Data Science and Artificial Intelligence

Machine Learning

Regression

Lecture No. 03

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

Topic — Mean

Topic — Variance

Topic — 1D data → $y = mx + c$

Topic — m, c

Topic

Covariance

Topics to be Covered

Topic — data with more than 1D

Topic — we will see St. line for data more than 1D

Topic

Topic

Topic

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

GATE 2026 FODENGE

Inner Motivation

REVISION

1. What is the Loss Function

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

we minimize loss fxn.

REVISION

3. Direct formulae for M and C.

$(y = mx + c)$

St. line for 1D data

No of Parameters $\Rightarrow$

2 Parameters m, c

dimension.

1D data

x	y
x_1	y_1
x_2	y_2
x_3	y_3
⋮	⋮

$y \rightarrow$ label

- $m = \dfrac{Cov(x, y)}{Var\, x}$
- $c = \bar{y} - m\bar{x}$

REVISION

4. Covariance :

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

5. Variance : $$Var(x) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$$

6. So the data that we were using has ___1___ number of dimensions and the straight line obtained by liner regression has ___2___ number of parameters.

What is Correlation Coefficient

X, Y are 2 variables

$$\rho_{XY} = \frac{Cov(x,y)}{Std.x \; Std_y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

(i) Value -1 to 1

(ii) Value = ±1, then X, Y are highly Correlated

(iii) Value = 0, X, Y are un Correlated

Representing the two equations in Matrix format

Eq 1

$aw + bz = c$
$dw + ez = f$

→ In matrix form

$$\begin{bmatrix} a & b \\ d & e \end{bmatrix} \begin{bmatrix} w \\ z \end{bmatrix} = \begin{bmatrix} c \\ f \end{bmatrix}$$

$$\begin{bmatrix} aw + bz \\ dw + ez \end{bmatrix} = \begin{bmatrix} c \\ f \end{bmatrix}$$

Linear Regression

GAATE 2026 FODENGE

Representing the two equations in Matrix format

1D data ✓

N Points	X	Y
	x_1	y_1
	x_2	y_2
	x_3	y_3
	⋮	⋮

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

Let fxn is $y = mx + c$ for any x, $y_i' = mx_i + c$

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

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

No Need to Remember

Steps hein No need to Remember

$$\frac{\partial L}{\partial c} \Rightarrow \cancel{2} \sum_{i=1}^{N} (y_i - mx_i - c)(\cancel{-1}) = 0$$

$$\Rightarrow \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} m x_i - \sum_{i=1}^{N} c = 0$$

$$\Rightarrow \left\{ m \sum_{i=1}^{N} x_i + c \sum_{i=1}^{N} 1 = \sum_{i=1}^{N} y_i \right\} \quad (1)$$

$$\frac{\partial L}{\partial m} = \cancel{2} \sum_{i=1}^{N} (y_i - mx_i - c)(\cancel{-}x_i) = 0$$

$$\sum_{i=1}^{N} \left(y_i x_i - m x_i^2 - c x_i \right) = 0$$

$$\left\{ m \sum_{i=1}^{N} x_i^2 + c \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i x_i \right\} \quad (2)$$

$$\Rightarrow \begin{cases} m\sum_{i=1}^{N} x_i + c\sum_{i=1}^{N} 1 = \sum_{i=1}^{N} y_i \\ m\sum_{i=1}^{N} x_i^2 + c\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i x_i \end{cases}$$

$$\left.\underbrace{\begin{bmatrix} \sum_{i=1}^{N} 1 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix}}_{X^T X} \begin{bmatrix} c \\ m \end{bmatrix} = \underbrace{\begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} y_i x_i \end{bmatrix}}_{X^T Y} \right\} \Rightarrow \text{(No need to Remember)}$$

No need.

GAATE 2026 FODENGE

PW

Representing the two equations in Matrix format

$y = mx + c$ → Function that give Predicted Value

Data

$$\begin{bmatrix} x & y \\ x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \\ \vdots & \vdots \end{bmatrix}$$ → actual values.

$y = mx + c$ → using matrix must.

