Data Science and Artificial Intelligence

Machine Learning

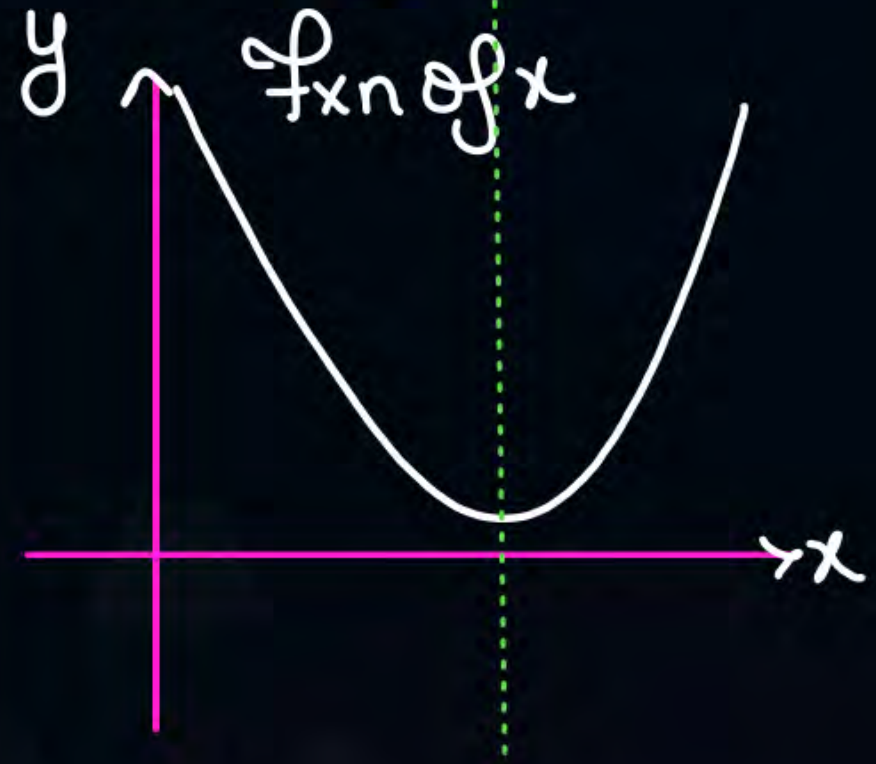
Linear Regression

Lecture No. 02

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- **Topic** Basic def
- **Topic** Model
- **Topic** Optimization
- **Topic** Training Process.
- **Topic** ML main purpose



Topics to be Covered

Topic 1D Linear regression

what is.

Topic Formulae

Topic mean

Topic Variance

Topic Covariance

About the Faculty

- AIR 1 GATE 2021, 2023 (ECE).
- AIR 3 ESE 2015 ECE.
- M.Tech from IIT Delhi in VLSI.
- Published 2 papers in field of AI-ML.
- Paper 1 : Feature Selection through Minimization of the VC dimension.
- Paper 2 : Learning a hyperplane regressor through a tight bound on the VC dimension.

By- SIDDHARTH SABHARWAL SIR

Success is walking from failure to failure with no loss of enthusiasm.

WINSTON CHURCHILL

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

Convex function

REVISION

Fill in the blanks :

1. The target/Goal of the ML is To learn the pattern of data, find relation b/w y and x and using this relation or function we can predict any y for new x values.
2. The best optimized model is that which minimize the error (Residue) in Training data
3. The problem with the simple model is Underfit (Cannot understand pattern of data)

REVISION

Fill in the blanks :

4. The problem with highly complicated model is ____________ Overfitting → Rote learning

5. The data is used to ____________ (Train) / Find. the ML model

6. The data is collected from ____________ Survey / experiment.

REVISION

Wo Teen bacchhe......

	under fit	over fit	Best fit
Training data →	high	zero	low
Test data →	high	high	low.

REVISION

What is a Residue

- diff b/w actual value and pred value
- $y_1 - \hat{y}_1$, $y_2 - \hat{y}_2$ --

data value → x

label → y

model ya fxn $y = mx + c$

x	y	$\hat{y}$
x_1	y_1	$\hat{y}_1 = mx_1 + c$
x_2	y_2	$\hat{y}_2 = mx_2 + c$
x_3	y_3	⋮

- absolute value of residue $|y_i - \hat{y}_i|$
- Square $(y_i - \hat{y}_i)^2$

How we do optimization (Absolute error & RSS)

model is $y = mx + c$

we have ∞ options for m and c values.

