Unit-4.

18/09/2023.

Supervised Learning : Regression

Introduction, Example of Regression, Common Regression Algorithms - Simple linear Regression, Multiple linear Regression, Assumptions in Regression Analysis, Main problems in Regression Analysis, Improving Accuracy of the Linear Regression Model, Polynomial Regression Model, Logistic Regression, Maximum Likelihood Estimation.

1. Regression Introduction:-

In this, we build concepts ~~of~~ on prediction of numerical variables, which is another key area of supervised learning. This area is known as "Regression", focuses on solving problems such as predicting value of real estate, demand forecast in retail, weather forecast etc.

Regression is a statistical analysis technique used to model the relationship used to model the relationship b/w a dependent variable (often denoted as "y") and one (or more independent variables (often denoted as "X").

It is commonly used for prediction and understanding the association b/w variables.

Regression analysis helps us to understand how the value of the dependent (target) variable is changing

corresponding to an independent variable.(predictor)
It predicts continuous / real values such as temperature, age, salary, price etc.

Regression is a supervised learning technique which helps in finding the correlation b/w variables and enable us to predict the output variable based on one or more predictor variables.

"Regression shows a line or curve that passes through all the data points on target - predictor graph in such a way that the vertical distance b/w the datapoints and the regression line is min".

Terminology:

Dependent Variable:

The main factor in regression analysis which we want to predict or understand is called the dependent variable. It is also called target variable.

Independent Variable:

The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variables, also called as a predictor.

Outliers:

Outlier is an observation which contains either very low value or very high value in comparison to other observed values.

Multicollinearity:

If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity.

Underfitting and Overfitting:

If our algorithm works well with the training dataset but not well with the test dataset, then such problem is called "Overfitting".

If our algorithm does not perform well even with training dataset then such problem is called "underfitting".

Types of Regression:

There are various types of regression which are used in data science and ML.

→ Linear Regression
→ Logistic Regression
→ Polynomial Regression
→ Support Vector Regression
→ Decision Tree Regression
→ Random Forest Regression
→ Ridge Regression
→ Lasso Regression

Regression can be extended to handle more complex relationships with techniques like multiple regression, polynomial regression, etc.

2. Simple Linear Regression :-

Simple linear Regression is a type of Regression algorithm that models the relationship b/w a dependent variable and a single independent variable.

The relationship shown by a Simple linear regression model is a linear or sloped SL, hence it is called simple Linear Regression.

The key point in Simple Linear Regression is that the dependent variable must be a continuous/ real value. However, the independent variable can be measured on continuous or categorical value.

Simple Linear Regression algorithm has mainly 2 objectives,

→ Model the relationship b/w the two variables.

→ Forecasting new observation.

