Data Science and Artificial Intelligence

Machine Learning

Regression

Lecture No. 02

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- Topic — Model
- Topic — optimization
- Topic — training Process
- Topic — ML-Problem statement.
- Topic

Topics to be Covered

- Topic — 1D Linear Regression
- Topic — Formulae,
- Topic — Questions
- Topic — mean, Variance, Covariance---
- Topic

Gate 2026.

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

WINSTON CHURCHILL

Basics of Machine Learning

REVISION

Fill in the blanks :

1. The target/Goal of the ML is _____ To learn Pattern of data, to predict y for a new x.
2. The best optimized model is that which minimize the error in _____ Training/available data
3. The problem with the simple model is _____ Underfit _____ (Kuch bhi padhai nahi ki)

Basics of Machine Learning

Fill in the blanks :

4. The problem with highly complicated model is _____Rotelearning, → Overfitting_____

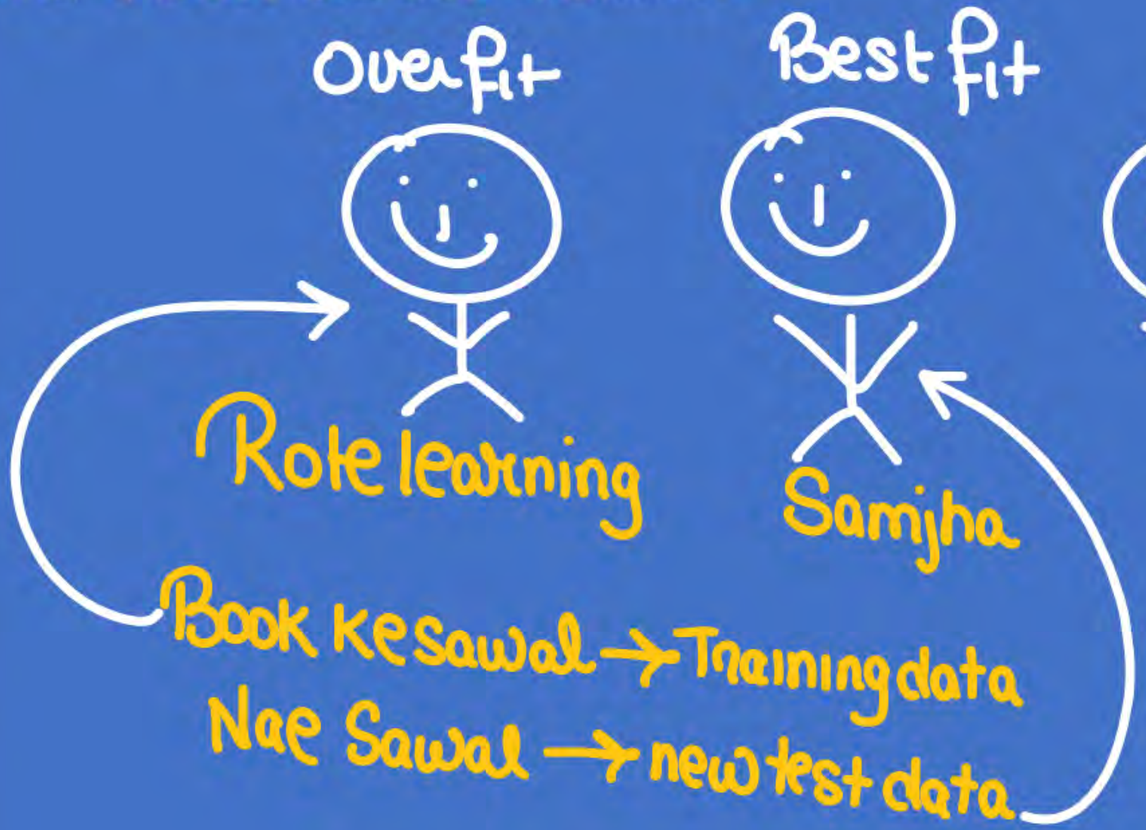
5. The data is used to _____Train_____ the ML model

6. The data is collected from _____Survey / experiment._____

Basics of Machine Learning

REVISION

Wo Teen bacchhe.....



