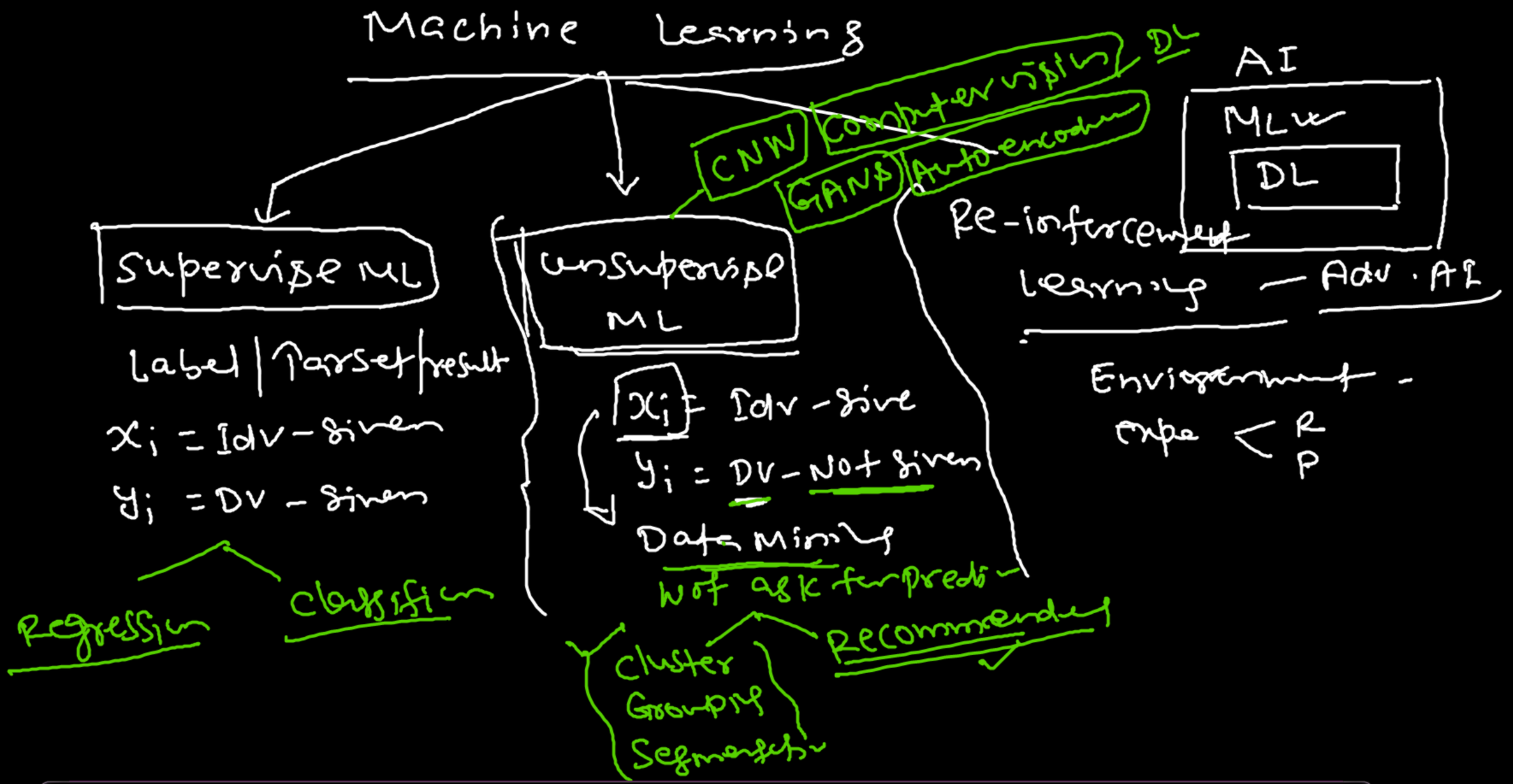
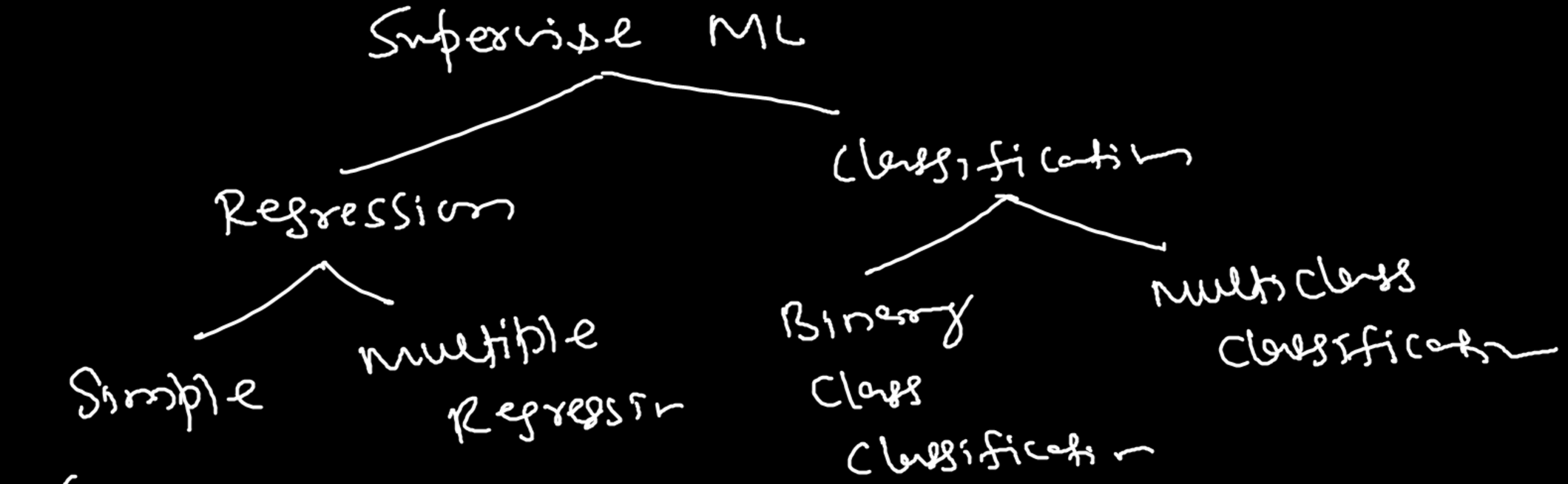