The linear regression model can be represented using the below equation,

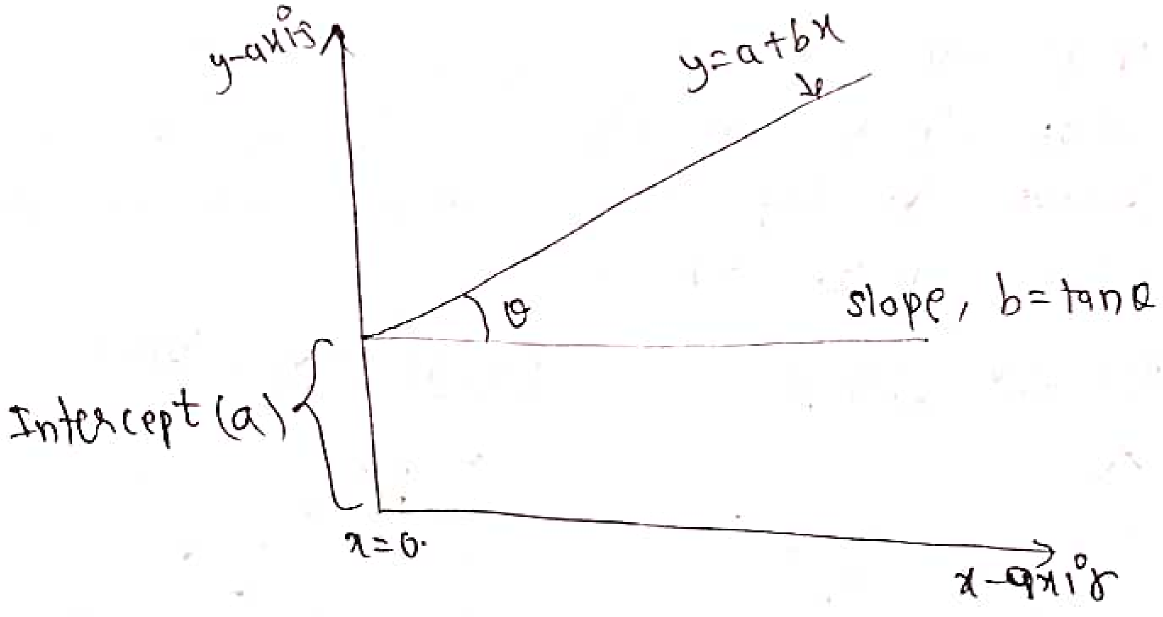
$$y = a_0 + a_1 x + \varepsilon$$

where, a_0 = Intercept of the regression line.

a_1 = solope of the regression line.

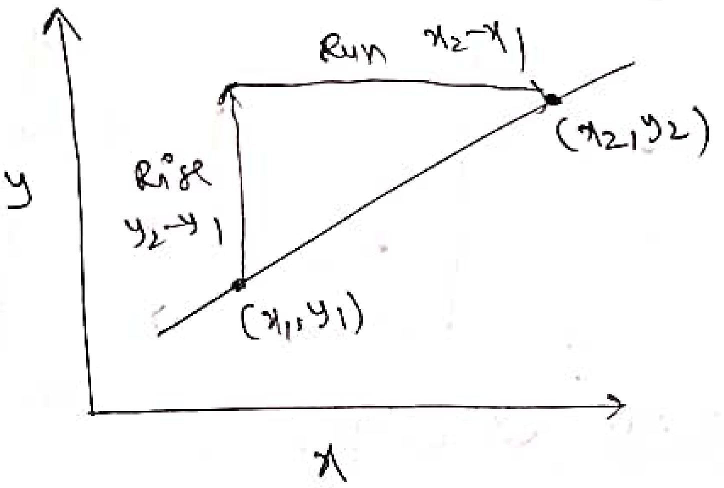
ε = The error term.

$$\overset{\text{Dependent variable}}{y} = \underset{\text{Intercept}}{a} + \underset{\text{Slope}}{b}\,\overset{\text{Independent variable}}{x}$$



The value of intercept indicates the value of 'y' when x = 0. It is known as "y-intercept".

$$\text{Slope} = \frac{\text{change in } y}{\text{change in } x} = \frac{\Delta y}{\Delta x} = \frac{y_2-y_1}{x_2-x_1}$$



$$\text{Slope} = \frac{\text{Rise}}{\text{Run}} = \frac{y_2-y_1}{x_2-x_1}$$

There can be 2 types of slopes in a linear regression model:

→ +ve slope

→ -ve slope.

Different types of regression lines based on the type of slope include:

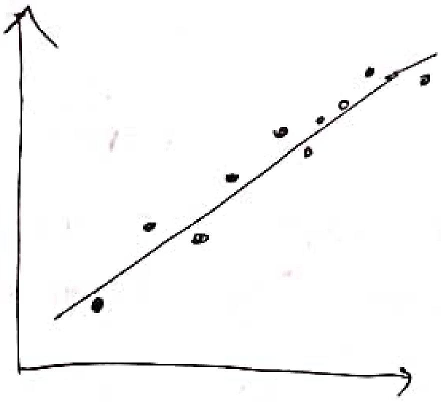
→ Linear +ve slope

→ Curve Linear +ve slope

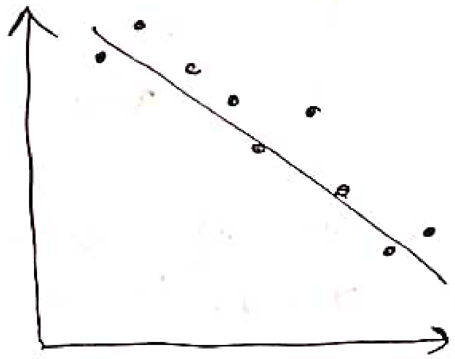
→ Linear -ve slope

→ Curve Linear -ve slope.