⇒ Perform Poor in Both.

Basics of Machine Learning

What is a Residue

error ⇒ diff b/w the actual and Predicted Value → $(y-y')$

$(y-y')$ Residue / error

$|y-y'|$ ⇒ Absolute error

$(y-y')^2$ ⇒ Square of error.

model $y=f(x)$ → So value from model ⇒ Predicted Value

REVISION

How we do optimization (Absolute error & RSS)

(N Training data points)

Refer

Best model

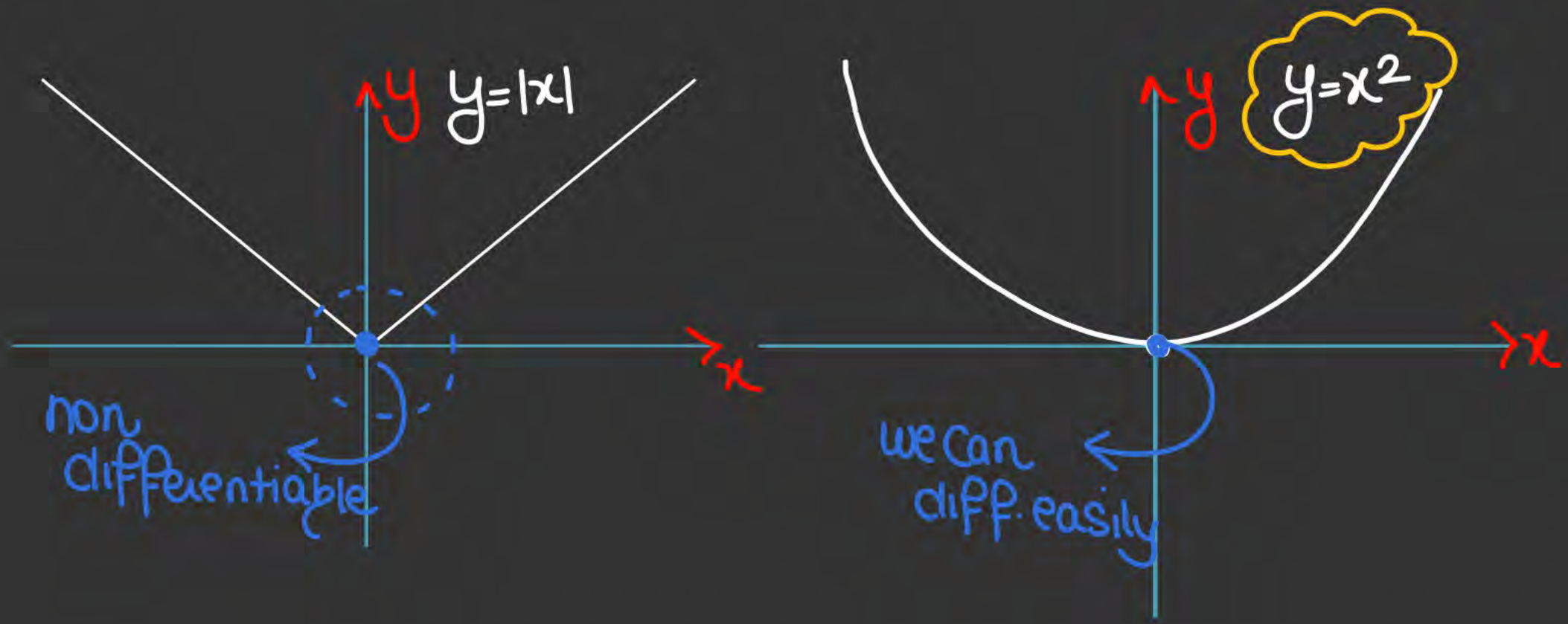
$$\min \sum_{i=1}^{N} |y_i - y_i'|$$

Sum of Absolute Error

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

Sum of Square of error
Sum of y y Residue

RSS
Residual Sum of Square


y
y=|x|
x
non
differentiable
y
y=x²
x
we can
diff. easily

Basics of Machine Learning

19. The output of training process in machine learning is

A. machine learning model ← we get model.

B. machine learning algorithm

C. null

D. accuracy

Basics of Machine Learning

34. In simple term, machine learning is

A. training based on historical data

B. prediction to answer a query

C. both a and b??

D. automization of complex tasks

$f(x) = 3x^2 + 4x + 5$

find x to min $f(x)$

$$\frac{d f(x)}{d x} = 0$$

$f(a,b) = 3a^2 + 4b^2 + 10ab + 40$

find a, b to min $f(a,b)$

