REGRESSION

Regression Analysis in Machine learning

- Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables.
- More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed.
- It predicts continuous/real values such as temperature, age, salary, price, etc.
- We can understand the concept of regression analysis using the below example:
- Example: Suppose there is a marketing company A, who does various advertisement every year and get sales on that. The below list shows the advertisement made by the company in the last 5 years and the corresponding sales

Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

- Now, the company wants to do the advertisement of $200 in the year 2023 and wants to know the prediction about the sales for this year.
- So to solve such type of prediction problems in machine learning, we need regression analysis.
- Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables.
- It is mainly used for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.

- In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data.
- In simple words, "Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum
- ." The distance between datapoints and line tells whether a model has captured a strong relationship or not.
- Some examples of regression can be as:
- o Prediction of rain using temperature and other factors
- o Determining Market trends
- o Prediction of road accidents due to rash driving.

Terminologies Related to the Regression Analysis:

- 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.
- o Independent Variable: The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a predictor.
- o Outliers: Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.
- Multicollinearity: If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity. It should not be present in the dataset, because it creates problem while ranking the most affecting variable.
- o Underfitting and Overfitting: If our algorithm works well with the training dataset but not well with test dataset, then such problem is called Overfitting. And if our algorithm does not perform well even with training dataset, then such problem is called underfitting.

Why do we use Regression Analysis?

- As mentioned above, Regression analysis helps in the prediction of a continuous variable. There are various scenarios in the real world where we need some future predictions such as weather condition, sales prediction, marketing trends, etc., for such case we need some technology which can make predictions more accurately.
- So for such case we need Regression analysis which is a statistical method and used in machine learning and data science.
- Below are some other reasons for using Regression
- analysis:
- o Regression estimates the relationship between the target and the independent variable.
- o It is used to find the trends in data.
- o It helps to predict real/continuous values.

- o By performing the regression, we can confidently determine the most important factor, the least important factor, and how each factor is affecting the other factors.

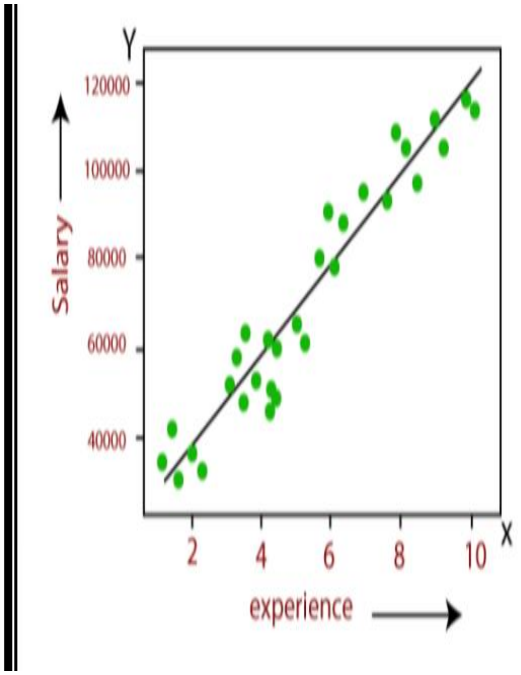
Types of Regression

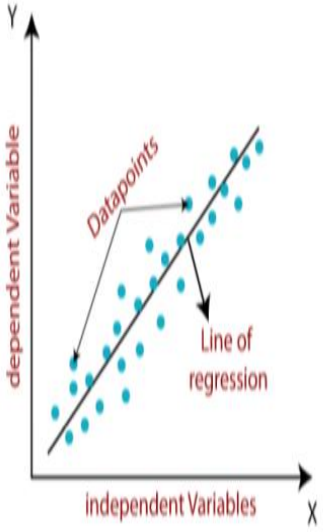
- There are various types of regressions which are used in data science and machine learning. Each type has its own importance on different scenarios, but at the core, all the regression methods analyze the effect of the independent variable on dependent variables.
- Here we are discussing some important types of regression which are given below:
- o Linear Regression
- o Logistic Regression
- o Polynomial Regression
- o Support Vector Regression
- o Decision Tree Regression
- o Random Forest Regression
- o Ridge Regression
- o Lasso Regression

LINEAR REGRESSION IN MACHINE LEARNING

Linear Regression:

- Linear regression is a statistical regression method which is used for predictive analysis. It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
- It is used for solving the regression problem in machine learning.
- In the simplest words, Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable.
- If there is only one input variable (x), then such linear regression is called simple linear regression. And if there is more than one input variable, then such linear regression is called multiple linear regression.
- Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc
- .The relationship between variables in the linear regression model can be explained using the below image.
- Here we are predicting the salary of an employee on the basis of the year of experience.