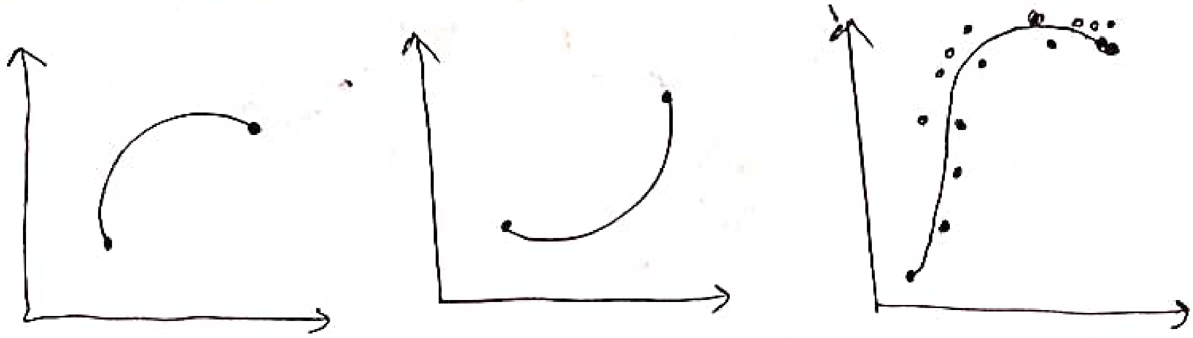
Linear +ve slope:



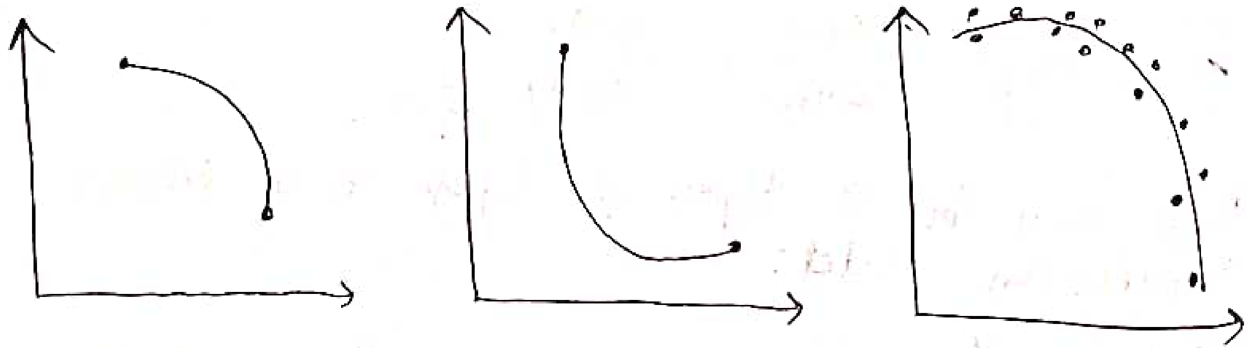
Linear -ve slope:



Curve linear +ve slope:

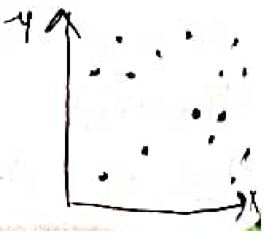


Curve linear -ve slope:



No relationship Graph:

It is very difficult to conclude whether the relationship b/w x and y is +ve (or) -ve.



errors in Simple Regression:

There will be some error value (ε) associated with it. This error is called marginal (or) residual error.

$$y = (a + bx) + \varepsilon .$$

The main goal is to find the best-fitting line (a and b values) that minimizes the difference b/w the predicted y values($\hat{y}$)and the actual y values in our dataset

simple linear regression is a fundamental technique in statistics and data analysis and serves as the basis for more complex regression methods like multiple regression.

It is commonly used in various fields for tasks like predicting sales based on advertising spending, estimating the impact of variables on outcomes.

