$\Rightarrow 85 + 40 + 7.7 + 4.4 = 137.1$

$\Rightarrow 137.1$

Line 2 = $y = a + bx$

Line 3 = $z = a + bx + cy$

Im. u = $k = a + bx + cy + dz$

$y_0 = \beta_0 + \beta_1 x_0 + \beta_2 x_2 + \beta_3 x_2$

$\beta_4 x_4 + \beta_5 x_0$

$y_{p+1} = \beta_1 x_1 + \beta_4 x_4 + \beta_p x_p$

demote (P+1) - value P

Date __/__/__

limitation of least square regration

- We the predictors & the target are not linear related so application of least square regration

least square meth = $\hat{y}_{reg} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_p x_p$

- The accuracy of model is not as desired Hence accuracy.
the result obtained by those models are not accepted - To improve the performance of the model we may regularise the parameters (β-values) obtained.

- in this situation, Here may exist some unnecessary prediction which affect the performance of the model. So by shrinking the impact of those predictors, the performance can be improve.

- there are several methods for this..
i) Ridge regression / Ridge regularisation.
ii) LASSO Regression / Regularisation

Date _/_/_

RIDGE Regression (Shrinkage Method)

Analytics: To reduce the impact of unnecesary predictors Ridge regression is usse d.
- it is one of the shrinkage method

Given: $D = \{x_i, y_i\}_{i=1}^{N}$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$

Predict y_{new} for $x_{new} \neq x_i$; $i = 1$ to N

with desired accuracy.

Now,

Step 1: Given matrix : $X =$

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ \vdots & & & & \\ x_{i1} & x_{i2} & x_{i3} & \cdots & x_{ip} \\ \vdots & & & & \\ x_{N1} & x_{N2} & x_{N3} & \cdots & x_{Np} \end{bmatrix}$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_N \end{bmatrix}$$

Date __/__/__

→ Step 2: Compute $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots \bar{x}_p$

by $\bar{x}_j = \frac{\sum_{i=1}^{N} x_{ij}}{N}$

and subtract $\bar{x}_j$ from corresponding element of that column.

$$\therefore \text{New } X = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \ldots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & & x_{2p} - \bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{i1} - \bar{x}_1 & x_{i2} - \bar{x}_2 & & x_{ip} - \bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{N1} - \bar{x}_1 & x_{N2} - \bar{x}_2 & & x_{Np} - \bar{x}_p \end{bmatrix}$$

→ Step 3: Compute.

$$\beta_0 = \frac{\sum_{j=1}^{N} y_i}{N}$$

→ Step 4: Then the updated $Y = \begin{bmatrix} y_1 - \beta_0 \\ y_2 - \beta_0 \\ \vdots \\ y_N - \beta_0 \end{bmatrix} = \begin{bmatrix} y_1 - \beta_0 & y_2 - \beta_0 & \ldots & y_N - \beta_0 \end{bmatrix}^T$

→ Step 5: Now we have to determine β by minimizing the below function.

$$\text{Ridge } RSS(\beta) = \underbrace{\sum_{j=1}^{N} \left[y_i - \left[\beta_0 + \sum_{j=1}^{p} \beta_i x_i \right] \right]}_{\text{Multiple Linear Regression.}} + \lambda \underbrace{\sum_{j=1}^{p} \beta_j^2}_{\text{Regularizer}}$$

λ is to be selected judiciously.

(K fold Cross Validation is used to estimate λ)

Date __/__/__

So by differentiable eq (*) for minimization

Ridge

$$\Rightarrow \frac{\partial RSS(\beta)}{\partial \beta} = -2X^T(Y - X\beta) + \lambda 2\beta$$

$$\Rightarrow \frac{\partial RSS(\beta)}{\partial \beta} = -2X^T(Y - X\beta) + \lambda 2\beta = 0 \quad (\because \text{Minimize})$$

$$\Rightarrow -2X^T(Y - X\beta) + 2\lambda\beta = 0$$

$$\Rightarrow -2X^TY + 2X^TX\beta + 2\lambda\beta = 0$$

$$\Rightarrow 2(X^TX + \lambda I)\beta = 2X^TY$$

$$\Rightarrow (X^TX + \lambda I)\beta = X^TY$$

$$\Rightarrow (X^TX + \lambda I)^{-1}(X^TX + \lambda I)\beta = (X^TX + \lambda I)^{-1}X^TY$$

$$\Rightarrow I\beta = (X^TX + \lambda I)^{-1}X^TY$$

$$\Rightarrow \hat{\beta} = (X^TX + \lambda I)^{-1}X^TY$$

extra only. compare to multiple lin Reg.

Date __/__/__

→ Pseudocode (Ridge Regression)

input: $\{x_i, y_i\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ &
a new predictor $x_{new} \in \mathbb{R}^p$

output: y_{new}.

step 1: choose $\lambda > 0$

step 2: create the matricess, X, β, and y

3: Compute $\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y$

4: Compute $\hat{\beta}_0 = \frac{1}{N} \sum_{j=1}^{N} y_j$

5. create $\tilde{x}_{new} = \begin{bmatrix} 1 \\ x_{new} \end{bmatrix}_{(p+1)\times 1}$

6. Generate $\tilde{\beta}_0 = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{bmatrix}$

7. Estimate $y_{new} = \underset{1\times(p+1)}{\tilde{\beta}_0^T} + \underset{(p+1)\times 1}{\tilde{x}_{new}}$

1×1

Note: ⇒ 1. the λ value adds a possitive constant to the diagonal elements of $X^T X$ before inverssion which makes $(X^T X + \lambda \cdot I)$ matrix non singular even if

Date _/_/_

(X^TX) is not a full rank matrix.

2. the cross validation method may be ussed to determine λ

3. Ridge regression shrinks the coefficients of sum of the predictors which have less significat effect on response target.