$$\frac{\partial f(a,b)}{\partial a} = 0$$

When we do partial derivative with a, then Consider b → Const

$$\frac{\partial f(a,b)}{\partial b} = 0$$

here a is a Const.

Basics of Machine Learning

Problem 1 – Predict Population of bacteria in a lab

We must create a model with following data

model
→ $y = mx + c$

Time (x)	Population (y)
0	50
10	200

$y_{predicted}$
c ✓
10m + c ✓

To find m, c we ⇒ $\min \left[(y_1 - y_1')^2 + (y_2 - y_2')^2 \right]$

⇒ $\min \left[(50 - c)^2 + (200 - (10m + c))^2 \right]$

→ We have to minimize this fxn ⇒ 2 unknown.

Now predict the population at t = 20

$$\mathcal{L} = \min \left[(50-c)^2 + (200-(10m+c))^2 \right]$$

To min $\mathcal{L}$ ⇒

$$\frac{\partial \mathcal{L}}{\partial c} \Rightarrow \frac{\partial}{\partial c}\left[(50-c)^2 + (200-10m-c)^2 \right]$$

$$\Rightarrow 2(50-c)(-1) + 2(200-10m-c)(-1) = 0$$

$$-(100-2c) - (400-20m-2c) = 0$$

$$-100+2c-400+20m+2c = 0$$

① → $20m+4c = 500$ ⇐

$$5m+c = 125$$

$$L = \min \left[(50 - c)^2 + (200 - (10m + c))^2 \right]$$

Process ✓
Ans ✓
m = 15, C = 50.

To min $L \Rightarrow$

$$\frac{\partial L}{\partial m} \Rightarrow \cancel{2(50-c)0} + 2(200 - 10m - c)(-10) = 0$$