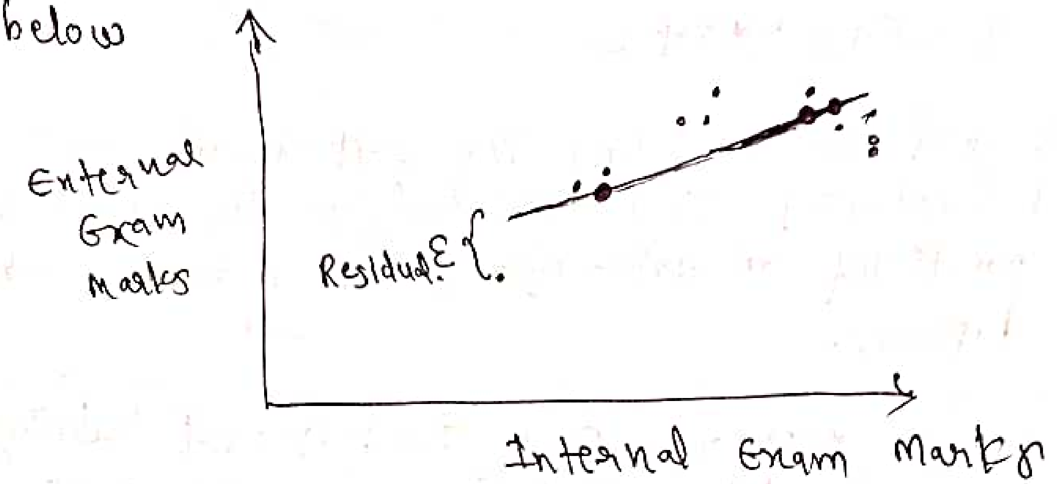
Example of Simple Regression:-

A college professor believes that if the grade for internal examination is high in a class, the grade for external examination will also be high.

A random sample of 15 students in that class was selected, and the data is given below.

Internal Exam	15	23	18	23	24	22	22	19	19	16	24	11	24
External Exam	49	63	58	60	58	69	60	63	60	52	62	30	59

A scatter plot was drawn to explore the relationship b/w the independent variable (internal marks) mapped to X-axis and dependent variable (external marks) mapped to Y-axis as shown below



Residual: Residual (ε) is the distance b/w the predicted point (on the regression line) and the actual point as depicted in above figure.

Ordinary Least Squares (OLS) is the technique used to estimate a line that will minimize the error (ε), which is the difference b/w the predicted and the actual values.

The Sum of the Squares of the Errors

$$(SSE) = \sum_i \varepsilon_i^2 .$$

$$b = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)} = \frac{\text{Sum of products}}{\text{Sum of squares}}$$

$$a = \bar{y} - b\bar{x} .$$

Sum of $x = 299$

Sum of $y = 852$

Mean $(\bar{x}) = 19.93$

Mean $(\bar{y}) = 56.8$

Sum of Squares = 226.9333

Sum of products = 429.8

$$y = bx + a$$

$$b = \frac{429.8}{226.9333} = 1.89395$$

$$a = \bar{y} - b\bar{x} = 56.8 - 1.89(19.93)$$

$$a = 19.0473$$

$$\therefore \hat{y} = 1.89395x + 19.0473$$

Marks in External Exam = 19.04 + 1.89 x (Marks in Internal Exam)

OLS Algorithms:

(i) Calculate the $\bar{x}$ and $\bar{y}$
(ii) Calculate the errors of x and y
(iii) Get the product
(iv) Get the summation of the products
(v) Square the difference of x
(vi) Get The sum of the squared difference.
(vii) Calculate slope (b)
(viii) Calculate 'a' using value of (b) (Intercept)

Maximum and Minimum point of curves :

Max and min points on a graph are found at points where the slope of the curve is zero.

The max point is the point on the curve of the graph with the highest y-coordinate and a slope of zero.

The min point is the point on the curve of the graph with the lowest y-coordinate and slope of zero.

---

4. Multiple Linear Regression :

Multiple Linear Regression is an extension of Simple Linear Regression, where it models the relationship b/w a dependent variable (Y) and multiple independent variables (X) by assuming a linear relationship.