5/02/24

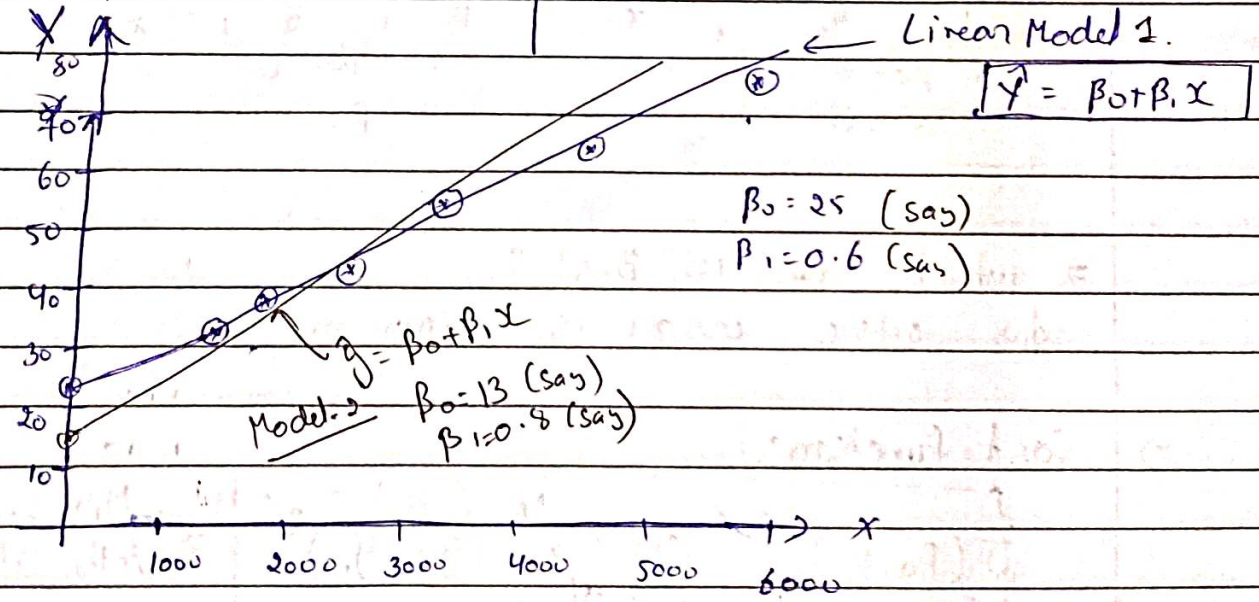
SIMPLE Linear Regression with one variable

Model Representation:

- It is supervised learning alg.
- Given any new sample, it is expected to estimate accurately the response / target.

- ex: Estimating Price of a House?

Size of feet^2 (x)	Price of House in Lakh (₹)
$x_1 \rightarrow$ 1474	34
$x_2 \rightarrow$ 1632	39
$x_3 \rightarrow$ 3407	56
$x_4 \rightarrow$ 4408	67
$x_5 \rightarrow$ 5569	78
$x_6 \rightarrow$ 2696	42
$x_7 \rightarrow$	



Date __/__/__

⇒ Training Set:

↓

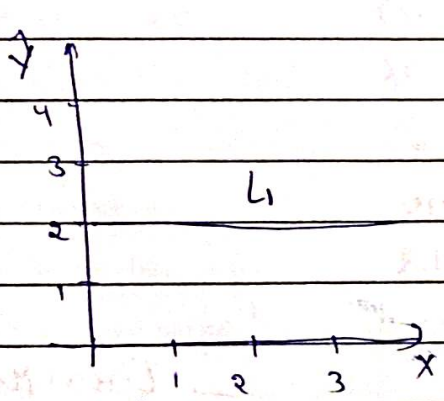
learning Algo.

↓

Size in feet² x_{new} → Hypothesis $\hat{y}$ → Price of Home y_{new}

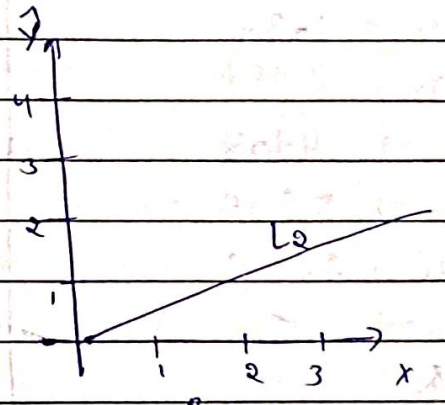
→ How to Represent $\hat{y}$

$$\hat{y} = \beta_0 + \beta_1 x$$



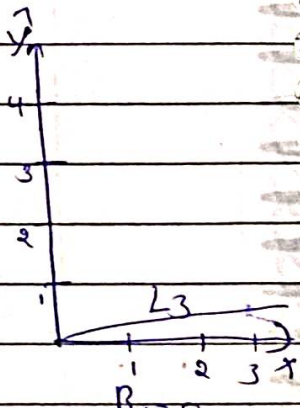
$\beta_0 = 2$

$\beta_1 = 0$



$\beta_0 = 0$

$\beta_1 = 0.8$



$\beta_0 = 0$

$\beta_1 = 0.5$

⇒ which one is Best ?

Ans: where error is Minimum

⇒ Cost function:

$$J(\beta_0, \beta_1) = \underset{\beta_0 \beta_1}{\text{Min}} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$J(\beta_0, \beta_1)$ → $RSS(\beta)$

$$\Rightarrow \underset{\beta_0 \beta_1}{\text{Min}} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

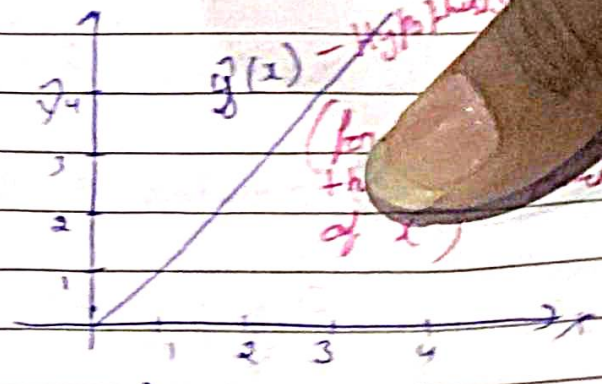
β is vector of two Paramh.