- Linear Regression
  - → Easy to understand
  - → Simple & elegant
  - → Geometrical, Loss approach

Linear Regression – Actual regression Problem

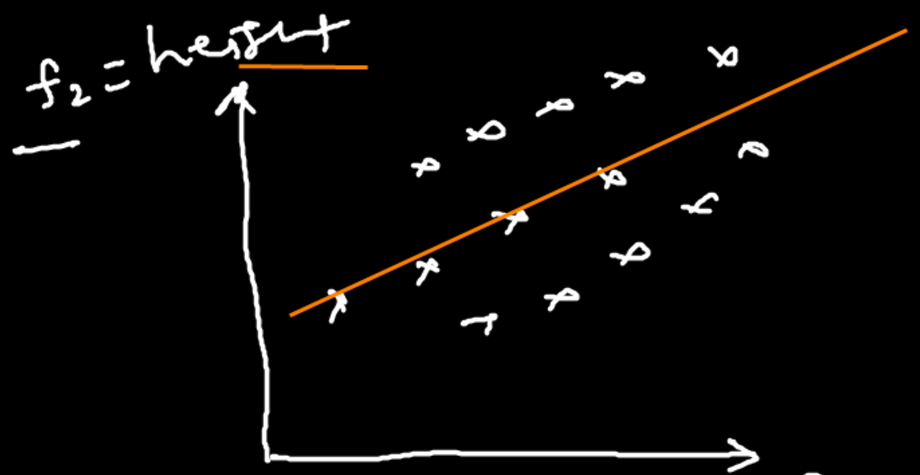
- Simple LR → only one Idv
- Multiple LR → more than one Idv
- Polynomial LR

Cgpa	Package

Exp	Salary

Geometrical intuition of Linear Regression

Eg:- Predict height $\in \mathbb{R}$ given weight, Gender, ethnicity hair color etc.



Task:- find a line/plane that fits the given data

$Y = mx + c$

height = w_0 + w_1 * weight

$y = c + mx_1$

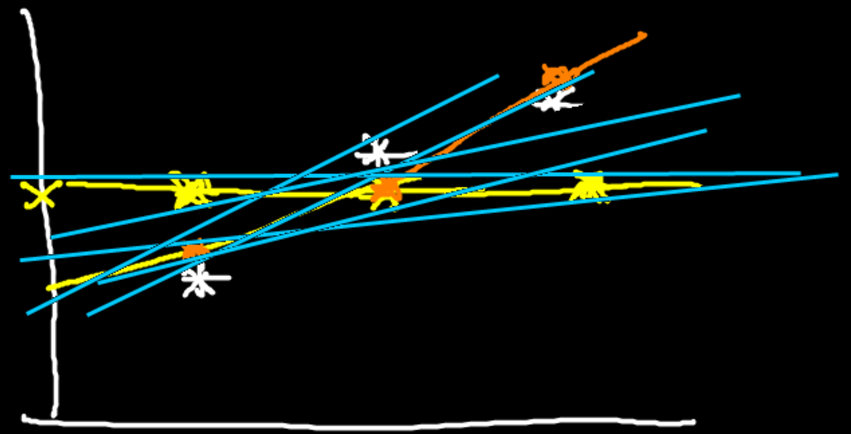
$Ax_1 + By_1 + C = 0$ – General eqn.