The multiple linear regression model is represented as

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

where, y = dependent variable

$x_1, x_2, \ldots x_n$ = Independent variables

b_0 = Intercept

$b_1, b_2, b_3, \ldots, b_n$ = Coefficients of respective x variables.

The main goal is to find the best fitting values of the coefficients $(b_0, b_1, b_2, \rightarrow b_n)$ that minimize the difference b/w the predicted Y values $(\hat{Y})$ and the actual 'y' values in your dataset.

The following expression describes the equation involving the relationship with a predictor variables.

$$\hat{y} = a + b_1x_1 + b_2x_2.$$

The model describes a plane in the 3-D space of $y, x_1,$ and x_2.

where, a = Intercept

b_1, b_2 = Partial Regression Coefficients.

In ~~simple~~ linear regression, where a single independent/predictor (x) variable is used to model the response variable (y).

For MLR, the dependent (or) target variable (y) must be the continuous/real, but the predictor (or) independent variable may be of continuous (or) categorical form.

5. Assumptions in Regression Analysis:

(i) The dependent variable (y) can be calculated/predicted as a linear function of a specific set of independent variables (x) plus an error term (ε).

(ii) The number of observations (n) is greater than the number of parameters (k) to be estimated. i.e, n>k.

(iii) Relationships determined by regression are only relationships of association based on the data set.

(iv) Regression line can be valid only over a limited range of data.

(v) Variance is the same for all values of 'X'. (Homoskedasticity).

(vi) The error term (ε) is normally distributed.

(vii) The values of the error (ε) are independent and are not related to any values of 'X'.

Given the above assumptions, the OLS estimator is the BLUE - Best Linear Unbiased Estimator, and this is called as Gauss-Markov Theorem.

The theory of linear regression is based on certain statistical assumptions.
It is crucial to check these regression assumptions before modeling the data using the linear regression approach.
Mainly there are 7 assumptions taken,

→ Linear Model

→ No multicollinearity in the data

→ Homoscedasticity of residuals or Equal variance

→ No autocorrelation in residuals

→ Number of observations greater than the number of predictors.

→ Each observation is unique.

→ Predictors are distributed normally.

Main Problems in Regression Analysis:-

In multiple regressions, There are 2 primary problems : multicollinearity and heteroskedasticity.

multicollinearity:

Two variables are perfectly collinear if there is an exact linear relationship b/w them.

Multicollinearity is the situation in which the degree of correlation is not only b/w the dependent variable and the independent variable, but there is also a strong correlation within (or) among the independent variables themselves.

One way to gauge multicollinearity is to calculate the Variance Inflation Factor (VIF), which assess how much the variance of an estimated regression coefficient increases if the predictors are correlated.

If no factors are correlated, the VIF will be 1.

Heteroskedasticity:

Heteroskedasticity refers to the changing variance of the error term.

If the variance of the error term is not constant across data sets, there will be erroneous predictions.

In general, for a regression equation to make accurate predictions, the error term should be independent, identically distributed.

Mathematically, it is represented as

$$Var(y_i|X) = \sigma^2 \text{ and}$$

$$cov(y_i y_j|X) = 0 \text{ for } i \neq j.$$

Here are some of the main issues may encount in regression analysis:

Assumption violations:

Linear regression assumes a linear relationship b/w the independent and dependent variables. Vidations of this assumption can lead to inaccurate results.

Overfitting:

Adding too many independent variables to a regression model, especially when they are not truly related to the dependent variable, can lead to overfitting.

This results in a model that fits ~~that~~ The training data well but performs poorly on new, unseen data.

Underfitting:

On the opposite end, an overly simplistic model may underfit the data and fail to capture important relationships b/w variables.

Outliers:

Outliers are the extreme data points that don't follow the general trend, can disproportionately influence regression results.

Endogeneity:
This problem arises when an independent variable is correlated with the error term, often due to reverse causality.

Addressing these problems often involves careful data preprocessing, model selection, diagnostic testing, and in some cases, using alternative regression techniques, such as robust regression, ridge regression, or non-linear regression, to better fit the data and mitigate these issues.