Cost function intution - 1

$\hat{y} = \beta_0 + \beta_1 x$

Parameters: β_0, β_1

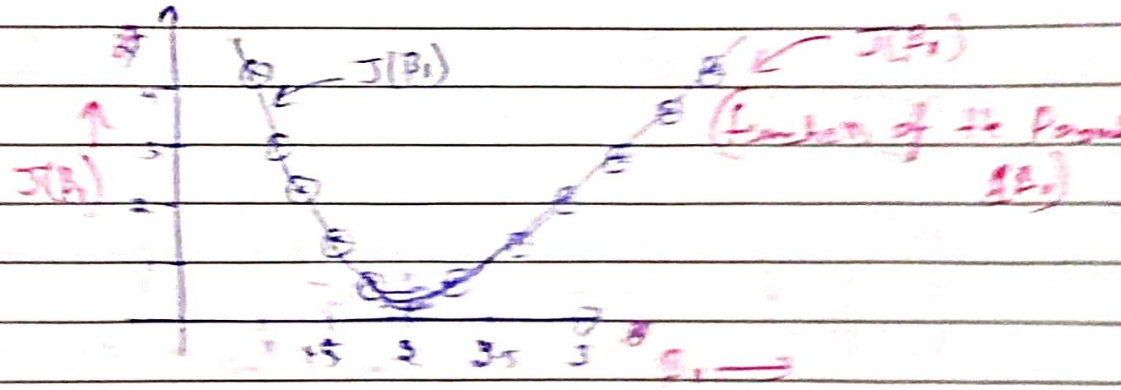
Cost function:



$$J(\beta_0, \beta_1) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Goal: $\min_{\beta_0, \beta_1} J(\beta_0, \beta_1)$

Simplied version:



$\hat{y} = \beta_1 x$

$$J(\beta_1) = \frac{1}{2N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Goal: $\min_{\beta_1} J(\beta_1)$

Date __/__/__

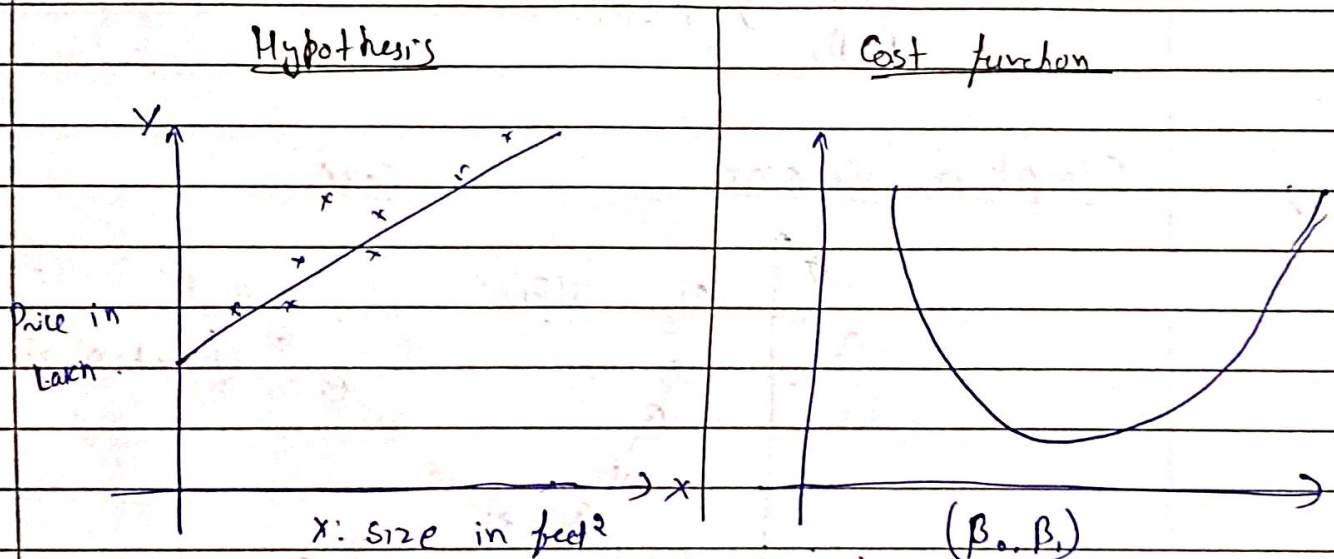
→ Cost function : Intuition - 2

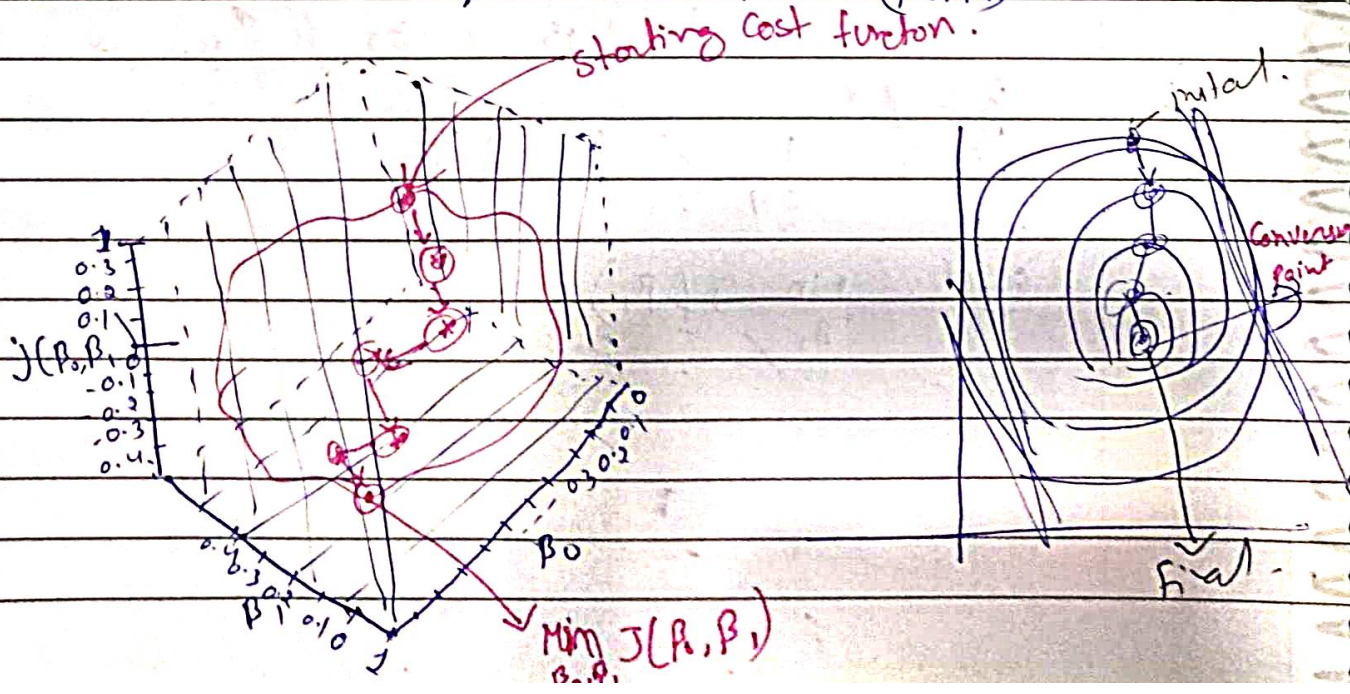
Hypothesis : $\hat{y} = \beta_0 + \beta_1 x$

Parameters : β_0, β_1

Cost function : $J(\beta_0, \beta_1) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$

Goal : $\min_{\beta_0, \beta_1} J(\beta_0, \beta_1)$





7/02/24

GRADIENT DESCENT

algo 1. start with the parameter

$\beta_0 = 0$

$\beta_1 = 0$

2. Go on changing the values of β_0 & β_1 s.t the cost function $J(\beta_0, \beta_1)$ get reduced unhill we reach a minimum.