$$\cancel{-20}(200 - 10m - c) = \cancel{0}$$

$$10m + c = 200 \quad -(2)$$

$$5m + c = 125 \quad -(1)$$

(1) Subtract 2 - 1

$5m = 75$ | $C = 50$

$m = 15$

$y = 15x + 50$

@ $t = 20$, $y = 20 \times 15 + 50$

$y = 350$

Basics of Machine Learning

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

H.W

We must create a model with following data

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

Ypred
30m+c
40m+c

$y = mx + c$ model

↓ Sale of iphone

↘ age

$$d \Rightarrow \min \left[(300 - (30m+c))^2 + (400 - (40m+c))^2 \right]$$

$\partial d/\partial m = 0$ } $\partial d/\partial c = 0$ H.W ✓

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

Basics of Machine Learning

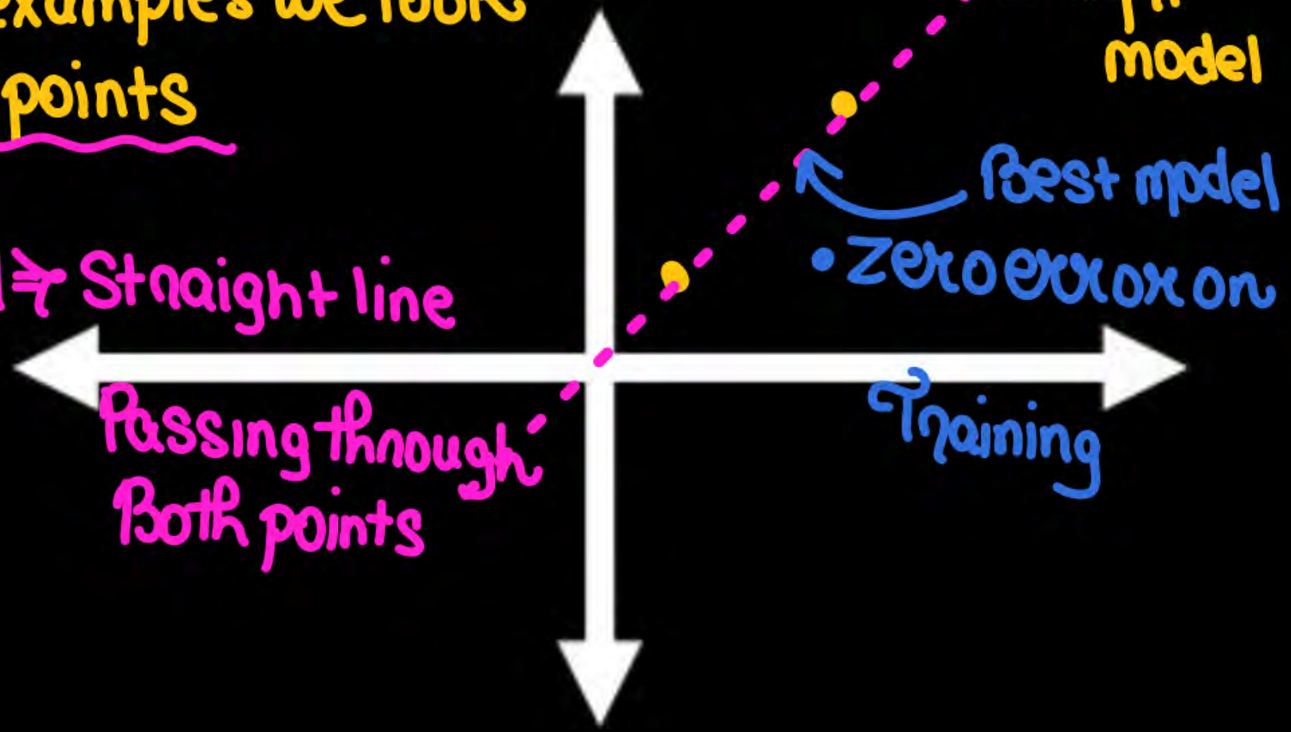
GAATE 2026 FODENGE

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 _____

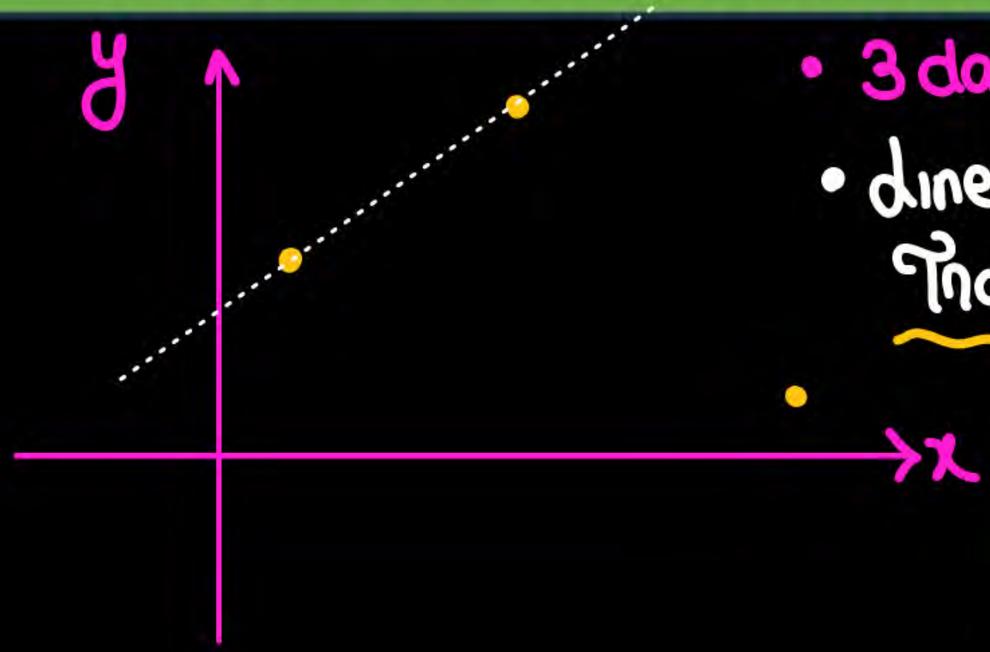
What is the best model now ?

