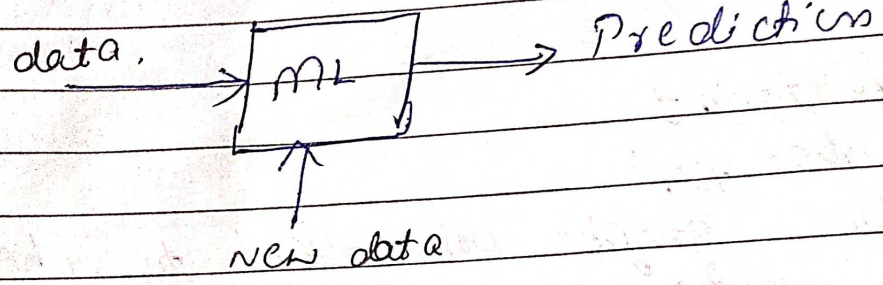
Machine learning

Machine learning is a branch of computer science which has several machine algorithms which learns from past data and makes predictions



Types of Machine learnings

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement [self driven or learns from punishment and rewards]

① Supervised learning :-

Train machine learning algorithms with both input & output variables such algorithms are called as supervised learning

Ex: predicting house prices, credit score etc.

x — input variable — featuring independent

y — output variable — target / label dependent variables

② Unsupervised learning :-

If we train machine learning with only input variable, such algorithms are unsupervised learning.

Ex: Market basket Analysis, clustering of customer feedback, market segmentation.

③ Reinforcement learning :-

is a machine learning training method is based on rewarding and punishing

Ex: self driven car

~~use~~

→ Supervised
- Regression
- Classification

a) Regression :- If target is continuous
Ex: house price prediction
flight price prediction

b) Classification :- If target is discrete or categorical.
Ex: Diabetic, Ratings, covid19

Linear Regression :-

- It comes under supervised Learning
- It comes under regression and it is used to predict the continuous target variable

Imp
- Linear Regression is used to find linear relationship b/w independent (x) and dependent (y) variables by fitting/drawing straight line.

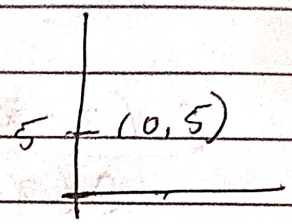
Types of Regression:

① **Simple linear regression:** If an algorithm uses single input variable and single output variable such an algorithm is called as Simple linear Regression.

$$y = mx + c$$

m = slope
c = y - intercept

$$m = \frac{dy}{dx}$$



② **Multi linear Regression:** If an algorithm uses multiple input variables and single output variables, such an algorithm is called as multi Linear Regression.

$$y = m_1x_1 + m_2x_2 + m_3x_3 + c$$

Assumption of linear Regression:

① **Linearity:** There should be a linear relationship b/w dependent & independent variable.

② Data should be normal.

③ Little or no multicollinearity [There should be very less correlation among input variables.

How linear Regression works?

x	y
No of hours studied	marks obtained
2	20
3	30
4	45
7	75
8	85
10	95
5	?

Linear Regression

straight line Eqn

$y = mx + c$

Find: $m = ?$

$c = ?$

$$m = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\Sigma (x - \bar{x})^2}$$

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

$\bar{x}$ & $\bar{y}$ = mean.

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
				Σ	Σ

after finding $m = 8$, $c = 5$

$$\hat{y} = 8x + 5$$

→ Regression model / Best fit line / Regression line.

Find for 5 hours,

$\hat{y} = 8 \times 5 + 5$

$\hat{y} = 45$

① Error / loss = actual value − Predicted value.

x	y	Predicted $\hat{y}$	Error
2	20	21	1
3	30	28	1
4	45	37	8
7	75	61	14
8	85	69	16
10	95	85	10
5	?		

Calculation Total Error:-

(1) MAE → mean absolute Error

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

y_i = actual, $\hat{y}_i$ = Predicted.

(2) MSE → mean Squared Error

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

MSE best because of it square no ture.

(3) RMSE:- Root mean squared Error

$$RMSE = \sqrt{MSE}$$

IMP

Note:-

→ If Error is less, we say its good model

→ If Error is high, we say its bad model.

Data preprocessing :-

- checks for missing values and handle them
- checks for duplicates
- checks for outliers and handle them

How to handle outliers :-

IQR

- This method is used to find outliers when data is skewed / not normal

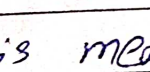
- Lower limit = $Q_1 - 1.5 * IQR$
- Upper limit = $Q_3 + 1.5 * IQR$
- $x <$ Lower and $x >$ upper then x is outlier

$$x \notin [Q_1 - 1.5 * IQR, \; Q_3 + 1.5 * IQR]$$

- If we are using IQR replace with median

3 - Sigma Rule

- This method is used to identify outliers when data is normal



- Lower limit = $\mu - 3\sigma$
- upper limit = $\mu + 3\sigma$

$\mu \Rightarrow$ mean

$\sigma \Rightarrow$ standard deviation

- $x <$ Lower and $x >$ upper then x is outlier

$$x \notin [\mu - 3\sigma, \; \mu + 3\sigma]$$

- If we are using 3-sigma replace with mean.