How to find best possible value

$(\hat{y}_i = mx_i + c)$

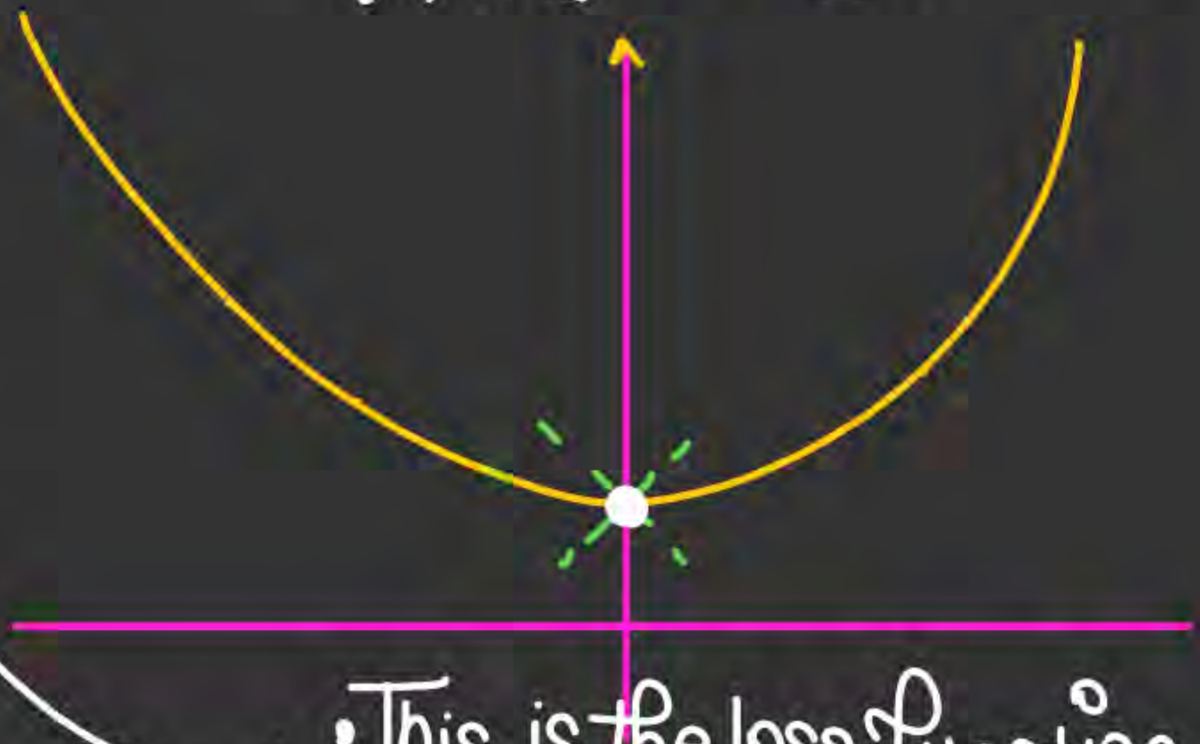
$$\min \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad \text{OR} \quad \min \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

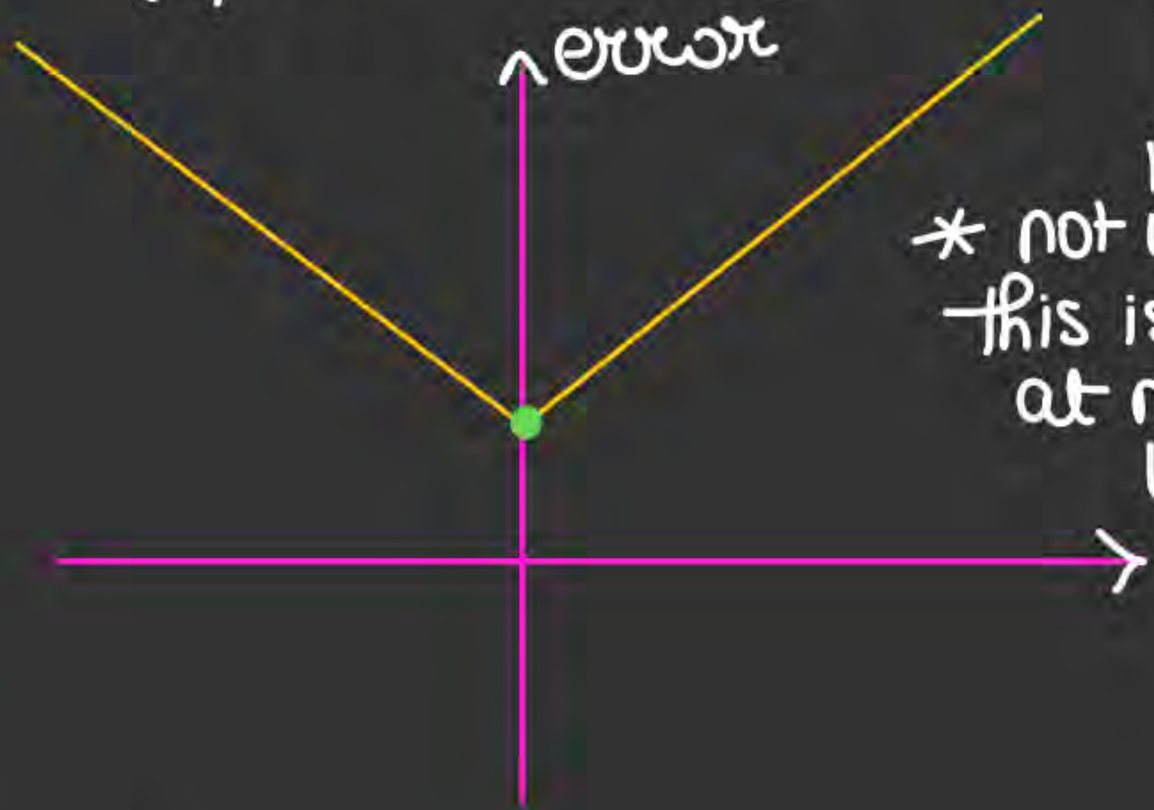
$$\min \sum_{i=1}^{N} \left(y_i - (mx_i + c)\right)^2 \quad \text{OR} \quad \min \sum_{i=1}^{N} |y_i - (mx_i + c)|$$

So we want that m, c that minimize these Quantities.