*last 2 examples we took
2 data points
↓
model ≠ Straight line



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

Now we have to find the best parameters..



- 3 data points
- linear - straight line model

Training error ≠ 0

- So $y = mx + c$ ← General Process Valid

To find best m, c

$$\min \sum_{i=1}^{3} \left(y_i^{o} - y_i^{c}\right)^2$$

Basics of Machine Learning

GAATE 2026 FODENGE

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

Now we have to find the best parameters..

So To find best model we follow

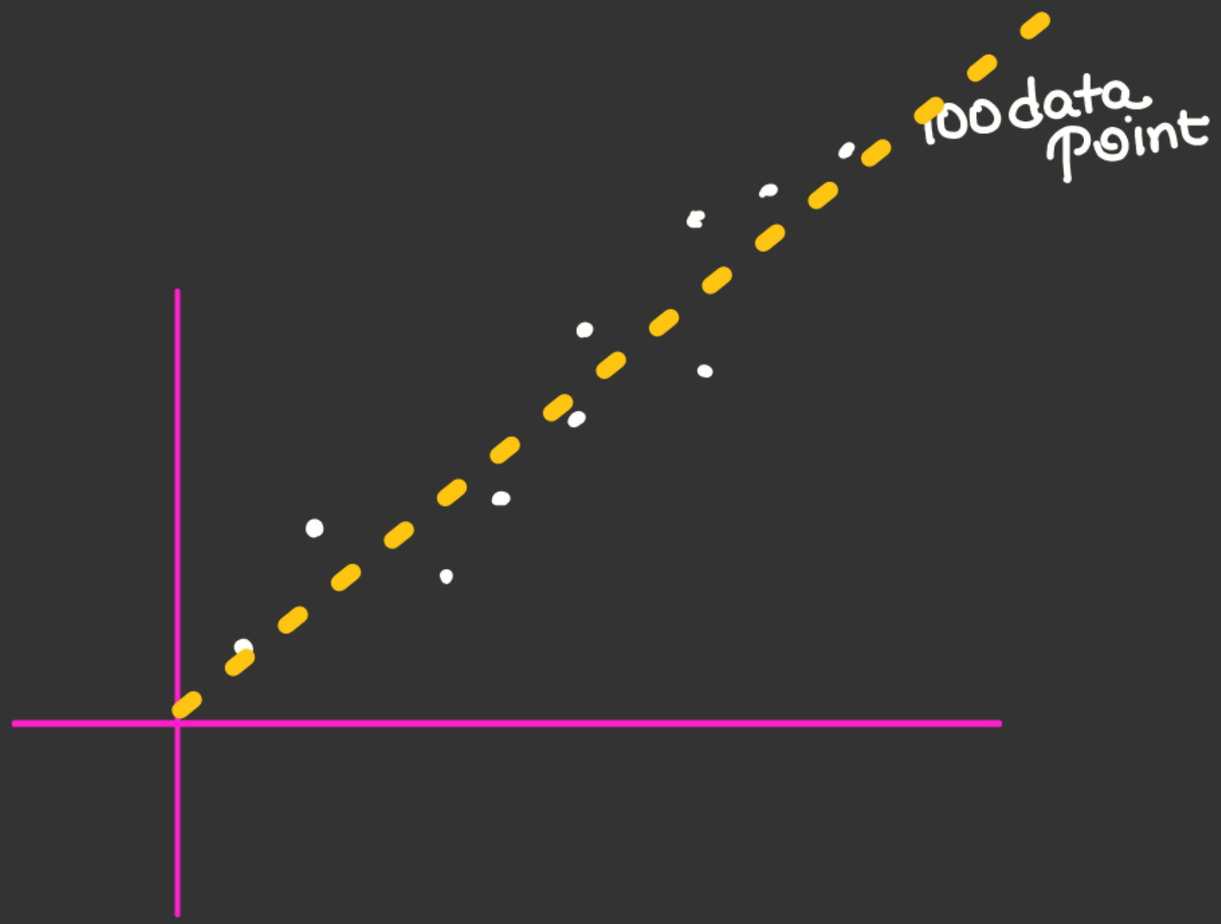
Same process

$y = mx + c$ (model)

$$L = \sum_{i=1}^{N} (y_i - y_i')^2$$

min L to get m, c

$$\frac{\partial L}{\partial m} = 0, \frac{\partial L}{\partial c} = 0$$


100 data Point

Basics of Machine Learning

GAATE 2026 FODENGE

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

We must create a model with following data

Age x	Sale of I-Phone (in a month) y
30	300
40	400
50	300

$y_{pred.}$

$30m+c$

$40m+c$

$50m+c$

model $y=mx+c$

$$d=\sum(y_i^o-y_i^{'})^2 \Rightarrow \left[(300-30m-c)^2+(400-40m-c)^2+(300-50m-c)^2\right]$$

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

$$L = \sum (y_i^0 - y_i^1)^2 \Rightarrow \left[(300 - 30m - c)^2 + (400 - 40m - c)^2 + (300 - 50m - c)^2 \right]$$

2 Variables m, c

$$\frac{\partial L}{\partial c} = \left[-2(300 - 30m - c) - 2(400 - 40m - c) - 2(300 - 50m - c) \right] = 0$$

$$300 - 30m - c + 400 - 40m - c + 300 - 50m - c = 0$$

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

$$L = \sum (y_i^o - y_i^l)^2 \Rightarrow \left[(300 - 30m - c)^2 + (400 - 40m - c)^2 + (300 - 50m - c)^2 \right]$$