$y_1' = mx_1 + c \Rightarrow \begin{bmatrix} 1 & x_1 \end{bmatrix} \begin{bmatrix} c \\ m \end{bmatrix}$

$y_2' = mx_2 + c \Rightarrow \begin{bmatrix} 1 & x_2 \end{bmatrix} \begin{bmatrix} c \\ m \end{bmatrix}$

1D data

$$\begin{bmatrix} x & y \\ x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \\ \vdots & \vdots \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & x_4 \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{bmatrix}$$

Column of 1

Let us have 4 Point 1D data

$$\begin{bmatrix} x & y \\ x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \\ x_4 & y_4 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \end{bmatrix}$$

$$X^T X \Rightarrow$$

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix}$$

$$\begin{bmatrix} \sum_{i=1}^{4} 1 & \sum_{i=1}^{4} x_i \\ \sum_{i=1}^{4} x_i & \sum_{i=1}^{4} x_i^2 \end{bmatrix}$$

Let us have 4 Point 1D data

$$\begin{bmatrix} x & y \\ x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \\ x_4 & y_4 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \end{bmatrix}$$

$X^T Y \Rightarrow$

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

$$\begin{bmatrix} \sum_{i=1}^{4} y_i \\ \sum_{i=1}^{4} x_i y_i \end{bmatrix}$$

Conclusion

Bas itna hi dimag mein Rakho $\Rightarrow$

1D data

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

model $\rightarrow$ $y = \underbrace{\beta_0}_{c} + \underbrace{\beta_1}_{m} x$

2 Parameters β_0, β_1

$\rightarrow \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{bmatrix}$$

$$\boxed{\begin{aligned} (X^TX)\,\beta &= X^TY \\ \beta &= (X^TX)^{-1}(X^TY) \checkmark \end{aligned}}$$

$$A = \begin{bmatrix} a & b \\ c & a \end{bmatrix}_{2\times 2}$$

$$\boxed{A^{-1} = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \frac{1}{ad-bc}}$$

Linear Regression

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

$y = mx + c$ find m, c

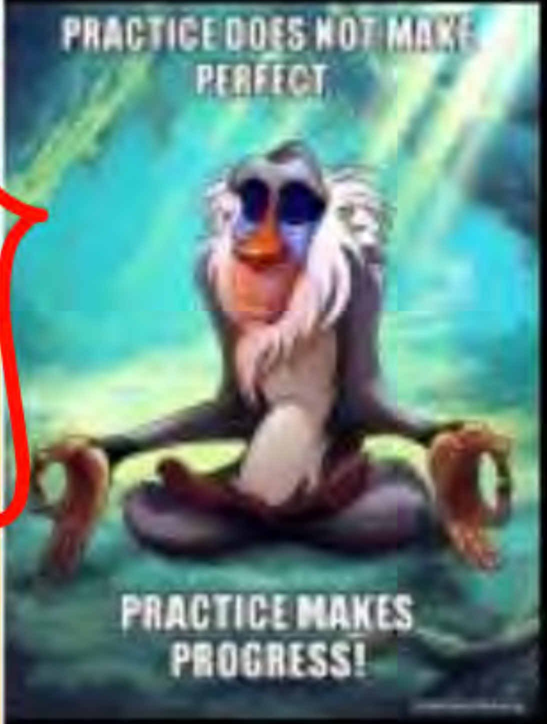
Valid

$\text{Cov}(x,y)/\text{Var}\,x$

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

$$X = \begin{bmatrix} 1 & 5 \\ 1 & 2 \\ 1 & 4 \\ 1 & 3 \end{bmatrix} \quad Y = \begin{bmatrix} 16 \\ 10 \\ 13 \\ 12 \end{bmatrix} \quad X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 5 & 2 & 4 & 3 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 4 & 14 \\ 14 & 54 \end{bmatrix} \quad X^T Y = \begin{bmatrix} 51 \\ 188 \end{bmatrix}$$



$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \begin{matrix} \to C \\ \to m \end{matrix}$$

$$X^T X (\beta) = (X^T Y)$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = (X^T X)^{-1} (X^T Y) \Rightarrow$$

$$\begin{bmatrix} 54 & -14 \\ -14 & 4 \end{bmatrix} \frac{1}{54 \times 4 - 14^2} \begin{bmatrix} 51 \\ 188 \end{bmatrix}$$

$$\Rightarrow \frac{1}{20} \begin{bmatrix} 54 & -14 \\ -14 & 4 \end{bmatrix} \begin{bmatrix} 51 \\ 188 \end{bmatrix} \Rightarrow$$

$$\frac{1}{20} \begin{bmatrix} 122 \\ 38 \end{bmatrix}$$

$$\beta_0 = C = 122/20$$

$$m = 38/20$$

Linear Regression

1D data

$y = (-.4 + 2.4x)$

$y = 2 + 2.4(x-1)$

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

$\bar{X} = 1 \quad \sigma_x = 3$

$\bar{Y} = 2 \quad \sigma_y = 9$

$\rho_{xy} = .8$

$\beta_0 = \bar{y} - \beta_1 \bar{x}$

$= 2 - 2.4(1)$

$= -.4$

$y = \beta_0 + \beta_1 x$

$\rho_{xy} = \dfrac{Cov(x,y)}{\sigma_x \sigma_y}$

$.8 = \dfrac{Cov(x,y)}{3 \times 9}$

$Cov(x,y) = 9 \times 3 \times .8 = 21.6$

$\beta_1 = \dfrac{Cov(x,y)}{Var(x)} = \dfrac{21.6}{\sigma_x^2} = \dfrac{21.6}{9} \Rightarrow 2.4$

In the regression model $\boxed{(y = a + bx)}$ 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 $\boxed{\text{parameter 'b'}}$ of the model is correct?