Convex Function

func ⇒ $\min \sum_{i=1}^{N} \left(y_i - (mx_i + c)\right)^2$ OR $\min \sum_{i=1}^{N} \left|y_i - (mx_i + c)\right|$





* not used bcoz this is not diff. at min location

- This is the loss function
- m and c are found by $\frac{\partial f}{\partial m} = 0$ $\frac{\partial f}{\partial c} = 0$
- Residual Sum of Squares RSS

So this is optimization :-

- model $y = mx + c$
- we can choose ∞ values of m, c
- But the best m, c is obtained by min RSS

Residual Sum of Square

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

- RSS : Convex fxn, $\frac{d}{dm} = 0$, $\frac{d}{dc} = 0$ and find m, c

REVISION

19. The output of training process in machine learning is

A. machine learning model $\Rightarrow y = f(x)$ function.

B. machine learning algorithm

C. null

D. accuracy

REVISION

34. In simple term, machine learning is

A. training based on historical data

B. prediction to answer a new query

C. both a and b??

D. automization of complex tasks

$$f(a,b) = 3a^2 + 6ab + 9b^2 + 10$$

$$\Rightarrow \frac{\partial f}{\partial a} = 6a + 6b$$

b is a Const

$$\Rightarrow \frac{\partial f}{\partial b} = 6a + 18b$$

a is Const

Problem 1 – Predict Population of bacteria in a lab

We must create a model with following data

$y_{predicted}$ x y

C

$(K\sqrt{10}+c)$

Time	Population
0	50
10	200

$y = (K\sqrt{x}+c)$ find K and C

$$\min\left[(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2\right] \Rightarrow \min\left[(50-c)^2+(200-K\sqrt{10}-c)^2\right]$$

$$\frac{\partial F}{\partial c}=0 \Rightarrow 2(50-c)(-1)+2(200-K\sqrt{10}-c)(-1)=0$$

$$\partial F/\partial K=0 \Rightarrow 2(200-K\sqrt{10}-c)(-\sqrt{10})=0$$

Now predict the population at t = 20

$$\frac{\partial F}{\partial c} = 0 \Rightarrow 2(50-c)(-1) + 2(200 - K\sqrt{10} - c)(-1) = 0$$

$$\partial F/\partial K = 0 \Rightarrow \cancel{2}(200 - K\sqrt{10} - c)(\cancel{-\sqrt{10}}) = 0$$