Repeat until convergence {

$$\beta_j = \beta_j - \alpha \frac{\partial}{\partial \beta_j} J(\beta_0, \beta_1)$$

}

for simulation purpose:

Consider $temp_0 = \beta_0 - \alpha \frac{\partial}{\partial \beta_0} J(\beta_0, \beta_1)$

$$temp_1 = \beta_1 - \alpha \frac{\partial}{\partial \beta_1} J(\beta_0, \beta_1)$$

$\beta_0 = temp_0$

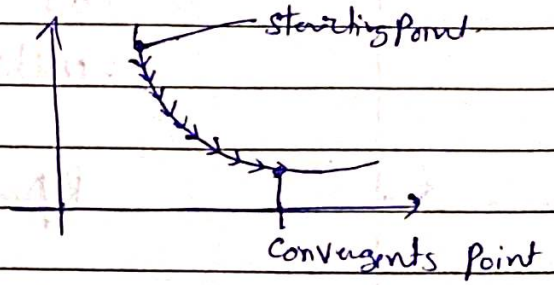
$\beta_1 = temp_1$

Date __/__/__

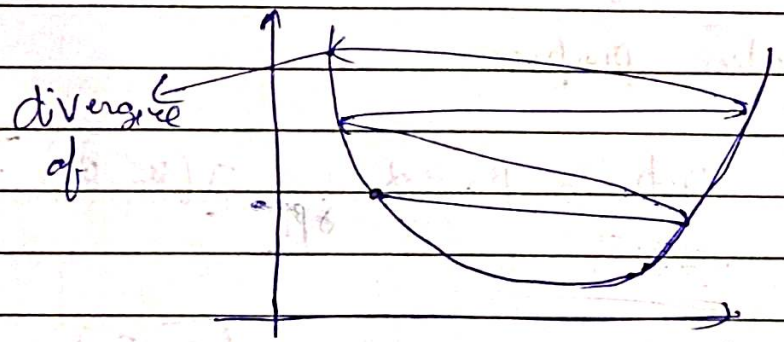
- Intuition of

1. it is a iterative optimization process to minimize objective function.

2. if α (is too small) $\Rightarrow$ gradient descent (GD) is slow.



3. if α (is too large) $\Rightarrow$ gradient descent may not converge. it may diverge.



4. Gradient descent may converge to a local minimum even with learning Rat (α) is fixed.

5. when we approach a local minimum G.D will automatically takes smaller steps

$\beta_j = \beta_j - \alpha \frac{\partial}{\partial \beta_j} J(\beta_0, \beta_1)$

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

$\frac{1}{2N} \sum_{i=1}^{N} e_i^2$

Date __/__/__

Types of Gradient Descent Method

1. Batch Gradient Descent (BGD)
2. Stochastic GD : (SGD)
3. Mini - Batch GD : (MBGD)

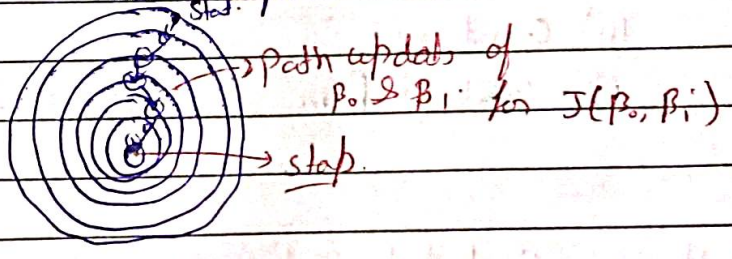
1# Batch Gradient Descent: (BGD):

i) in case of BGD the whole training sample is used to compute the Parameter updates.

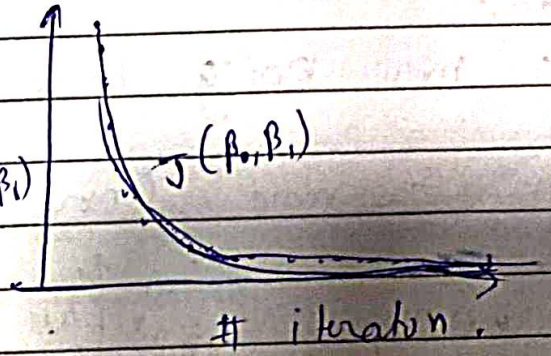
ii) that is the parameter β_0, β_1 are updated after Computing the gradient of the error w.r.t the entire trainning set.

iii) time complexity is more

iv) it makes smooth updates in model parameter that is



v) when no. of iteration increases



Date __/__/__

2# Stochastic GD (SGD)

Pseudocode (SGD):

1. Initialization $\beta_0 \leftarrow$ random$_0$
   $\beta_1 \leftarrow$ random$_1$

2. Define K: No. of max iteration
   α : learning Rate : (0 to 1)
   ϵ : error tolerance : >0
   $\epsilon x = 0.0001$
   Count = 0

3. do
   {/* Repeated until convergence or reaches max, iteration */}

   i) $X \leftarrow X$ shuffle
   ii) $\beta_{j\,new} = \beta_{j\,old} - \alpha \frac{\partial}{\partial \beta_j} (\beta_0, \beta_1)$
   iii) Count ++;
   iv) while $\left(\| \beta_{new} - \beta_{old} \| \geq \epsilon \;\; \underset{\text{or}}{\|} \;\; \text{Count} \leq K \right)$

1) it is computational efficient that BGD
2) as it selects the training sample randomly that's why it is Stochastic
3) generally it is more noisy than BGD
4) No. of iteration are more (more efficient)
5) it avoid local minima.
6) However it is sensitive to learning Rate.
* Due to less expensive it is Prefered over BGD.