2 Variables m, c

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

$$9000 - 900m - 30c + 16000 - 1600m - 40c$$

$$15000 - 2500m - 50c = 0$$

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

$$40000 = 5000m + 120c \quad -(11) \longrightarrow 1000 = 125m + 3c$$

$1000 = 120m + 3c$

Subtract

$5m = 0$

$m = 0$

$c = 1000/3$

$y = \frac{1000}{3}$

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

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

$(-1, -2), (0, 0), (1, 3)$ ← (x, y) data given

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

PW

$y = ax + b$ model

x	y (actual)	y_p (Pred)
-1	-2	$-a+b$
0	0	b
1	3	$a+b$

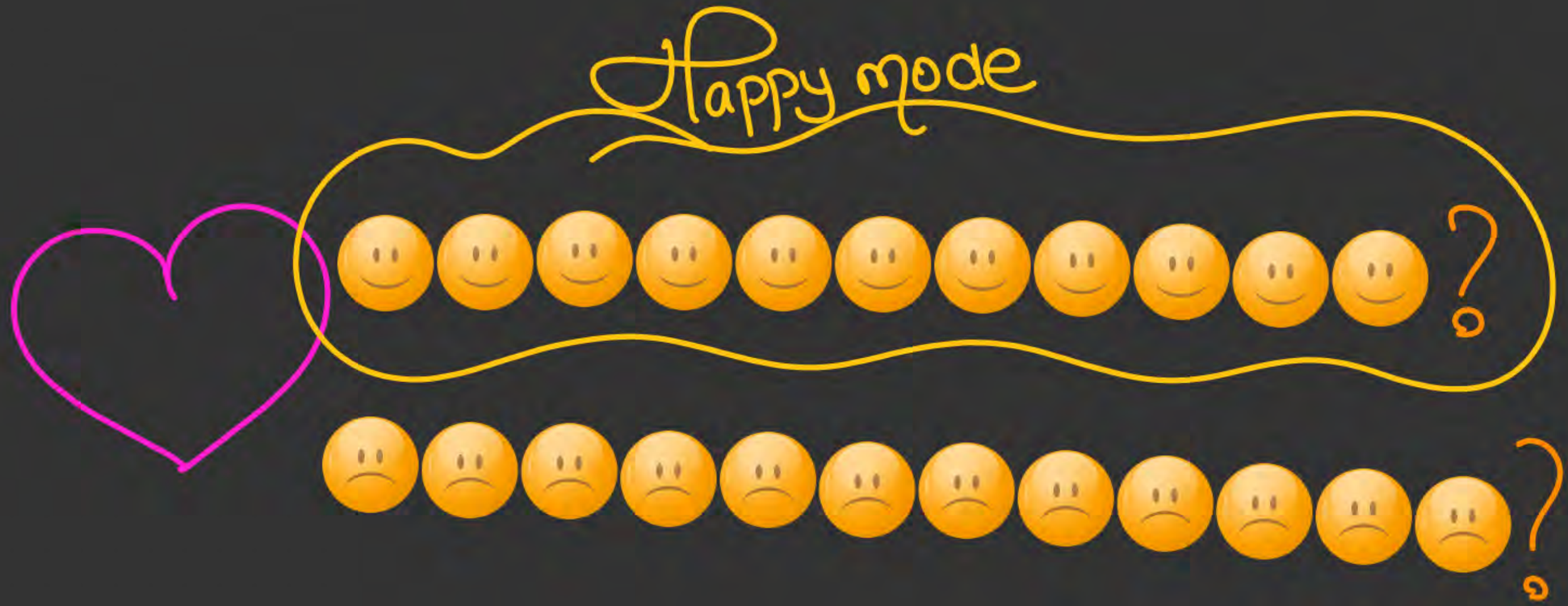
$$L = \sum (y_p - y_i)^2$$

$$L = (-2-(-a+b))^2 + (0-b)^2 + (3-(a+b))^2$$

min L to get a, b

$\frac{\partial L}{\partial a} = 0$ ✓

$\frac{\partial L}{\partial b} = 0$ ✓


Happy mode

Basics of Machine Learning

RSS → done

GATE 2026 FODENGE

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

Creating the best model

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

$$L = \sum_{i=1}^{N} (y_i - y_i')^2$$

This RSS is also called loss fxn

$\min L$

Read

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**

- for diff algo loss fxn are different.

Basics of Machine Learning

GAATE 2026 FODENGE

Now how to find the best parameters ??

mark	Student
40	1
45	2
43	3
44	4

- Std deviation ⇒ $\sqrt{\text{Variance}}$

- mean value of marks ⇒ $\frac{\text{Sum of all Values}}{\text{Total No Of Values}}$

Variance and mean...

mean ⇒ $\frac{40+45+44+43}{4}$

⇒ 43

Variance ⇒ $\frac{(43-40)^2+(43-45)^2+(43-43)^2+(43-44)^2}{4}$

⇒ $\frac{14}{4}$

Basics of Machine Learning

Now how to find the best parameters ??

Expression

dimension

label

data

X	Y
x_1	y_1
x_2	y_2
x_3	y_3
⋮	⋮
x_N	y_N

data value

- @ each point we have single data value

→ 1 Dimension data
1D-data

Variance and mean...