IQR:

1. Find Q_1 and Q_3
2. $IQR = Q_3 - Q_1$
3. Find lower limit and upper limit

lower limit $= Q_1 - 1.5 \times IQR$, upper $= Q_3 + 1.5 \times IQR$

(VVI)

⇒ handle only outlier is less than 5%. other wise don't touch it.

Scaling:-

Scaling is one of the preprocessing method which is applied on continuous data to bring all the values into a certain scale.

- scaling is applied only on input variables
- There are two types

(1) min max scaler	(2) standard scaler
• It is called as normalization	• It is called as standerdization
• $Minmax = \frac{x - x_{min}}{x_{max} - x_{min}}$	• standardization $z = \frac{x - \mu}{\sigma}$
• range of minmax $= [0, 1]$ It will reduce all the values to lie b/w $[0, 1]$	• range $= [-3, 3]$

Sklearn :-

Sklearn is one of the scientific library which includes all the fn / pckages related to machine learning, data science, preprocessing etc

→ Feature Engineering :-

- Drop irrelevent column
- create New columns
- select best features with target

  → Find corr
  → plot heat map

- Include the colums which has high correlation with target
- drop columns which has less correlation (<30%) with target
- Drop

→ No @ little Multi collinearity

- make sure the correlation b/w input variables should be less
- If correlation is high among input variables drop one of them.

random_state = 4, (shuffle data)

⇒ Split data for training and testing

70 % train , 30 % (or) test

80 % , 20 %

x = x-train, x-test

y = y-train, y-test

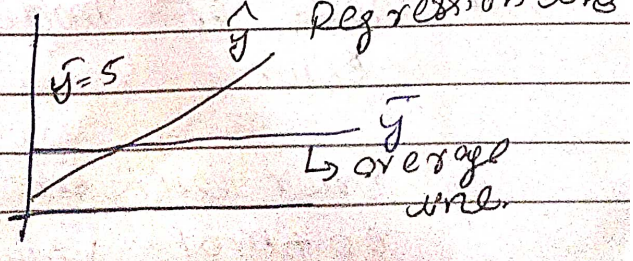
⇒ Training (x-train, y-train)

⇒ testing / prediction ⇒ x-test.

(1) ⇒ R-squared / R^2 score

- It is used to measure the stregth of the model
- It tells us about how good is the model.
- $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y}_i)^2}$
- RSS ⇒ Residual sum of Square.
- TSS ⇒ Total sum of squares
- R^2 score compares regression line with average

R^2 score = [-1, 1]



(5) Adjusted R^2-score:

Disadvantage of R^2:-

R^2 score increases as the number of independent variables increases which has very less relationship with target variable.

To over come above in issue we use adjusted R^2

$$adj = 1 - \frac{(1-R^2)(N-1)}{(N-P-1)}$$

$N \Rightarrow$ No of observation in test data

$R^2 \Rightarrow$ R^2-score

$P \Rightarrow$ No of independent variables

(VVIMP)

if Adjusted R^2-score $<$ R^2-score.
Then the model is good.

→ Cost function:

$$J = \frac{1}{2n} \Sigma (y_i - \hat{y}_i)^2$$

- Goal of model is to reduce cost function error.
- Error can be minimised by choosing optimal values from m & c.

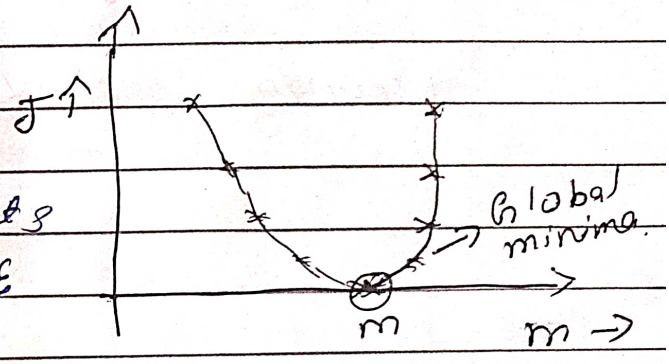
- How to choose optimal values for m & c?

Gradient descent:-

Gradient descent is one of the optimization method which helps in selecting optimal values for m & c such that error is minimum.

How gradient desent works?

- plot graph of error v/s slope



- Gradient descent starts with random slope & works iteratively to reach global minima where error is minimum.

- m value at the global minima will be the optimal value.

$$m_{new} = m_{old} - \eta \frac{\partial J}{\partial m}$$

$$c_{new} = c_{old} - \eta \frac{\partial J}{\partial c}$$

$\eta \rightarrow$ Learning rate (or) α.

η - Learning rate:- It tells us about how many steps it has taken to reach Global minima.

If η is large :- It overshoots and never reaches global minima.

η - is small :- It takes more time to reach global minima.

How to choose η?

Initially take larger steps and then take smaller steps when near to global minima.