Date __/__/__

02/02/24

#3. Mini - Batch Gradient Desent Alg (MBGD)

i) The Parameters are updated after computing the Gradient of the error w.r.t the subset of training set.

ii) The updates are faster in comparison to BGD & slower in comparison to SGD (if it uses less no. of training sample)

iii) Depending upon the BATCH-SIZE the updates can be of less noisy.

-i.e Greater is the Batch size lesser is the Noise.

Algo. MBGD

initialization:

$\beta_0 = \beta_{0init}$

$\beta_1 = \beta_{1init}$

K = Max. iteration (No. of epoches)

for i=1 to K

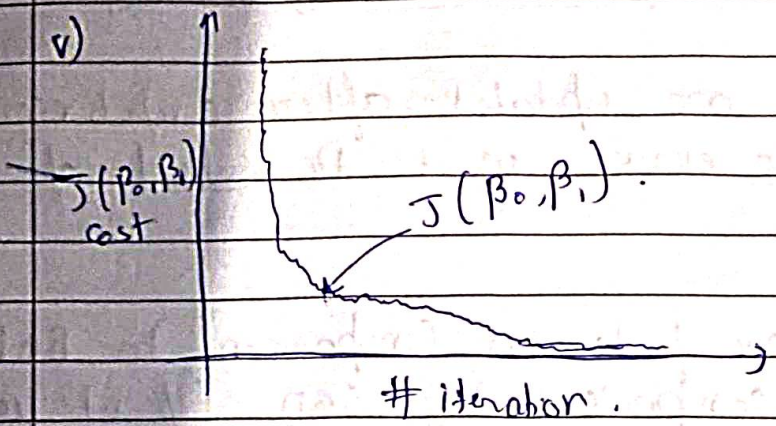
for mini-Batach (X_min, Y_min).

$$\beta_{jnew} = \beta_{jold} - \alpha \frac{\partial}{\partial \beta_j} J(\beta_0, \beta_1)$$

iv) The Convergence:

Date __/__/__

v)



Feature scaling

i) Normalization
ii) Standardization

- Feature scaling is important for building accurate and effecting ive ml. model by using either Normalization or standardization processes.

- that is feature scaling is the process of transforming the data to specific ranges. so as to reduce the impact of outliers and to reduce dominance of huge data of one feature.

Advantages:

i) improve performance of the model.

ii) Reduces the impact of outlier.

iii) Ensuring data in the same scale.

Date __/__/__

for ex:

Student	CGPA
A	3.5 in the scale of 0 to 5
B	5.2 " " " 0 to 10
C	75%

↳ Performance of the students

- Here we need to scale to performace of the students in some standardized method (scale)

↳

Student	% of final exam X_1	Salary in the company (lakhs) X_2
A	3.5	60
B	3.2	70
C	4.0	55
D	4.8	62

range [3, 5] range [55, 100]

The two feature X_1 & X_2 are not in respective scale (Range)

→ Why should we need feature scaling & when? (models)

[∵ Some ML Algo are sensitive to feature scaling where as some are insensitive to feature scaling

#1 Gradient Desent Algo (BGD, MBGD, SGD)

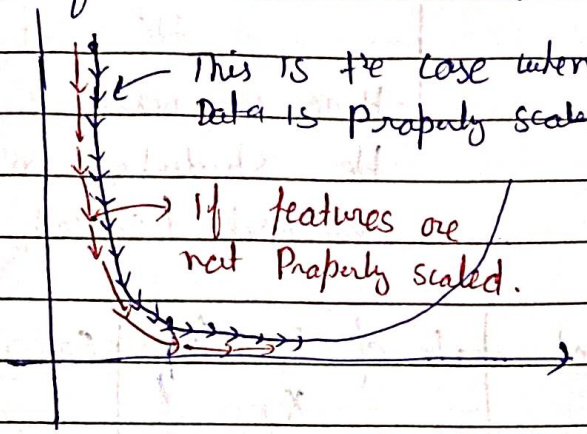
- The ML model such as LR, Logistic regression, Neural N/W, PCA (Principal component Analysis) that use Gradient Desent Algo as an optimization technique. require data to be scale.

Date __/__/__

$$\beta_j = \beta_j - \alpha \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i) x_j^{(i)}$$

- the presence of feature X that is $x_j^{(i)}$ will effect the step size of Gradient descent.

affects stepsize



- to insure that GD moves smoothly towards the minimum & the steps of GD are updated at the same Time & Rate for all the feature. We have to scale the data before ~~fitting~~ to the model feeding (taking) input)

#2. Distance based Algo:

the ML models KNN, K means, SVM. are effected by the range of the features. So scaling is important for those algo. also.

activity:

for ex:

Student	X_1 CGPA	X_2 Salary (lakh)
A	8.5	66 → scale (0,1)
B	3.2	70
C	4.0	55
D	4.8	62

scal(0,1)

Raj [3,5) Raj [50,100]

The two feature X_1 & X_2 are of different scale (Range)

Date __/__/__

#3. Tree - Based Algo.

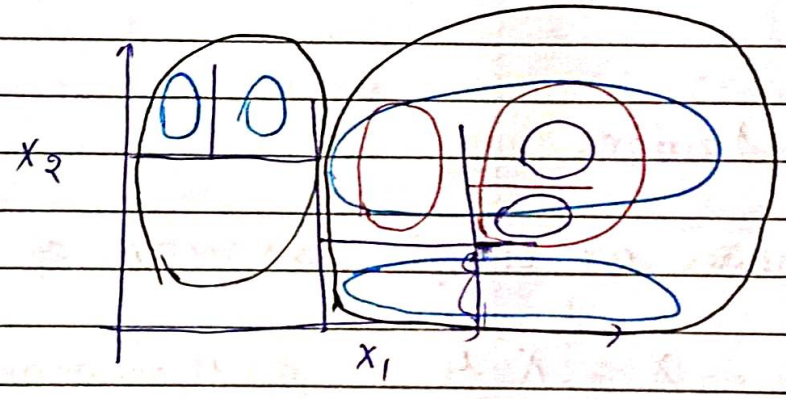
ML models like decision tree, Random are fairly insensitive to the scale of features

- the decision tree splits a node by selecting the correct feature & then ~~puting~~ spliting a node at a suitable spliting point.

- in this case by splitting it increases the homogeneity of the node.