i) $(200 - \sqrt{10}\,K - c) = 0$

ii) $2(50-c)(-1) + 0 = 0$

$$\boxed{\begin{aligned} C &= 50 \\ K &= 150/\sqrt{10}. \end{aligned}}$$

Problem 2 – Predict Sale of I-phone based on Age of customer

We must create a model with following data

H.W

Age Age (x)	Sale of I-Phone (in a month) Sale of iphone (y)
30	300
40	400

Predict Sale of iphone ⇒ y

H.W model! $y = mx + c$

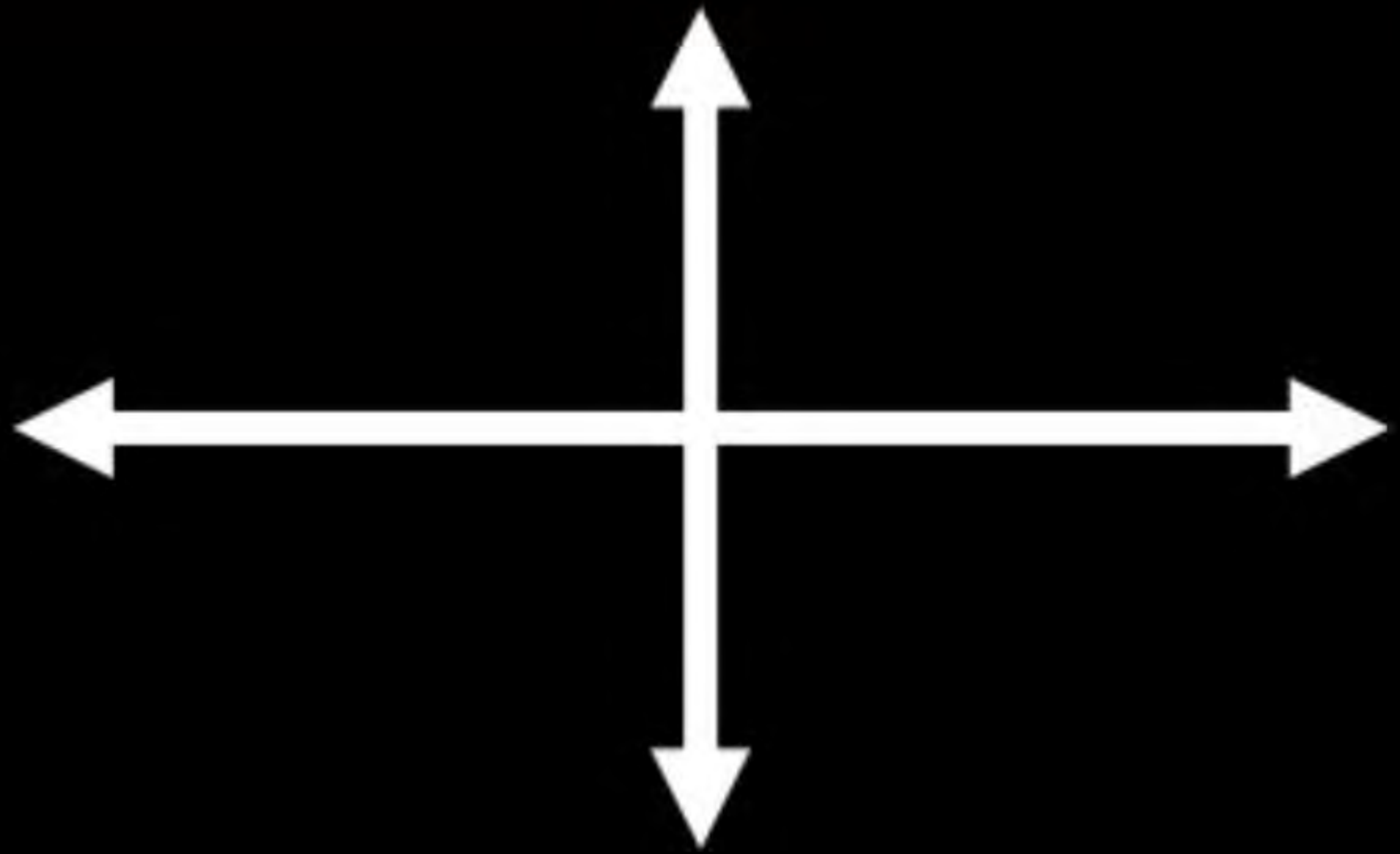
find m, c

Now predict the Sale of I-Phone at Age = 20

Problem 2 – Predict Sale of I-phone based on Age of customer

We don't have any expert now, and data has only two Points.
So ____________________

What is the best model now ?



GAATE 2026 FODENGE

Problem 3 – Predict Sale of I-phone based on Age of customer

We must create a model with following data

Predicted	Age (x)	Sale of I-Phone (in a month)
30m+c	30	300
40m+c	40	400
50m+c	50	300

Sale of iphone (y)

model $y = mx + c$

Loss fxn = RSS = $(300-30m-c)^2 + (400-40m-c)^2 + (300-50m-c)^2$

Now predict the Sale of I-Phone at Age = 20

model $y = mx + c$

Loss fxn = RSS = $(300-30m-c)^2 + (400-40m-c)^2 + (300-50m-c)^2$

$$\frac{\partial f}{\partial m} = 0 \Rightarrow 2(300-30m-c)(-30) + 2(400-40m-c)(-40) + 2(300-50m-c)(-50) = 0.$$

$$2\left[\begin{array}{l} -9000-16000-15000 \\ +900m+30c+ \\ 1600m+40c+ \\ 2500m+50c \end{array}\right] = 0$$

$$(-40000 + 5000m + 120c) = 0$$

$$\frac{\partial f}{\partial c} = 0 \Rightarrow 2(300-30m-c)(-1) + 2(400-40m-c)(-1) + 2(300-50m-c)(-1) = 0.$$

$$-2(1000-120m-3c) = 0 \quad -(1)$$

$$\boxed{\begin{array}{c} 12c + 500m - 4000 = 0 \\ 1000 - 120m - 3c = 0 \end{array}}$$

$$12C + 500m - 4000 = 0$$

$$1000 - 120m - 3C = 0$$

$\times 4$

$$12C + 500m - 4000 = 0$$

$$4000 - 480m - 12C = 0$$

$$20m = 0$$

- $m = 0$

- $1000 - 12\cancel{0m}^{0} - 3C = 0$

$$\underline{C = 1000/3}$$

So model

$$\left(y = \frac{1000}{3}\right)$$

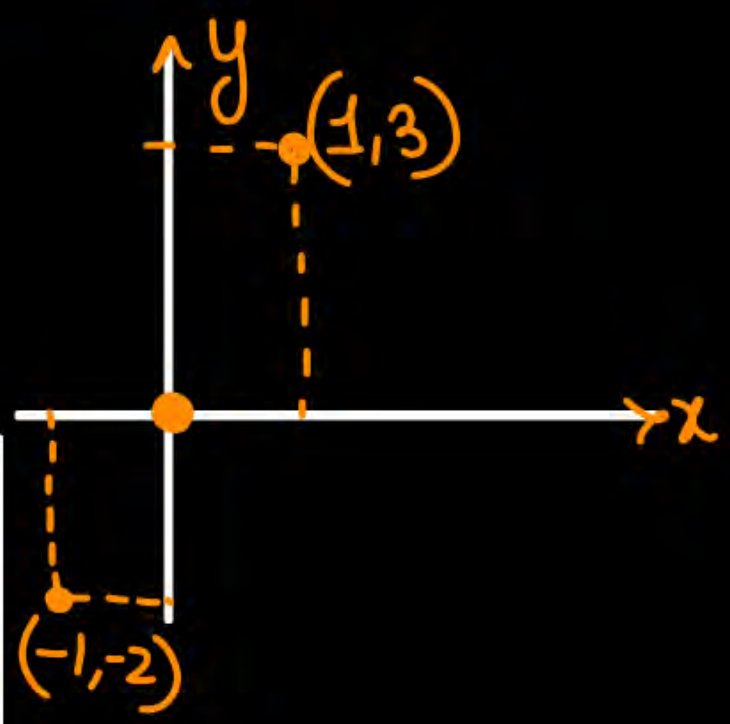
Problem 3 – Predict Sale of I-phone based on Age of customer