Improving accuracy of the linear regression model:-

The concept of bias and variance is similar to accuracy and prediction.
Accuracy refers to how close the estimation is near the actual value, where as prediction refers to continuous estimation of the value.

High bias = low accuracy
High variance = low prediction

Low bias = high accuracy
Low variance = high prediction

If the variance increases (low prediction), the spread of our data points increases, which results in less accurate prediction.

As the bias increases (low accuracy), the error b/w our predicted value and the observed values increases.

The accuracy of linear regression can be improved using the following 3 methods.

→ Shrinkage Approach

→ Subset selection

→ Dimensionality (Variable) Reduction.

Shrinkage (Regularization) Approach:-

By limiting (shrinking) the estimated coefficients, we can try to reduce the variance at the cost of a negligible increase in bias.

Few variables used in the multiple regression model are in fact not associated with the overall response and are called as irrelevant variables.

However, the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinking has the effect of reducing the overall variance.

The 2 best-known techniques for shrinking the regression coefficients towards zero are

→ Ridge Regression

→ Lasso (Least Absolute shrinkage selector Operator)

Ridge regression performs L2 regularization. i.e, it adds penalty equivalent to square of the magnitude of coefficients.

Lasso regression performs L1 regularization. i.e, it adds penalty equivalent to the absolute value of the magnitude of coefficients.

Subset Selection:

Identify a subset of the predictors that is assumed to be related to the response and then fit a model using OLS on the selected reduced subset of variables.

There are 2 methods in which subset of the regression can be selected:

→ Best Subset selection. (2^k)

→ Stepwise subset selection

(i) Forward stepwise selection (0 to k)

(ii) Backward stepwise selection (k to 0).

In best subset selection, we fit a seperate least squares regression for each possible subset of the k predictors. It considers 2k possible models containing subsets of the 'p' predictors.

Forward stepwise selection begins with a model containing no predictors, and then predictors are added one by one to the model until all the 'k' predictors ~~are added one by one to the model~~. are included in the model.

Backward stepwise selection begins with the least squares model which contains all 'k' predictors and then iteratively removes the least useful predictor one by one.

Dimensionality Reduction:

In dimensionality reduction, predictors (X) are transformed and the model is set up using the transformed variables after dimensionality reduction.

The number of variables are reduced using the dimensionality reduction method.

Ex:- PCA - Principal Component Analysis.

8. Polynomial Regression Model:-

Polynomial regression model is the extension of the simple linear model by adding extra predictors to a power.

$$f(x) = c_0 + c_1 x^1 + c_2 x^2 + c_3 x^3$$

where, c_0, c_1, c_2 & c_3 are the coefficients.

Polynomial regression is a type of regression analysis used when the relationship b/w the independent variable (X) and the dependent variable (Y) is not linear but can be better approximated by a polynomial function.

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + --- + a_n x^n.$$

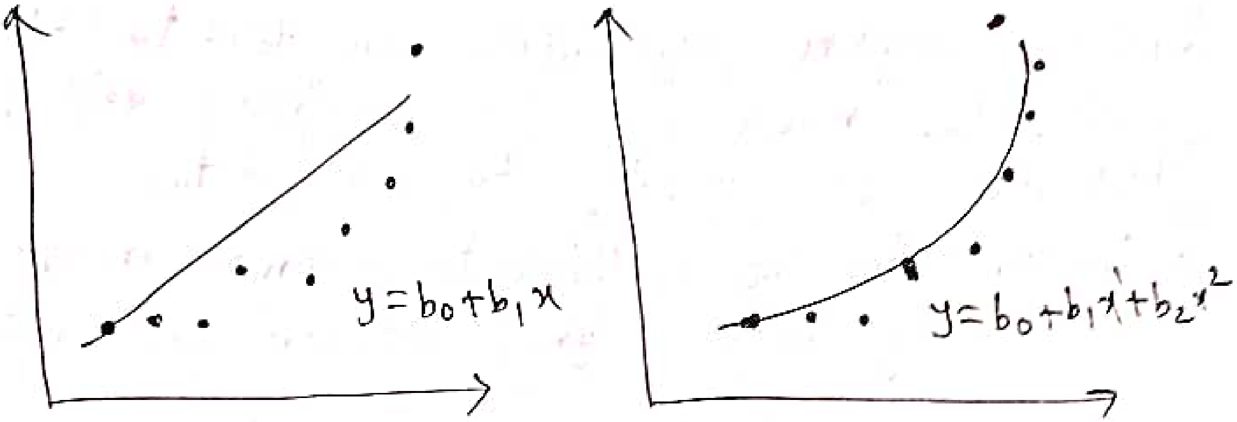
Here, n represents the degree of the polynomial, which determines the complexity of the curve.