- other features don't influence this split of feature.

ex:



12/02/24 W Date: _/_/_

1. Normalization:

Normalization for feature scaling

In this method the features are rescaled to the range [0, 1]. it is also known as min-max feature scaling method.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

i.e initially the range of feature was $[X_{min}, X_{max}]$ and after Normalization the range becomes [0,1]

2. Standardization:

The features are closed to mean of the corresponding feature.

$$X' = \frac{X - \mu}{\sigma}, \quad \mu = \text{mean}$$

σ = Standard Deviation.

$$= \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

Note: Variance $= \sigma^2$

$$= E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

$$= E[(x^2)] - (E[x])^2$$

for 3-marks

Date: __/__/__

Q. When we should use Normalization and when to use standardization in feature scaling

Normalization	Standardization
1) The features are scaled to the range [0,1].	1) The features are scaled within range $[\mu-\sigma, \mu+\sigma]$
2) When the distribution of data set is unknown & not Gaussian.	- Distribution ~~Gaussian~~ unknown & Gaussian
3) Retains the shape of the distribution	- original distribution is not retained
4) Sensitive to outliers	- less sensitive to outlier
5) $X' = \frac{X - X_{min}}{X_{max} - X_{min}}$.	$x' = \frac{x - \mu}{\sigma}$
6) The relationship b/w the data point may not be preserved	- relationship b/w data point is preserved.

Date __/__/__

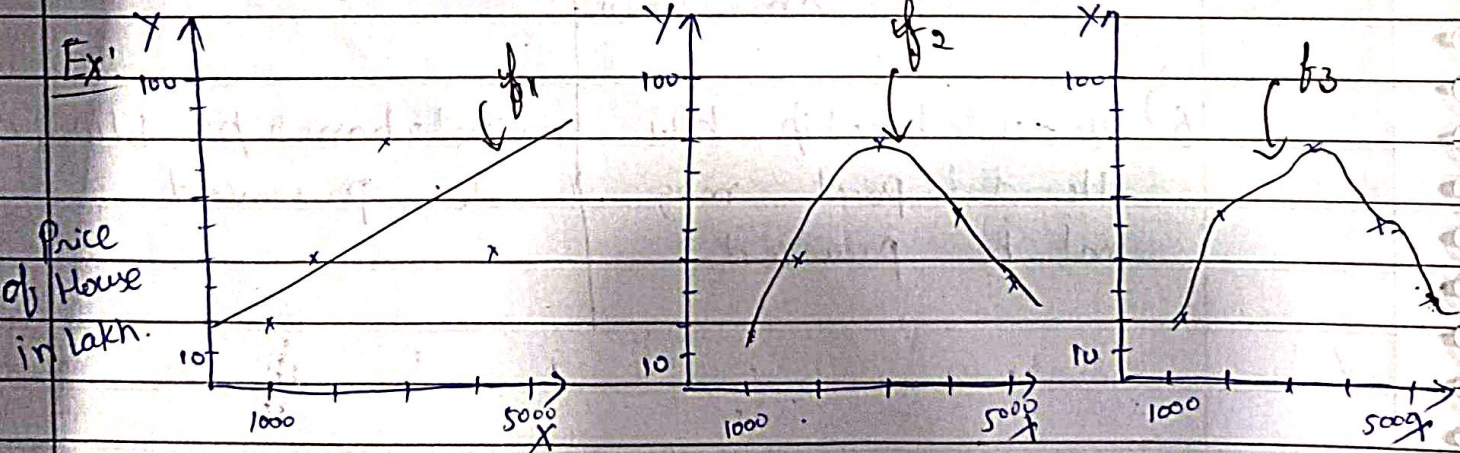
~~Prob~~ Overfitting & Underfitting.

Underfitting: Vs Overfitting

Underfitting: When a complex model / Problem is modeled by using a simple method then the model may not fit to the data set & may not achive the desire accuracy.

Overfitting: When No. of features in data set is large the learned Hypothesis may fit the dataset very-well (~~i.e $J(\beta) = \frac{1}{N}$~~ (i.e $J(\beta) = \frac{1}{2N}\sum_{i=1}^{n}(y_i - \hat{y}(x_i))^2 \approx 0$) but it couldn't approximate any new x_{new} accurately.

Ex:



$f_1 = \beta_0 + \beta_1 x_1$
degree - 1
- Under fit

$f_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2$
degree - 2
Just fit

$f_3 = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x_4^4$
degree - 4
Overfit

Date __/__/__

Underfit	Just fit	Overfit
- Training error is large - Test error large ✗	- Training & test error at par. (both accepted) ✓	- Training error is minimum - Test error is large ✗

How to address overfit ? Method

i) reduce no. of features.
- Manually select the suitable features to keep
- apply model selection Algo.

2) Regularization
- Keeping all the features & reducing the magnitudinal value of the parameters β_j

The methods are :- i) Ridge Regularization) Parameters are regularized
ii) LASSO Regularization

- Scale or Re-scale the features so that all the features can contribute uniformly i.e, $X_1, X_2, X_3 \dots X_p$ should contributed to y uniformly (∵ By regulating the ranges of features) → features are regulated)

Date __/__/__

BIAS & Variance Trade off

Objective: TO decide for a any given set of Data which method produces the best ~~rest~~ result i.e selection of a suitable ML model for a given set of data is the most challenging task for a data Scientist w.r.t Statistical learning in Practice.

How to?

1) We have to measure the Quality of fit of the ML model (least error).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{f}(x_i)\right)^2$$

Training Mean Square error — $\hat{f}(x_i)$ estimated.