$Ax_1 + By_1 + C$

x_1 & y_1 = variables
A, B = co-eff
C = Int = Constant

(x) - given (y) - predict

Exp	Salary	Avg
2	10800	12500
3	13000	12500
4	14500	12500
0	?	
10	?	
5	?	



Slope = ?

Intercept = ?

Avg Sal = 10 + 13 + 14.

$$\text{Salary} = \text{Intercept}\ (w_0) + \text{slope}(w_1) * \text{Exp} \quad \text{- eqn.}$$

Salary = ?

$$Y = w_0 + w_1 * x$$

w_0 & w_1 = ?

Exp = given

$$\text{Slope / co-efficient} = \frac{\sum_{i=1}^{n} (Cov(x) * Cov(y))}{\sum_{i=1}^{n} Cov(x)^2} = \frac{4500}{2} = 2250$$

$$Cov(x) = (x_{actual} - x_{Avg})$$

$$Cov(y) = (y_{act} - y_{avg})$$

$$y = mx + c$$

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

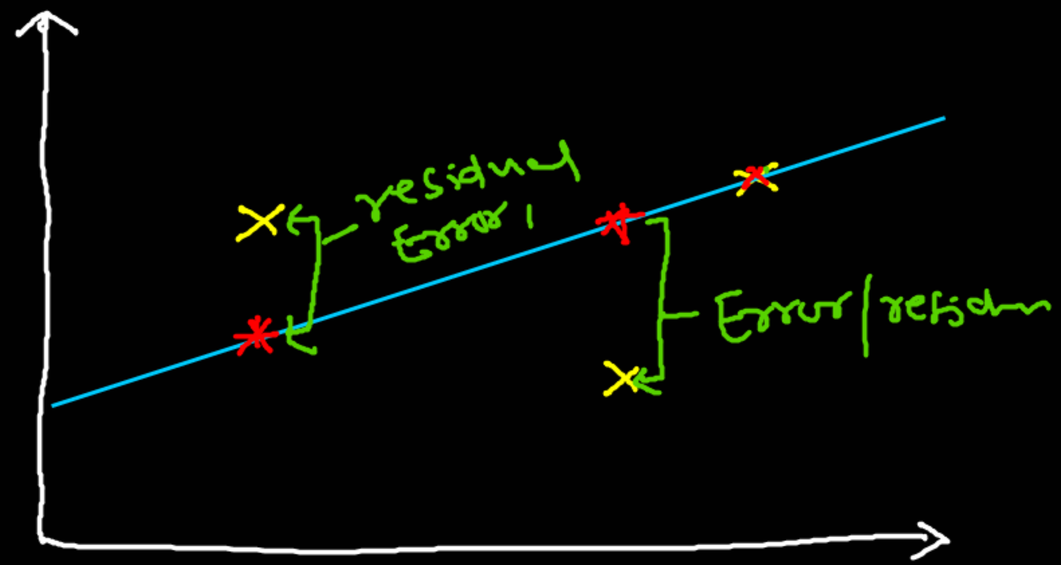
$$y_{Avg} = 12500$$

$$x_{Avg} = 3$$

$$m = 2250$$

$$c = 12500 - 2850 * 3$$

Find a line/plane that best fits the data point



Yellow color = Actual data
Red color = Predicting

Error 1 = Actual − Predicted
(Error 1)² = +ve = +ve
Error 2 = Actual − Predicted
(Error 2)² = −ve = +ve
Error 3 = Act − Pred
(Error 3)² = zero = zero