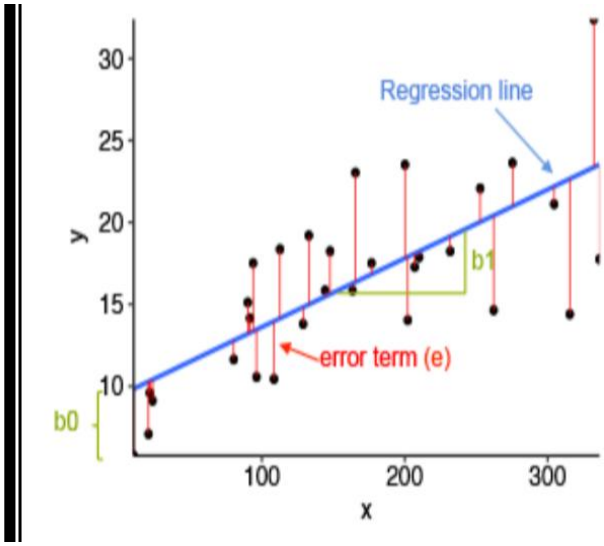




- Below is the mathematical equation for Linear regression:

  Y= aX+b

- Here,
- Y = dependent variables (target variables),
- X= Independent variables (predictor variables),
- a and b are the linear coefficients
- A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.
- Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.



- In the above diagram,
- x is our dependent variable which is plotted on the x-axis and y is the dependent variable which is plotted on the y-axis.
- Black dots are the data points i.e the actual values.
- bo is the intercept which is 10 and b1 is the slope of the x variable.
- The blue line is the best fit line predicted by the model i.e the predicted values lie on the blue line.
- The vertical distance between the data point and the regression line is known as error or residual. Each data point has one residual and the sum of all the differences is known as the Sum of Residuals/Errors.

Mathematical Approach:

- Residual/Error = Actual values – Predicted Values
- Sum of Residuals/Errors = Sum(Actual- Predicted Values)
- Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))2
- i.e

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

S

- Some popular applications of linear regression are:
- o Analyzing trends and sales estimates
- o Salary forecasting
- o Real estate prediction
- o Arriving at ETAs in traffic.

Types of Linear Regression

- Linear regression can be further divided into two types of the algorithm:
- o **Simple Linear Regression**: If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- o **Multiple Linear regression**: If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression

- **Equation of Simple Linear Regression**, where bo is the intercept, b1 is coefficient or slope, x is the independent variable and y is the dependent variable.

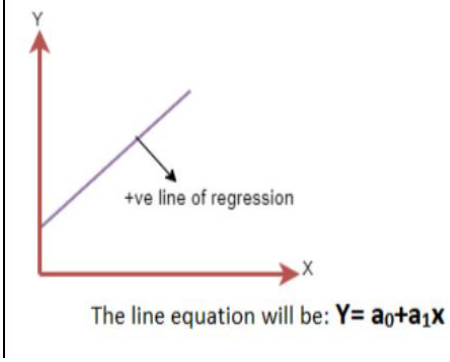
$$y = b_o + b_1 x$$

- **Equation of Multiple Linear Regression**, where bo is the intercept, b1,b2,b3,b4…,bn are coefficients or slopes of the independent variables x1,x2,x3,x4…,xn and y is the dependent variable.

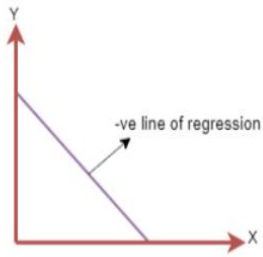
$$y = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots . + b_n x_n$$

Linear Regression Line

- A linear line showing the relationship between the dependent and independent variables is called a regression line.
- A regression line can show two types of relationship:
- o Positive Linear Relationship:
- If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.
- o Negative Linear Relationship:
- If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



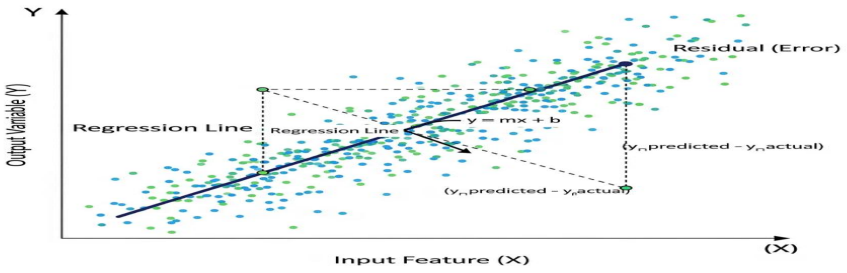
The line equation will be: $Y = a_0 + a_1 x$