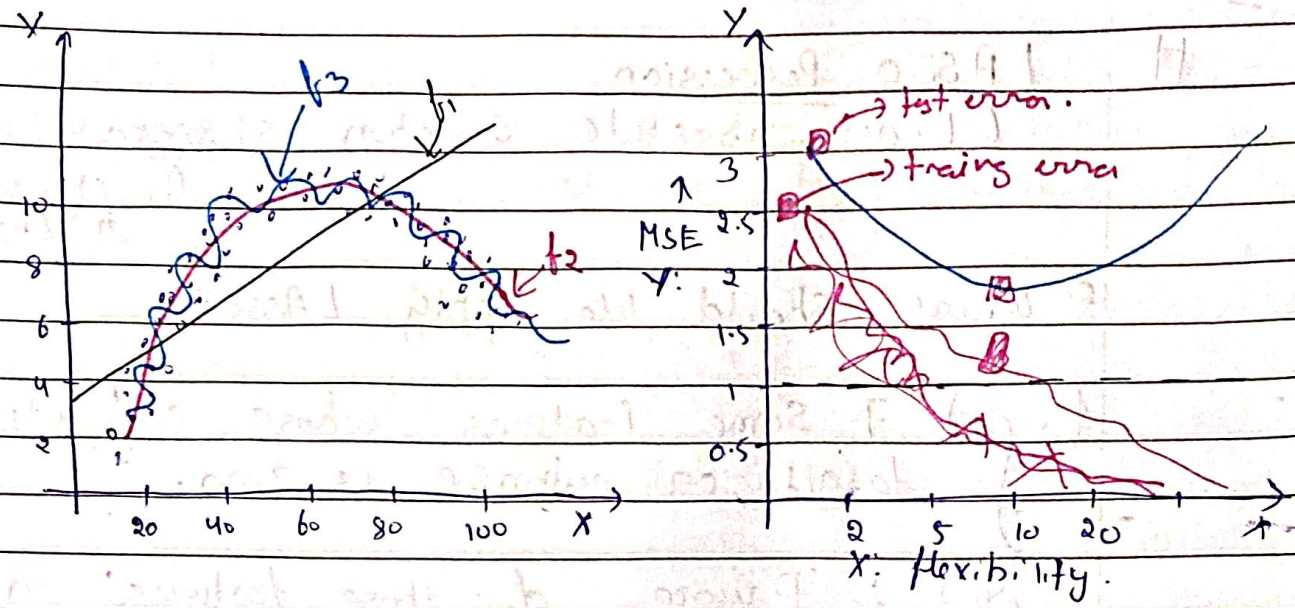
Analysis: Here we are not intrested where or not the model accurately predicts the training Dataset.
We are interested in accurately Predicting on the new Sample.

So, for given $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$, $x_i \in R^p$, $y_i \in \{0,1\}$

we obtain the estimate $\hat{f}$

- We can then compute $\hat{f}(x_1), \hat{f}(x_2), \hat{f}(x_3) \ldots \hat{f}(x_N), \hat{f}(x_{new})$

Date __/__/__



- f_1 → linear Regression
- f_2 → approximated spline. - traing error.
- f_3 → approximated spline.

W Date __/__/__

19/02/24

LASSO Regression:

(Least absolute selection shrinkage operator Algo)

Q) i) Ridge (near to zero)
ii) LASSO. (zero)

Q Where should we apply LASSO?

Ans: i) ∃ Some features whose contribution towards target or response is zero.

ii) ∴ Presence of those features in the ML model incurs
a) Unnecessary Complexity.
b) affect the performance of the model.
c) Desired accuracy is not achieved

→ There unnecessary features should be made to ZERO

Note: LASSO is also considered as Features selection.

∵ X_1	X_2	X_3	X_4	X_5	...	X_p	Y
	= 0	= 0		= 0			

K - features are selected out of p features:

$K < p$.

Date __/__/__

How To?

In case of LASSO regression the parameters β_j ($j=1$ to p) are estimated by minimizing the function as:

$$\text{LASSO RSS}(\beta) = \sum_{i=1}^{N} \left[y_i - \left(\beta_0 + \sum_{j=1}^{P} \beta_j x_j \right) \right]^2 + \lambda \sum_{j=1}^{P} |\beta_j|$$

error term

$\lambda > 0$

Regularization term.

L^1 - Regularization

$\Rightarrow$ L^2 - Regularization
: Ridge Regression.

or: Minimizing:

$$RSS(\beta) = \sum_{j=1}^{N} \left[y_i - \left(\beta_0 + \sum_{j=1}^{P} \beta_j x_j \right) \right]^2.$$

Subject to $\sum_{j=1}^{P} |\beta_j| < t$, where $t > 0$.

t is constant

lasso is better than Ridge:

- it discards unnecessary feature.
- Performance is better.
- Accuracy is also good

Date __/__/__

LOGISTIC Regression (classification)

* it is a supervised ML model for classification.
* it is a statistical model.

Goal: To Predict the class label of any new data x_{new} by estimating prob $(Y=1 \mid X=x_{new}) = ?$

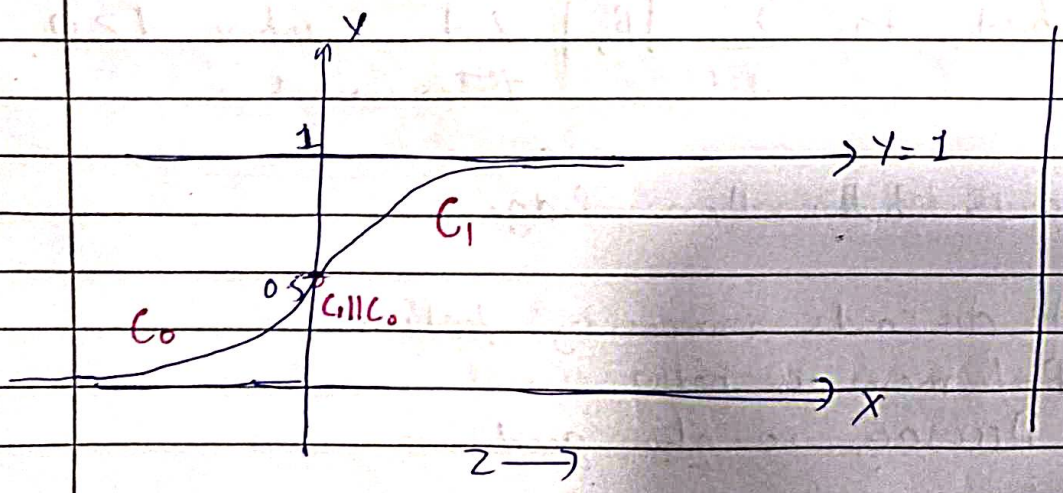
if $(P(Y=1 \mid X=x_{new}) > 0.5) \Rightarrow x_{new} \in C_1$ (class 1)

else if $(P(Y=1 \mid X=x_{new}) < 0.5) \Rightarrow x_{new} \in C_0$ (class 0)

else $(P(Y=1 \mid X=x_{new}) = 0.5) \Rightarrow x_{new} \in C_1 \| C_0$

To estimate the Probability SIGMOIDAL function is used.