Find the best (least squares) straight line fit to the three points:

* we need a line that minimize the RSS

(x, y)

$$(-1,-2),(0,0),(1,3)$$



Let $y = ax + b$ be the straight line.

model

$$RSS = \left(-2-(-a+b)\right)^2 + (0-b)^2 + \left(3-(a+b)\right)^2$$

x	y	$\hat{y}$
-1	-2	$-a+b$
0	0	b
1	3	$a+b$

$$RSS = \left(-2-(-a+b)\right)^2 + (0-b)^2 + \left(3-(a+b)\right)^2$$

$$\frac{\partial f}{\partial a} = 2\left(-2-(-a+b)\right)(1) + 0 + 2\left(3-(a+b)\right)(-1) = 0$$

$$\frac{\partial f}{\partial b} = 2\left(-2-(-a+b)\right)(-1) + 2b + 2\left(3-(a+b)\right)(-1) = 0$$

$$2\left[-2+a-\cancel{b}-3+a+\cancel{b}\right] = 0$$

$$2a-5 = 0 \quad -(1)$$

$$a = 5/2$$

$$2\left[2-\cancel{a}+b+b-3+\cancel{a}+b\right) = 0$$

$$3b-1 = 0$$

$$b = 1/3$$

$$\boxed{\begin{array}{l} a = 5/2 \\ b = 1/3 \\ \quad\ 9 \end{array}}$$

Problem 3 – Predict Sale of I-phone based on Age of customer

Creating the best model

Loss Functions ?? (RSS- Residual Sum of Squares)