It is also called the special case of multiple linear regression in ML.

It is a linear model with some modification in order to increase the accuracy.

"In polynomial regression, the original features are converted into polynomial features of required degree (2, 3, ---, n) and then modeled using a linear model".

Where data points are arranged in a non-linear fashion, we need the polynomial regression model.



Simple linear model

Polynomial model

A polynomial regression algorithm is also called polynomial linear regression, because it does not depend on the variables, instead, it depends on the coefficients, which are arranged in a linear fashion.

The simple and multiple linear equations are also polynomial equations with a single degree, and the polynomial regression equation is linear equation with the nth degree.

The choice of the degree 'n' is crucial. Higher degree polynomials can fit the data more closely but are prone to overfitting.

Polynomial regression that can better capture non-linear relationships b/w variables.

Advantages:

→ Polynomial regression can model complex relationship that linear regression can not.

→ It is flexible and adaptable to various data shapes.

Disadvantages:

→ Higher degree polynomials can lead to overfitting, where the model fits the training data perfectly but fails to generalize to new data.

→ Interpreting the coefficients becomes more challenging with higher degree polynomials.

To combat overfitting, you can apply regularization techniques like Ridge (or) Lasso regression.

Use cross-validation to select the degree of the polynomial and asses model performance

Careful model selection and evaluation are essential to get the most accurate and reliable results

4. Logistic Regression :-

Logistic regression is both classification and regression technique depending on the scenario used.

Logistic regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable similar to OLS regression.

In logistic regression, dependent variable (Y) is binary (0,1) and independent variable (X) are continuous in nature.

In the logistic regression model, there is no R^2 to gauge the fit of the overall model, however, a chi-square test is used to gauge how well the logistic regression model fits the data.

The goal of logistic regression is to predict the likelihood that Y is equal to 1 for given certain values of 'X'.

Logistic regression is a supervised ML algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class or not.

It is referred to as regression because it takes the output of the linear.

It uses a sigmoid function to estimate the probability for the given class.

The difference b/w linear regression and logistic regression is that linear regression o/p is the continuous value that can be anything while logistic regression predicts the probability that an instance belongs to a given class or not.

Logistic Function (Sigmoid Function):

The sigmoid function is a mathematical function used to map the predicted values to probabilities.

It maps any real value into another value within a range of '0' and '1'., So it form a curve like the "S" form.

The S-form curve is called the sigmoid function.

In logistic regression, we use the concept of the threshold value, such as values above the threshold value tends to 1, and i value below the threshold values tends to '0'.

Types of Logistic Regression:

On the basis of the categories, logistic regression can be classified into three types.

(i) Binomial: In this, there can be only two possible types of the dependent variables such as 0 (or) 1.

(ii) Multinomial: In this, there can be 3 (or) more possible unordered types of the dependent variable. such as cat, dog, sheep.