Simple: $a = \bar{y} - b\bar{x}$

$1.5 = 5.5 - b(2.5)$

$b = 1.6$

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

GAATE 2026 FODENGE

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. ________

$2x = 3$ ✓

$4x = 1$

$2x=3$

$4x=1$

we can see that Single x Cannot Satisfy both.

error

Let x' is value of x that min the Sum of Square error

$$\left\{\begin{array}{l} 2z=3 \\ 4z=1 \end{array}\right\}$$

Actual

Actual

$2z'$

$4z'$

$(z = z')$

$\text{error} \Rightarrow (2z'-3)$

$\text{error} \Rightarrow (4z'-1)$

$$L = (2z'-3)^2 + (4z'-1)^2$$

$$\frac{\partial L}{\partial z'} \Rightarrow 2(2z'-3)\times 2 + 2(4z'-1)\times 4 = 0$$

$$8z' - 12 + 32z' - 8 = 0$$

$$40z' = 20, \; z' = .5$$

Linear Regression

Q. let $5z = 1$

$2z = 5$

$10z = 3$

find z that minimize the sum of square of error in above eq ⇒

let z' is the correct value

$$L \Rightarrow (5z'-1)^2 + (2z'-5)^2 + (10z'-3)^2$$

$$\min L \Rightarrow 2(5z'-1) \times 5 + 2(2z'-5) \times 2 + 2(10z'-3) \times 10 = 0$$

$$50z' - 10 + 8z' - 20 + 200z' - 60 = 0$$

$$258z' = 90, \quad z' = 90/258 = 15/43$$

Linear Regression

Attribute / features

Considering data of $\boxed{\text{2 Dimensions}}$

Attributes, Features, Dimensions...

2 Dimension
2 Attributes.

Label y

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)

GAATE 2026 FODENGE

How to write the 2 D inputs ??

1D
$y = \beta_0 + \beta_1 x$

Now I want a linear model :

$$y = \underbrace{\beta_0 + \beta_1 (\text{income}) + \beta_2 (\text{age})}$$

→ 2D data
→ 2 Attributes data

→ Parameters ⇒ 3 Parameters.

GAATE 2026 FODENGE

Linear model will have ________ number of parameters

2D data Representation

dataPoint	2 dimension x^1	x^2	label y
1	x_1^1	x_1^2	y_1
2	x_2^1	x_2^2	y_2
3	x_3^1	x_3^2	y_3
4	x_4^1	x_4^2	y_4
⋮			

dimension → Superscript

datapoint → Subscript

model ⇒

$$\Rightarrow \left(y = \beta_0 + \beta_1 x^1 + \beta_2 x^2\right)$$

Linear model will have ________ number of parameters

2D data Representation

	2 dimension		label
dataPoint	x^1	x^2	y
1	x_1^1	x_1^2	y_1
2	x_2^1	x_2^2	y_2
3	x_3^1	x_3^2	y_3
4	x_4^1	x_4^2	y_4
⋮			

$$X = \begin{bmatrix} 1 & x_1^1 & x_1^2 \\ 1 & x_2^1 & x_2^2 \\ 1 & x_3^1 & x_3^2 \\ \vdots & \vdots & \vdots \end{bmatrix}_{N\times 3} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \qquad \begin{bmatrix} X^T \end{bmatrix}_{3\times N}$$

$$\boxed{\begin{array}{l} (X^TX)\beta = (X^TY) \quad \text{Valid} \\ \bullet\ \beta = (X^TX)^{-1}(X^TY) \end{array}}$$

GAATE 2026 FODENGE

The optimisation method and equation will be ...

3D $\Rightarrow$

dataPoint	x^1	x^2	x^3	y
1	x_1^1	x_1^2	x_1^3	y_1
2	x_2^1	x_2^2	x_2^3	y_2
3				
⋮				

$$X = \begin{bmatrix} 1 & x_1^1 & x_1^2 & x_1^3 \\ 1 & x_2^1 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}_{N\times 4}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \end{bmatrix}_{N\times 1} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$\beta = (X^T X)^{-1} (X^T Y)$$

THANK - YOU