RSS is loss fxn

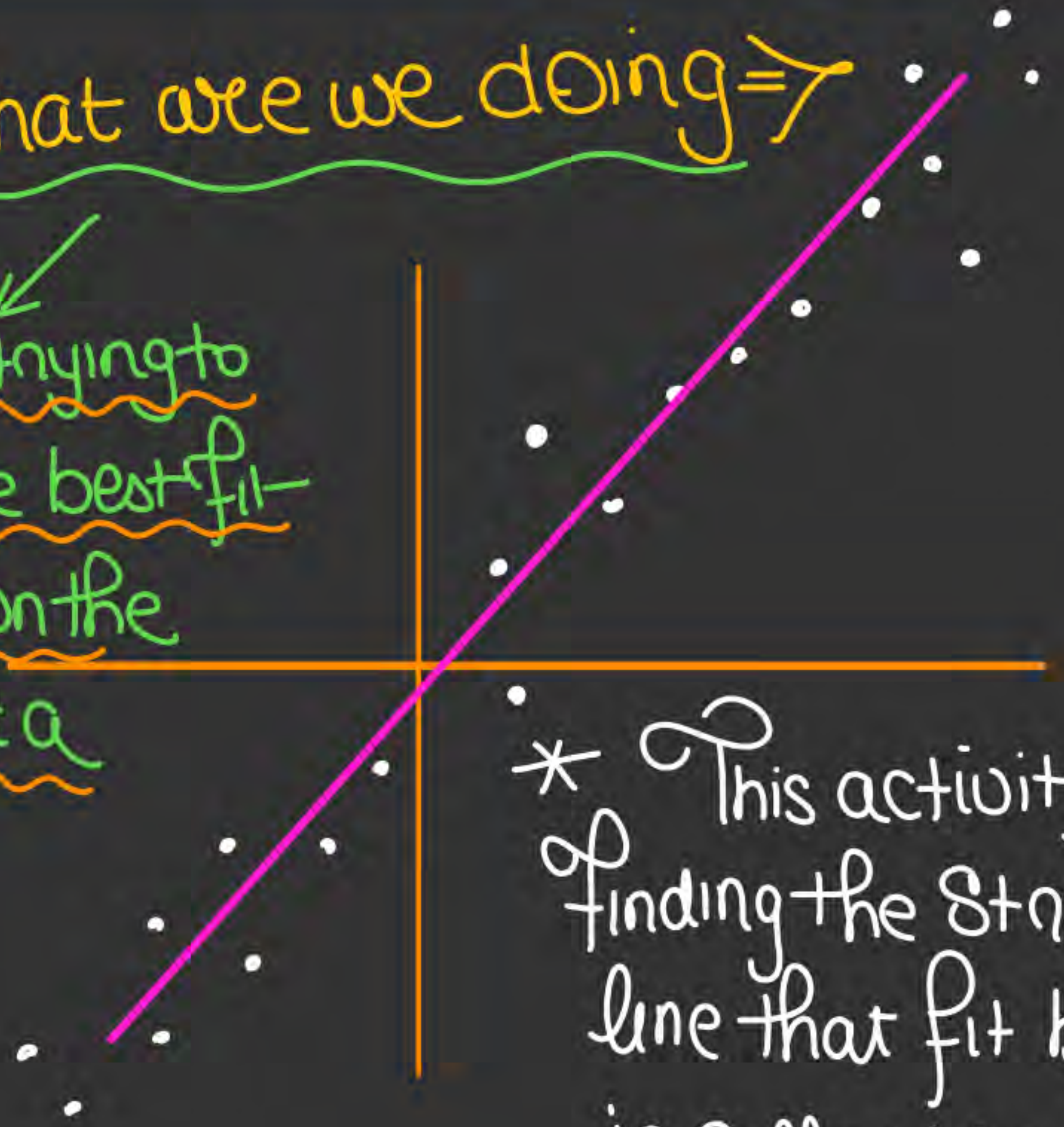
RSS

- The residual sum of squares (RSS), also known as the sum of squared residuals (SSR) or the sum of squared estimate of errors (SSE), is the sum of the squares of residuals

RSS = SSE = SSR

What are we doing ⟹

* we are trying to find the best fit straight line on the data

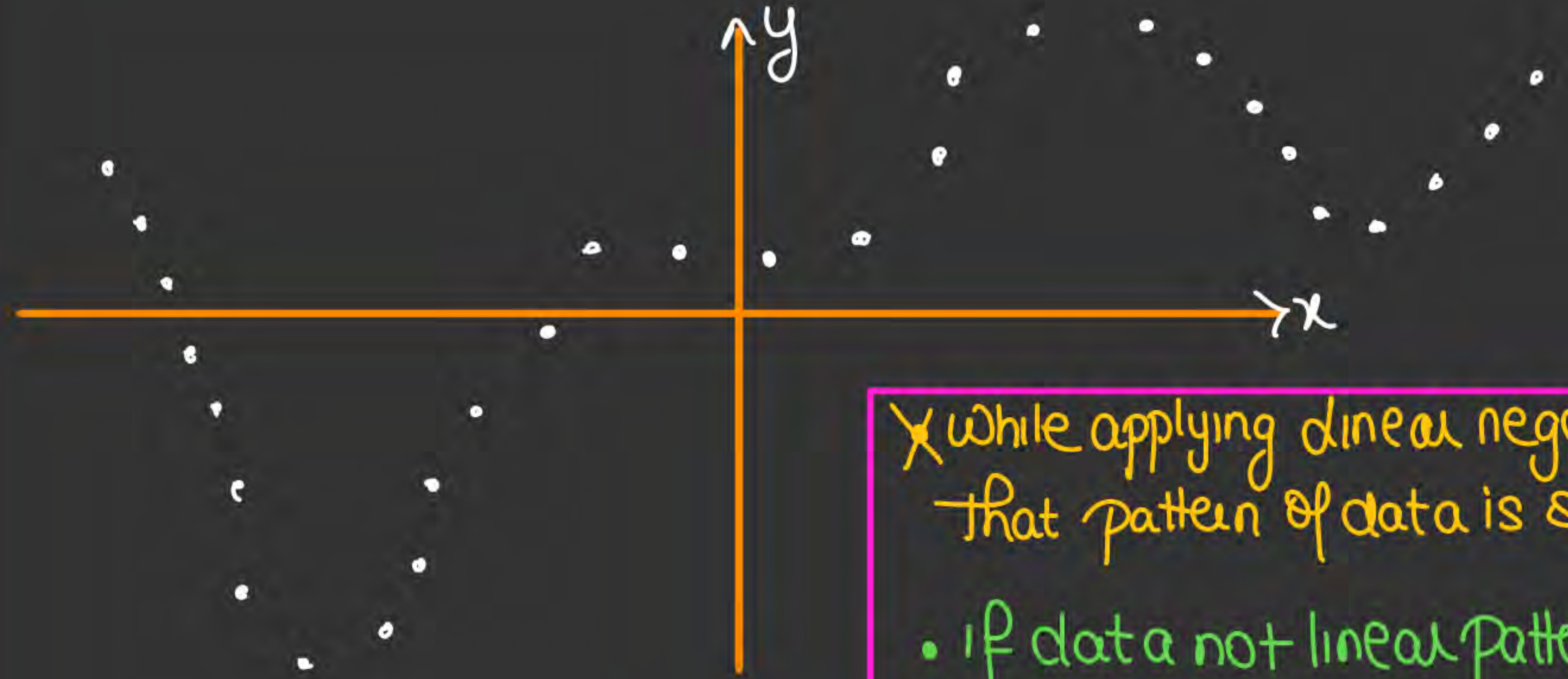


* This activity of finding the straight line that fit best on data is called linear regression

data

x	y
x_1	y_1
x_2	y_2
x_3	y_3
⋮	⋮

model $y = mx + c$
straight line



- While applying linear regression we assume that pattern of data is straight line

- If data not linear pattern then model become underfit.