To remove error, we have to use square

$|-3| = 3$

Gradient

$|x_1 - x_2| = abs$ $\quad \frac{\partial |x|}{\partial x} =$ Can't find it

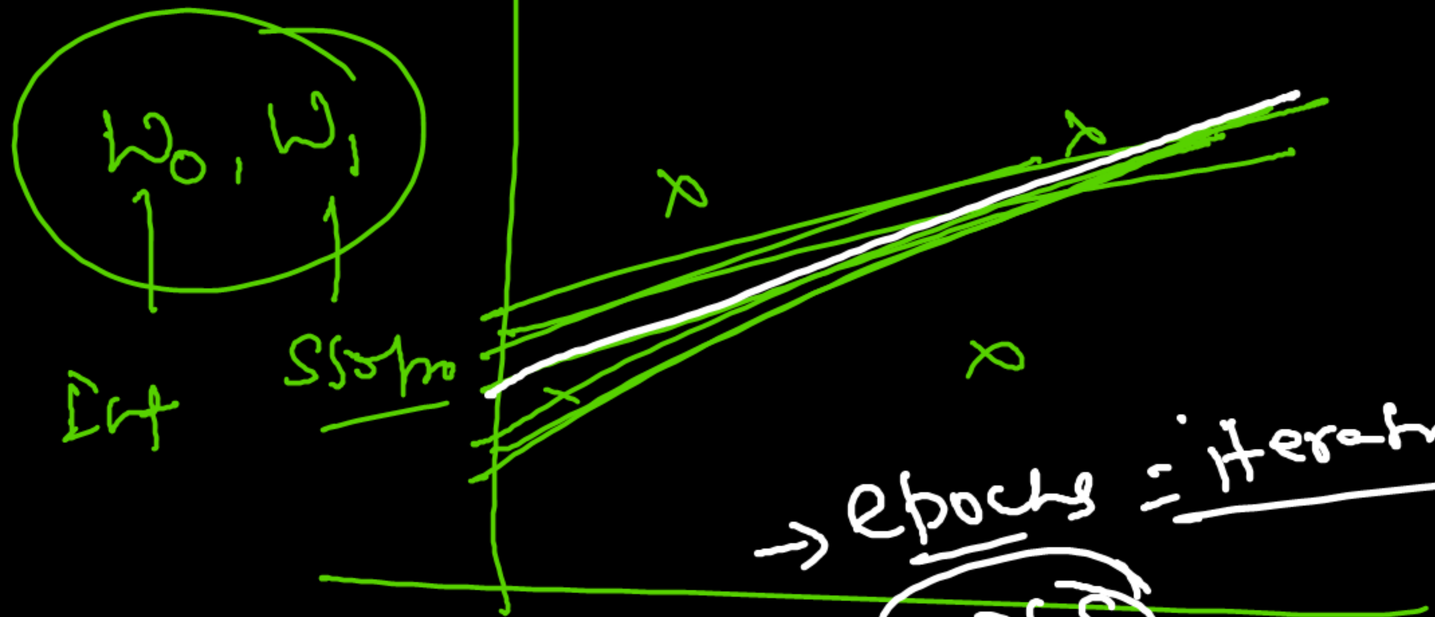
$(x_1 - x_2)^2 = sq.$ $\quad \frac{\partial (x^2)}{\partial x} = 2x$

Gradient descent

hyperparam — Learning rate (0 to 10)

$$\text{new weight} = \text{old weight} - \eta * \frac{\partial L}{\partial W_{old}}$$