So $\overline{X}$ ⇒ we have N values of x

$\overline{X}$ ⇒ $\frac{\text{add all x values}}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$

σ_x^2 : Variance of X ⇒ $\frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N}$

σ_x : Std dev of X = $\sqrt{\text{Var}}$

Basics of Machine Learning

Now how to find the best parameters ??

Expression

dimension

label

data

X	Y
x_1	y_1
x_2	y_2
x_3	y_3
⋮	⋮
x_N	y_N

data value

- @ each point we have single data value → 1 Dimension data / 1D-data

Variance and mean...

So $\overline{Y}$ ⇒ we have N values of y

⇒ $\frac{\text{add all y values}}{N} = \frac{\sum_{i=1}^{N} y_i}{N}$

σ_Y^2 : Variance of y ⇒ $\frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{N}$

σ_y : Std dev of y = $\sqrt{\text{Var}}$

Basics of Machine Learning

GAATE 2026 FODENGE

Now how to find the best parameters ??

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

Variance and mean...

Basics of Machine Learning

Now how to find the best parameters ??

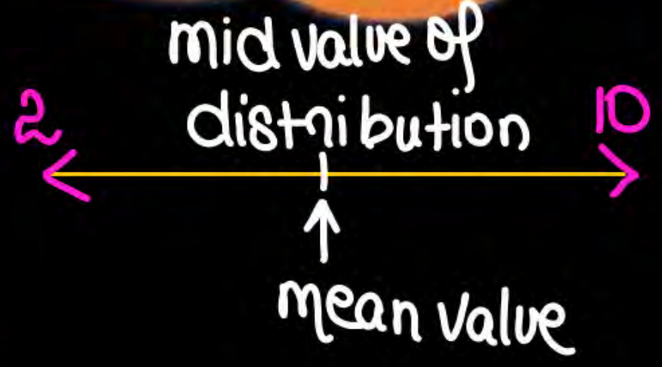
- What is mean value of a variable ⇒
  - Expectation value

- What is variance of a variable ⇒
  - Spread of data wrt mean

High variance → more spread

Low variance → less spread

Formulae to find direct value of m and c



Basics of Machine Learning

GAATE 2026 FODENGE

Now how to find the best parameters ??

Formulae to find direct value of m and c

Covariance → Show relation b/w 2 variables x, y

x inc, y inc → Cov +ve

x dec, y dec → Cov +ve

x inc, y dec → Cov -ve

x dec, y inc → Cov -ve

Now how to find the best parameters ??

For 1D data

Linear line $y = mx + c$

Direct formula

$$\Rightarrow \begin{cases} \bullet\ m = \dfrac{Cov(x,y)}{Var(x)} \\ \bullet\ c = \bar{y} - m\bar{x} \end{cases}$$

Formulae to find direct value of m and c

Basics of Machine Learning

GAATE 2026 FODENGE

PW

Now how to find the best parameters ??

Formulae to find direct value of m and c

Basics of Machine Learning

$$Var(y) = \frac{(10-5)^2+(10-7)^2+(10-9)^2+(10-12)^2+(10-17)^2}{5} \Rightarrow 17.6$$

Example

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

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

- $N \Rightarrow 5$
- $\bar{X} = \frac{2+3+4+5+6}{5} \Rightarrow 4$.
- $\bar{Y} = \frac{5+7+9+12+17}{5} \Rightarrow 10$

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

$$Cov(X,Y) \Rightarrow \frac{(2-4)(5-10)+(3-4)(7-10)+(4-4)(9-10)+(5-4)(12-10)+(6-4)(17-10)}{5} \Rightarrow (5.8)$$

Example

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

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

- $N \Rightarrow 5$
- $\bar{X} = \frac{2+3+4+5+6}{5} \Rightarrow 4.$
- $\bar{Y} = \frac{5+7+9+12+17}{5} \Rightarrow 10.$

$Cov(X,Y) = 5.8$

$$y = mx + c$$

$$m = \frac{Cov(x,y)}{Var\,x} = \frac{5.8}{2} = 2.9$$

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

$$c = -1.6$$

THANK - YOU