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



if $z=0$, $g(z)=0.5$

if $z \to \infty$, $g(z) \to 1$

if $z \to -\infty$, $g(z) \to 0$

Date __/__/__

types of logistic Regression

i) Binomial : 2-class classification Problem (0 or 1)

ii) Multinomial : No. of classes are 3 or more & not ordered

eg: Tiger, Lion, Cat, Dog.

iii) ordinal : 3 or more number of ordered classes.

eg: low, Medium, High

: F, D, C, B, A, E, O (grades)

- Conditions for implementing logistic Regression (Assumption)

i) Features should be independent. (independent features)

ii) Binray dependent variables.

iii) ∄ any outliers.

iv) Large Data size

v) ∃ linear relationship b/w independent variables & log odds (class label)

Date __/__/__

→ Pseudocode (logistic Regression).

Given the dataset $D = \{x_i, y_i\}_{i=1}^{N}$, $x_i \in \mathbb{R}^p$, $y_i \in \{0,1\}$

Predict the class label of x_{new} ?

1. initialize β^{old}, choose some small $\epsilon > 0$ (Eg: $\epsilon = 0.0001$), $\alpha \in (0,1)$

2. Compute Y, X

3. for $i = 1$ to N

$$P_i = \frac{1}{1 + e^{\beta^{old^T} \times \tilde{x}_i}}, \text{ where } \tilde{x}_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix}$$

4. Compute P: $P_1, P_2, P_3, \dots, P_N$, where $P_i = P(x_i, \beta^{old})$

5. Compute $\beta^{new} = \beta_{old} + \alpha X^T (Y - P)$ → (/* gradient ascent */)

6. if $\| \beta^{new} - \beta^{old} \| < \epsilon \Rightarrow$ Return β → /* It is done */

7) else $\beta^{old} = \beta^{new}$ and goto step-3

8) $$P(Y = 1 \mid x = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}}$$

9) if $(P(Y = 1 \mid x = x)) > 0.5 \Rightarrow x$ is in class 1.

10) else if $(P(Y = 1 \mid x = x)) \leq 0.5 \Rightarrow x$ is in class 0

Date __/__/__

Explanation:

$$|| \beta^{new} - \beta^{old} ||$$

$$\Rightarrow \left(\beta_1^{new} - \beta_1^{old}\right)^2 + \left(\beta_2^{new} - \beta_2^{old}\right)^2 + \cdots + \left(\beta_p^{new} - \beta_p^{old}\right)^2$$

Now if ( ) ≤ 0.0001

$\Rightarrow$ Return $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$

β is vector.

Date: _/_/_

Error analysis:

Why?
We need to develop a
i) reliable
ii) robust
iii) least bias
iv) least variance
v) generalized ML model for real-life Problem.

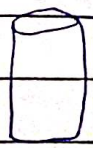
To achieve such a model the following parameters are considered for models performance evaluation.

The parameters are:
i) accuracy of the models
ii) Precision " " "
iii) Recall " " "
iv) F1-score " " "

1. error analysis in detail

It is the process of isolating, observing, & diagnosing the erroneous ML predictions so as to determine the high & low performing ML model.

Case-study-I: Image Recognition ML model for Dog detection.



Benchmark → In day light accuracy: 100% | In Night 70% | In Rainy night 65% | In a water pool 80%

Date __/__/__

Alg.	86%	62%	60%	46%
Alg.	85%	85%	63%	74%
Alg.	85%	68%	58%	41%

○ → accepted
○ → May be accepted
○ → Rejected

2. Error identification & diagnosis

Case study - Tiled (features are not apparent)
when the types are text, graphic, audio, image...

ML model: Cat classification.

Suppose we train a model on: 5000 images
(those images including cat in there different scenario)

No. of test sample: 1000
% of accuracy: 85%

850 images are correctly classified

150 images are miss classified.

Confusion matrix:

Total = P+N = 1000	Actual: Cat (P)	Not Cat (N)
Predicted: Cat (PP)	700 (TP)	100 (FP) → Type I error
Predicted: No cat (PN)	50 (FN) → Type-2 error	150 (TN)

Here, the confussion matrix doesn't provide clarity when and what is being miss-classified.

- Consider separate situation & analyse the % of error.

Soln

Case-Study II. (when data is more apparent)

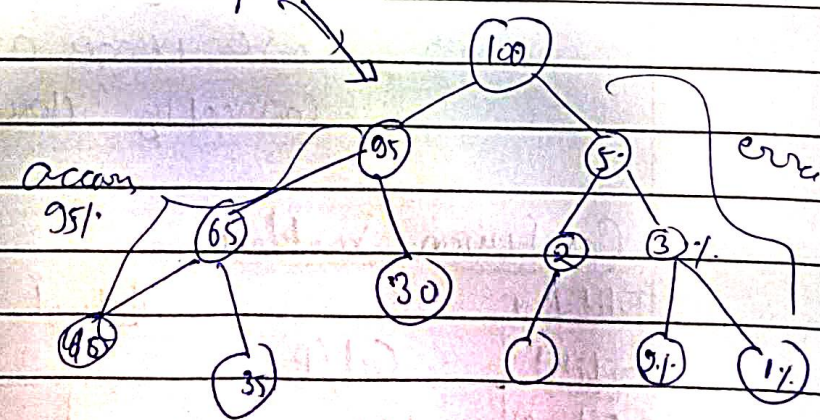
eg: Predicting the Mileage of a car
estimating the Credit Card limit of a person
Recommending a product

the features should be clearly categorized

→ Segregation of features: can help identifying & further diagnose the error distribution across features & certain values of those features.

methods are: i) Heat maps
ii) Tree maps

x_1	x_2	x_3	x_4	. .	x_p
8%	10%	5%	21%		9%



Date __/__/__

Short 9

#3. How to resolve the error.:

after identifing & analysing the errors we need to resolve those errors

Step 1: Collection more data where errors are happing

Step 2. Augment the data (By transformation)
↓
(i) translation, ii) Scaling, iii) Rotation, iv) Shearing)

Step 3: Use other ML model or tune the Hyper Parameters differently.

Problem solving (using Confusion Matrix)

i) Type-1 error (FP): over estimation (False alarm)
ii) Type-2 error (FN): Under estimation. (Miss)
iii) Precision: $\frac{TP}{PP} \Rightarrow \frac{TP}{TP + FP}$

iv) Recall: $\frac{TP}{P} \Rightarrow \frac{TP}{TP + FN}$ (Sensitivity)

v) accuracy: $\frac{TP + TN}{P + N}$

vi) F-1-Score = $\frac{2 * Precision * Recall}{Precisson + Recall}$