new weight — old weight — slope



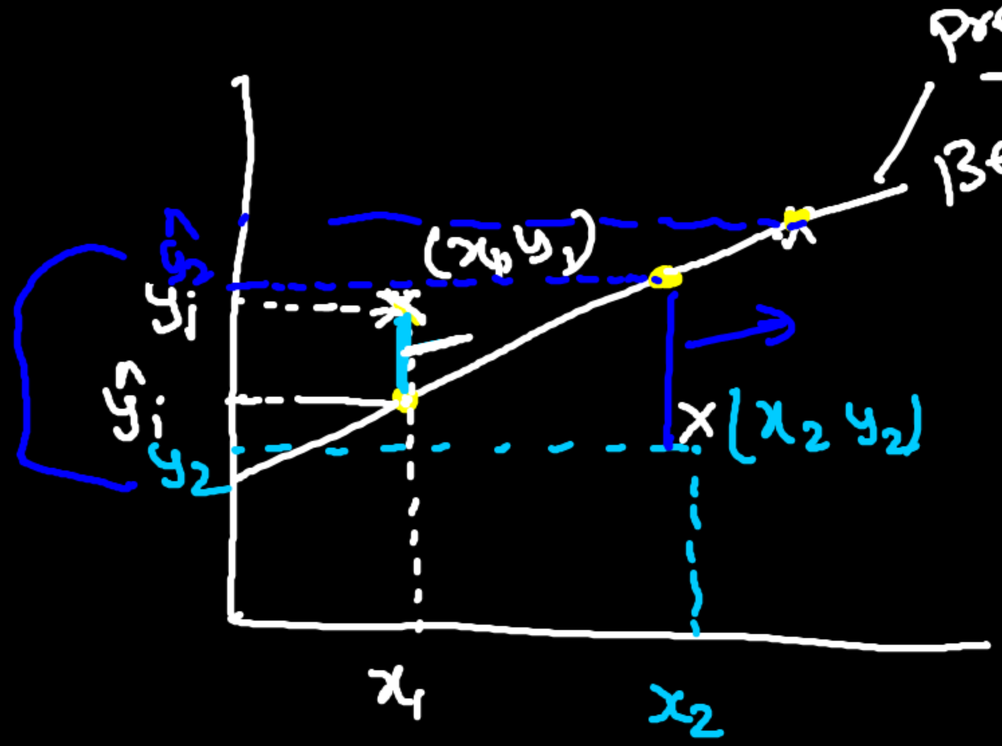
error1 = 500
error2 = 300 — best fit

min Sum of error across our training data

→ epochs = Iteration

OLS

2250 / 2000 / 1800 — Slope



error/residual $= (y_1 - \hat{y}_1)^2 = 3$

error/residual 2 $= (y_2 - \hat{y}_2)^2 = 4$

error 3 $= (y_3 - \hat{y}_3)^2 = 0$

7

Concept = Min Sum of error = best fit line

$\hat{y}_i$

$y = w_1 x_1 + w_0$ Int $= w_0$ Idv $= x_1 =$ SLR

Int $= w_0$ Slope $= w_1, w_2 w_3$

$y = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_0$ Idv $= x_1 x_2 x_3$ – MLR

Linear Regression

Algorithm or

OLS : Ordinary Least Square

LLS = Linear Least Square

$$w_0, w_1^* = \underset{w_0\ w_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

error is min, best fit line

Scaler (Intercept)

vector (Slope)

Input = given

find = w_0 & w_1

$$\hat{y}_i = w_i^T x_i + w_0$$

$y = mx + c$ $\quad m = w_1, \; c = w_0$

dot product $[w_1 \; w_2 \; w_3 \; w_4 \; \dots \; w_n] \; [x_1 \; x_2 \; x_3 \; \dots \; x_n]$

$w_1^T x_i$

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} [x_1 \; x_2 \; x_3 \; \dots \; x_n]$$

$w_1 x_1 + w_2 x_2 + w_3 x_3 + \dots$

$$\hat{y}_i = w_i^T x_i + w_0$$ — Notation

Optimization method — SQ-loss function

$$(w_i^*, w_0) = \underset{w^0, w_i^*}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{ y_i - (w_i^T x_i + w_0) \right\}^2$$

Part 1 - Regularization
- Lasso
- Ridge
- ElasticNet

Part 2 - Gradient descent

Part 3 :- Assumption :
- Linearity
- Normality of Residual
- Homoscedasticity
- No Autocorrelation
- No or little Multicollinearity
- No endogenity Problem

Linear Reg
- OLS / SS-Loss
- R-Sqrd / Adj-R-Sqrd
- P-Value

Actual Code Run + case study Multiple LinearReg

Life cycle of model

1. collect/extract data
2. Preprocessing of the data
3. Divide the dataset into train & test
4. EDA — Perform Descriptive analytics
5. Build Regression model
6. Estimate regression Parameter (Int, slope, ...)
7. Perform regression model diagnostics
8. Validate the model by using test data
9. Deployment on the cloud -

Df (200 data point)

x = Idv = Cgpa
y = dv = package

70%

30%

Building a model ← Train

140

1	6	8
3	8	9
4	9	9.5
5	5	7

SST, SSR, R2, mse

IDV → X_train

DV → y_train

Test ← Predict

60

x_test

2	7	8.9
6	8.	6.5

DV

y_test

OLS

7 = ?

8 = ?

70%	30%	by default
75%	25%	
80%	20%	
65%	35%	

	Cgpa	Package	
Train 1	6	8	
2	7	8.9	Test
Train 3	8	9	
Train 4	9	9.5	
Train 5	5	7	
6	8	6.5	Test

ML - Problems

1. Underfitting Problem | high bias
2. Overfitting Problem | high variable

training	test	variance	
~90-%	~70-%.	20-%.	Overfitting Prob | high variance more than 10%. — high variance high variance
~75-%.	~90%	15-%.	
~50-%.	~50%	less than 70%.	Underfitting Problem | high bias
~90-%.	~90	Actual fit	Target Actual fit

Data Leakage