The line of equation will be: $Y = -a_0 + a_1 x$

Finding the best fit line

- When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.
- Simple Linear Regression: -
- Simple Linear Regression is a type of Regression algorithms that models the relationship between a dependent variable and a single independent variable. The relationship shown by a Simple Linear Regression model is linear or a sloped straight line, 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 values.
- Simple Linear regression algorithm has mainly two objectives:
- o Model the relationship between the two variables. Such as the relationship between Income and expenditure, experience and Salary, etc.
- o Forecasting new observations. Such as Weather forecasting according to temperature, Revenue of a company according to the investments in a year, etc.
- Linear regression is one of the most fundamental yet essential models in machine learning. It is used to model the relationship between input variables (independent variables) and an output variable (dependent variable) through a straight line



1. What Is a Linear Model?

A linear model represents the relationship between input and output in the form of a straight-line equation:

$$y=wx+b$$

- x = input (e.g., marketing cost)
- y = output or target (e.g., sales)
- w = slope/regression coefficient
- b = intercept
- If the data forms a straight line, the relationship is called linear. If it curves, the relationship is non-linear.

- Example: Predicting monthly sales based on marketing budget. With historical data, we can build a linear model to estimate sales at different budget levels.

2. Error in Linear Regression

- A linear model will never be perfectly accurate. There will always be an error (difference) between predicted and actual values:
- e=y-(wx+b)
- If errors are simply added up, positive and negative errors may cancel each other out. To solve this, errors are usually squared, leading to the Residual Sum of Squares (RSS):

$$RSS = \sum (y - (wx + b))^2$$

- RSS measures how far the predictions deviate from the actual data. A good model has the smallest possible RSS.

3.Finding the Best Parameters (w and b)

- Since x (input) and y (output) are fixed, the only values we can adjust are the parameters w and b. The goal of linear regression is to find the combination of w and b that minimizes RSS.

Two common approaches:

1.Ordinary Least Squares (OLS):

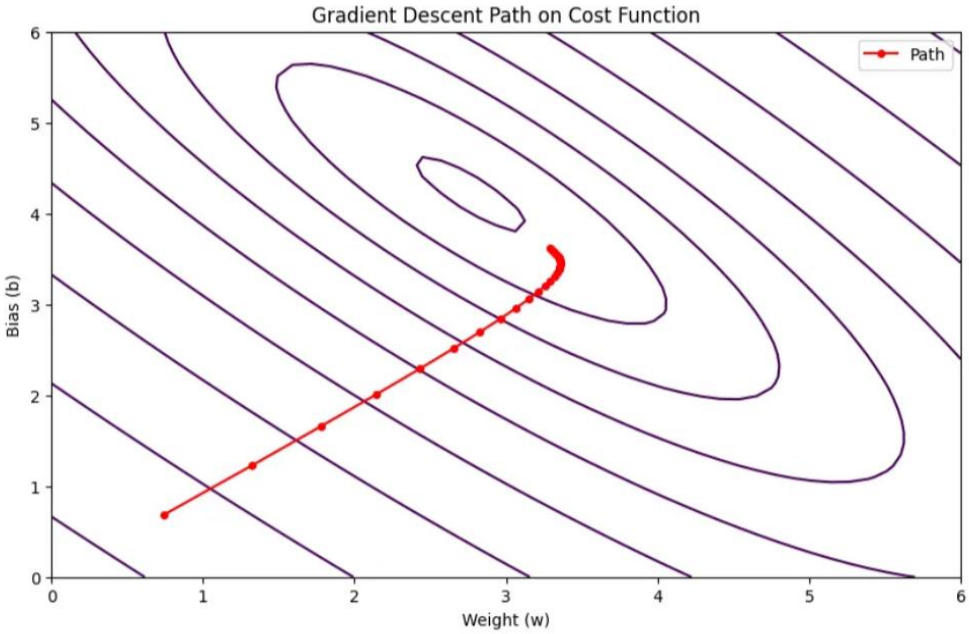
- Computes a direct solution using calculus.
- Takes the derivative of RSS with respect to the parameters and sets it to zero to find the minimum.
- Efficient for small to medium datasets.

2. Gradient Descent:

- An iterative optimization method.
- Starts from an initial guess, then updates parameters step by step in the direction that reduces error.
- Suitable for large datasets where OLS can be computationally expensive (due to large matrix operations).

4. Intuition Behind Gradient Descent

- Imagine walking down a mountain:
- If the slope (gradient) is positive → move left.
- If the slope is negative → move right.
- If the gradient is zero → you've reached the minimum point.
- This process repeats until the model finds the optimal parameters that minimize the error.



-

5. <u>Model Interpretation</u>

- Once the best parameters are found:
- Intercept (b): the predicted value when input = 0.
- Slope (w): the change in output for every 1-unit increase in input.

Example regression equation:

- Interpretation:
- When the budget = 0, sales are predicted to be 5.
- For every additional 1 unit in marketing budget, sales increase by 1.2.

$$Sales = 5 + 1.2 \times (Marketing\ Budget)$$