GAATE 2026 FODENGE

Now how to find the best parameters ??

Variance and mean...

Class 4 Students

Marks
40
50
60
70

mean value of marks $= \frac{40+50+60+70}{4}$

average of marks $\Rightarrow \frac{220}{4} = 55$ marks.

Variance of marks $\Rightarrow \frac{1}{4}\sum\left(\text{Marks}-\overline{\text{Marks}}\right)^2$

$\Rightarrow \frac{1}{4}\left[(40-55)^2+(50-55)^2+(60-55)^2+(70-55)^2\right]$

Variance $\Rightarrow \left(\frac{15^2+5^2+5^2+15^2}{4}\right)$

GAATE 2026 FODENGE

Now how to find the best parameters ??

Variance and mean...

	x	y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
4	x_4	⋮
⋮	⋮	⋮

- N number of data points

mean value of x

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Mean of y

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$$

Variance of x $\Rightarrow$ $$\sigma_x^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$$

Now how to find the best parameters ??

Variance of $y \Rightarrow \sigma_y^2 = \dfrac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N}$.

Variance and mean...

$\boxed{\begin{array}{l} \sigma_x = \text{Standard deviation of } x \\ \sigma_y = \text{Standard deviation of } y \end{array}}$

Variance of $x = \sigma_x^2$

Variance of $y = \sigma_y^2$.

Now how to find the best parameters ??

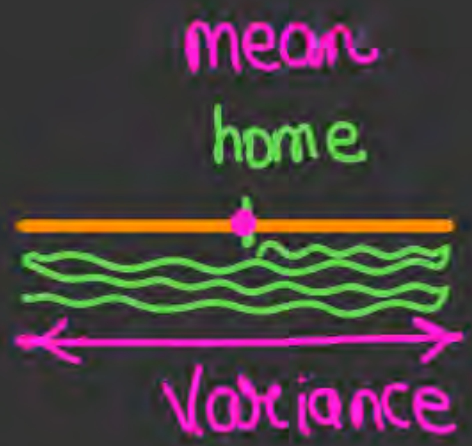
Formulae to find direct value of m and c

x	y
x_1	y_1
x_2	y_2
x_3	y_3
$\vdots$	$\vdots$

N Values

Covariance of x, y

$$\Rightarrow Cov(x, y) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N}.$$



ex

7student

mark	marks
20	32
30	32.5
40	33
50	36
60	32
70	35
80	33

Covariance show relation b/w two variables.

$\mathrm{Cov}(x,y)$ = Relation b/w x, y.

if $\mathrm{Cov}(x,y) > 0$ then   if x inc then y also inc
x, y likely related   if x dec then y also dec

if $\mathrm{Cov}(x,y) < 0$ then   if x inc y dec
x, y are opposite relation   if x dec y inc

Now how to find the best parameters ??

Formulae to find direct value of m and c

For 1 dimension data

- Each data point has single x
- 1D data

Only single 'x' values.

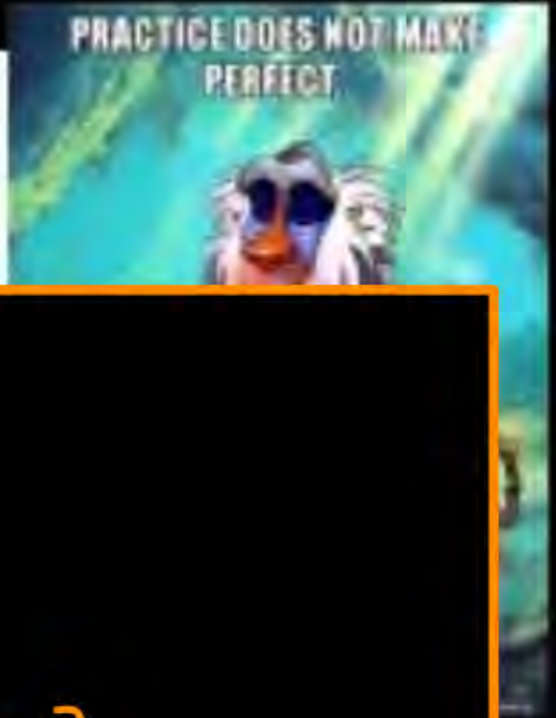
x	y
x_1	y_1
x_2	y_2
x_3	y_3
⋮	⋮

linear regression model

$$y = mx + c$$

$$\Rightarrow \boxed{m = \frac{Cov(x,y)}{Var(x)}}$$

$$\boxed{c = \bar{y} - m\bar{x}}$$



Example

Obtain a linear regression for the data in below table assuming that y is the independent variable.