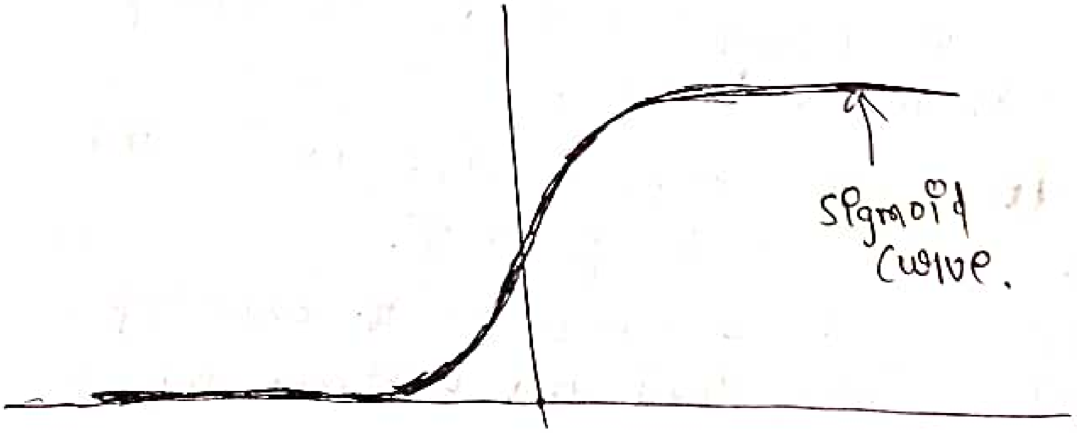
(iii) Ordinal: In this, there can be 3 (or) more possible ordered types of dependent variables such as, low, medium, high,

Odds: It is the ratio of something occurring to something not occurring.

Log-odds: It is also known as logit function, is the natural logarithm of the odds.

Maximum likelihood Estimation:

This method used to estimate the coefficients of the logistic regression model, which maximizes the likelihood of observing the data given the model.



Sigmoid function

Assumptions for Logistic Regression:

→ Independent observation
→ Binary dependent variables.
→ Linearity relationship b/w independent variables and log odds.
→ There exists a linear relationship b/w logit function and indepent variables.
→ No outliers
→ Large sample size

→ The error term follows binomial distribution [n,p]
p = probability of success.
n = number of records in the data.

→ The error terms (E), are independent from one another and identically distributed.
→ The dependent variable (Y) must be categorical (1/0) and take binary value.

10. Maximum Likelihood Estimation :-

The coefficients in a logistic regression are estimated using a process called maximum likelihood Estimation (MLE).

First let us understand what is likelihood function before moving to MLE.

A fair coin outcome flips equally heads and tails of the same number of times.

If we toss the coin 10 times, it is expected that we get 5 times Head and 5 times tail.

The probability (P) is > 0.5, it is said to be in favour of Head.

The probability (P) is < 0.5, it is said to be in favour of Tail.

Let us represent 'n' flips of coin as $X_1, X_2, \rightarrow X_n$.

$X_i = 1$ if Head

$X_i = 0$ if Tail

Bernoulli distribution represents each flip of the coin.

$$f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

iid = independent and identically distributed

The likelihood equation is

$$L(\theta|x) = \sum_{i=1}^{n} f(x_i|\theta)$$

But the likelihood function is not a probability. The likelihood for some coins may be 0.25 (or) 0 (or) 1.

MLE is about predicting the value for the parameters that maximizes the likelihood function

$$\log L(\theta|x) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

Maximum likelihood is an approach commonly used for such density estimation problems., in which a likelihood function is defined to get the probabilities of the distributed data.

The concept of maximum likelihood ~~est~~ as it is one of the primary and core concepts essential for learning other advanced ML and DL techniques and algorithms.

Likelihood: The likelihood is a function that tells us how likely the specific data point suits the existing data distribution.

Difference b/w Probability and Likelihood:

Likelihood is a function that defines (or) tells us how accurate the particular data point is valuable and contributes to the final algorithm in data distribution.

where as, probability that describes the chance of some event (or) thing happenning concerning other circumstances (or) conditions, mostly known as conditional probability

It is clear to us that a higher likelihood is desired for every model to get an accurate model and has accurate results.

So, here, the term maximum likelihood represents that we are maximizing the likelihood function, called the maximization of the likelihood function.

The max likelihood estimation is a base of some ML and DL approaches used for classification problems

Ex: Logistic regression, where the algorithm is used to classify the data points using the best-fit on the graph.

The same approach is known as the perceptron trick regarding DL algorithms.

Applications:

→ MLE is used in various statistical models, including linear regression, logistic regression, exponential distribution, poisson distribution, and many other.

→ It is also a fundamental concept in maximum likelihood based ML algorithms like MLE for probabilistic models.

→ MLE allows for the estimation of confidence intervals for the parameter estimates and hypothesis testing about the parameters values.