x	y
2	5
3	7
4	9
5	12
6	17

5datapoint

$$\bar{x} = \frac{2+3+4+5+6}{5} = 4$$

$$\bar{y} = \frac{5+7+9+12+17}{5} = 10$$

$$Var(x) = \frac{(2-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(6-4)^2}{5} = 2$$

x	y
2	5
3	7
4	9
5	12
6	17

5 data point

$$\bar{x} = \frac{2+3+4+5+6}{5} = 4$$

$$\bar{y} = \frac{5+7+9+12+17}{5} = 10$$

$$y = mx + C$$
$$m = 2.9$$
$$C = -1.6$$

$$Var(x) = \frac{(2-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(6-4)^2}{5} = 2$$

$$Cov(x,y) = \frac{(2-4)(5-10)+(3-4)(7-10)+(4-4)(9-10)+(5-4)(12-10)+(6-4)(17-10)}{5}$$

$$= \frac{(-2)(-5)+(-1)(-3)+0+(1)(2)+2(7)}{5}$$

$$= 29/5 = 5.8$$

$$m = \frac{Cov(x,y)}{Var(x)} = \frac{5.8}{2}$$

$$C = \bar{y} - m\bar{x}$$
$$= 10 - \frac{5.8}{2} \times 4$$

$$\boxed{\begin{matrix} m = 2.9 \\ C = -1.6 \end{matrix}}$$

Basics of Machine Learning

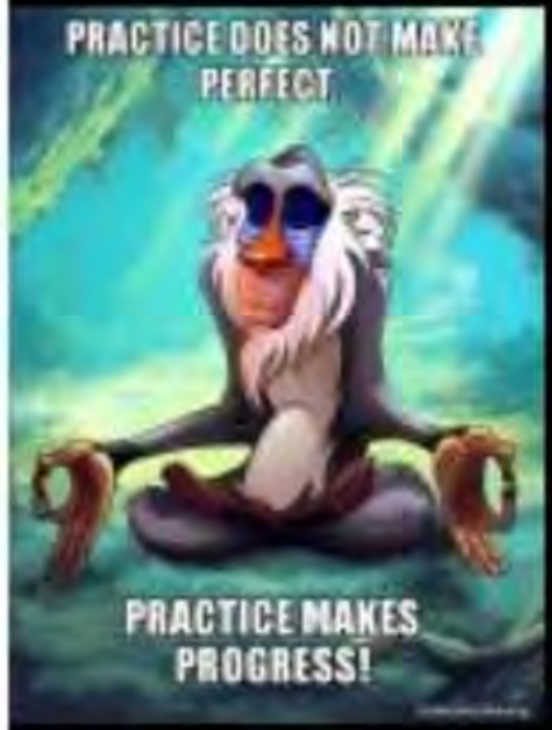
A set of observations of independent variable (x) and the corresponding dependent variable (y) is given below.

x	5	2	4	3
y	16	10	13	12

Based on the data, the coefficient a of the linear regression model

y = a + bx is estimated as 6.1

The coefficient b is ____________ . (round off to one decimal place)



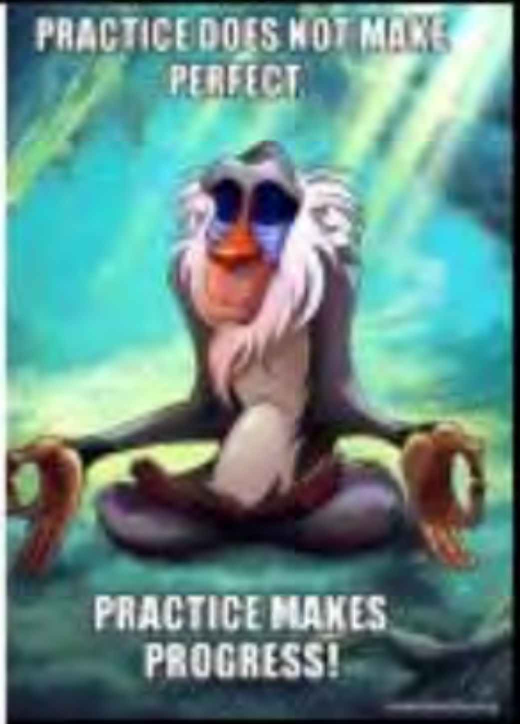
Basics of Machine Learning

For a bivariate data set on (x, y), if the means, standard deviations and correlation coefficient are

$\bar{x} = 1.0, \bar{y} = 2.0, s_x = 3.0, s_y = 9.0, r = 0.8$

Then the regression line of y on x is:

1. $y = 1 + 2.4(x - 1)$
2. $y = 2 + 0.27(x - 1)$
3. $y = 2 + 2.4(x - 1)$
4. $y = 1 + 0.27(x - 2)$



In the regression model ($y = a + bx$) where $\bar{x} = 2.50$, $\bar{y} = 5.50$ and $a = 1.50$ ($\bar{x}$ and $\bar{y}$ denote mean of variables x and y and a is a constant), which one of the following values of parameter 'b' of the model is correct?

1. 1.75
2. 1.60
3. 2.00
4. 2.50

There is no value of x that can simultaneously satisfy both the given equations. Therefore, find the 'least squares error' solution to the two equations, i.e., find the value of x that minimizes the sum of squares of the errors in the two equations. ________

$2x = 3$

$4x = 1$


We can expect
one
Question from
here in
GATE exam

Considering data of 2 Dimensions

Attributes, Features, Dimensions...

Till now we have seen a simple case of 1 D data, now let's see 2 D Data

Income (LPA)	Age	Sale of I-Phone (in a month)
20	30	300
50	40	400
70	50	300
We have N Data points		

Now the input data is 2 D (age and income)

THANK - YOU