Machine Learning – Webinar 2

BITS Pilani

Pilani Campus

innovate achieve lead

BITS Pilani

Pilani Campus

Case Study – Regression Analysis

Overview

In today's session, we will cover the below topics,

- Recap on What is Regression Analysis, Why we need Regression Analysis and Types of Regression Analysis
- Case study on Linear Regression
- Case study on Multiple Linear Regression
- Case study on Polynomial Regression

Regression Analysis

- It 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.
- 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.

Regression Analysis - Example

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

- 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:

- Now, the company wants to do the advertisement of $200 in the year 2019 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.

Terminologies Related to the Regression Analysis

- Dependent Variable (target variable)
- Independent Variable (predictor)
- Outliers
- Multicollinearity
- Underfitting and Overfitting

Why do we use Regression Analysis?

- Regression estimates the relationship between the target and the independent variable.
- It is used to find the trends in data.
- It helps to predict real/continuous values.
- 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

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

Linear Regression

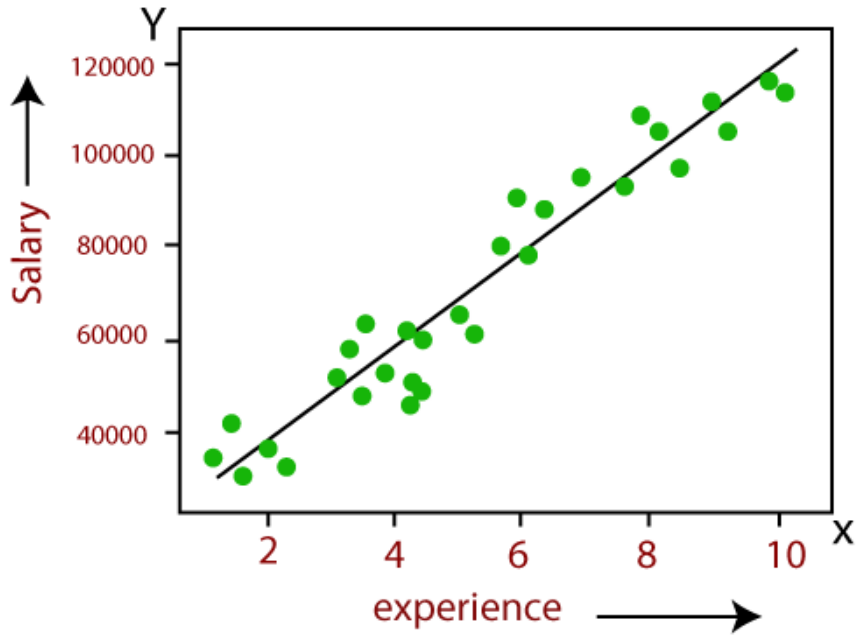
- It shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.

- 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.

- Mathematical equation for Linear regression is $Y = aX + b$, Where Y = dependent variables (target variables), X= Independent variables (predictor variables), a and b are the linear coefficients

Linear Regression

- 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.



Simple Linear Regression

- It is a type of Regression algorithms that models the relationship between a dependent variable and a single independent variable.

- 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.

- Mathematical equation for simple linear regression $y = a0 + a1x + \varepsilon$,

  Where, a0= It is the intercept of the Regression line (can be obtained putting x=0)

  a1= It is the slope of the regression line, which tells whether the line is increasing or decreasing

  ε = The error term. (For a good model it will be negligible)

Implementation of Simple Linear Regression Algorithm using Python

- Problem Statement: Here we are taking a dataset that has two variables: salary (dependent variable) and experience (Independent variable). The goals of this problem is:
  - We want to find out if there is any correlation between these two variables
  - We will find the best fit line for the dataset.
  - How the dependent variable is changing by changing the independent variable.
  - In this section, we will create a Simple Linear Regression model to find out the best fitting line for representing the relationship between these two variables.

Steps to follow for Simple LR modeling

- Step-1: Data Pre-processing
- Step-2: Fitting the Simple Linear Regression to the Training Set
- Step-3: Prediction of test set result
- Step-4: Visualizing the Training set results
- Step-5: Visualizing the Test set results

Multiple Linear Regression

- An extension of Simple Linear regression as it takes more than one predictor variable to predict the response variable
- MLR assumes little or no multicollinearity (correlation between the independent variable) in data.
- MLR equation, $Y= b0+b1x1+ b2x2+ b3x3+..... bnxn$ ..... (a)

  Where,

  Y= Output/Response variable

  b0, b1, b2, b3 , bn....= Coefficients of the model.

  x1, x2, x3, x4,...= Various Independent/feature variable

Implementation of Multiple Linear Regression Algorithm using Python

- We have a dataset of 50 start-up companies. This dataset contains five main information: R&D Spend, Administration Spend, Marketing Spend, State, and Profit for a financial year. Our goal is to create a model that can easily determine which company has a maximum profit, and which is the most affecting factor for the profit of a company.

- Since we need to find the Profit, so it is the dependent variable, and the other four variables are independent variables.

Model Evaluation

- Evaluation metrics for a linear regression model
  - Mean Squared Error (MSE)
  - Mean Absolute Error (MAE)
  - R-squared or Coefficient of Determination
  - Root Mean Squared Error (RMSE)

- Model selection & Subset Regression
  - Adjusted R-squared
  - Stepwise Regression

Types of Model Build

- All-in
- Backward Elimination
- Forward Selection
- Bidirectional Elimination
- Score Comparison

Steps of Backward Elimination

- Step-1: Firstly, We need to select a significance level to stay in the model. (SL=0.05)
- Step-2: Fit the complete model with all possible predictors/independent variables.
- Step-3: Choose the predictor which has the highest P-value, such that.
  - If P-value > SL, go to step 4. Else Finish, and Our model is ready.
- Step-4: Remove that predictor.
- Step-5: Rebuild and fit the model with the remaining variables.

Polynomial Regression

- Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:

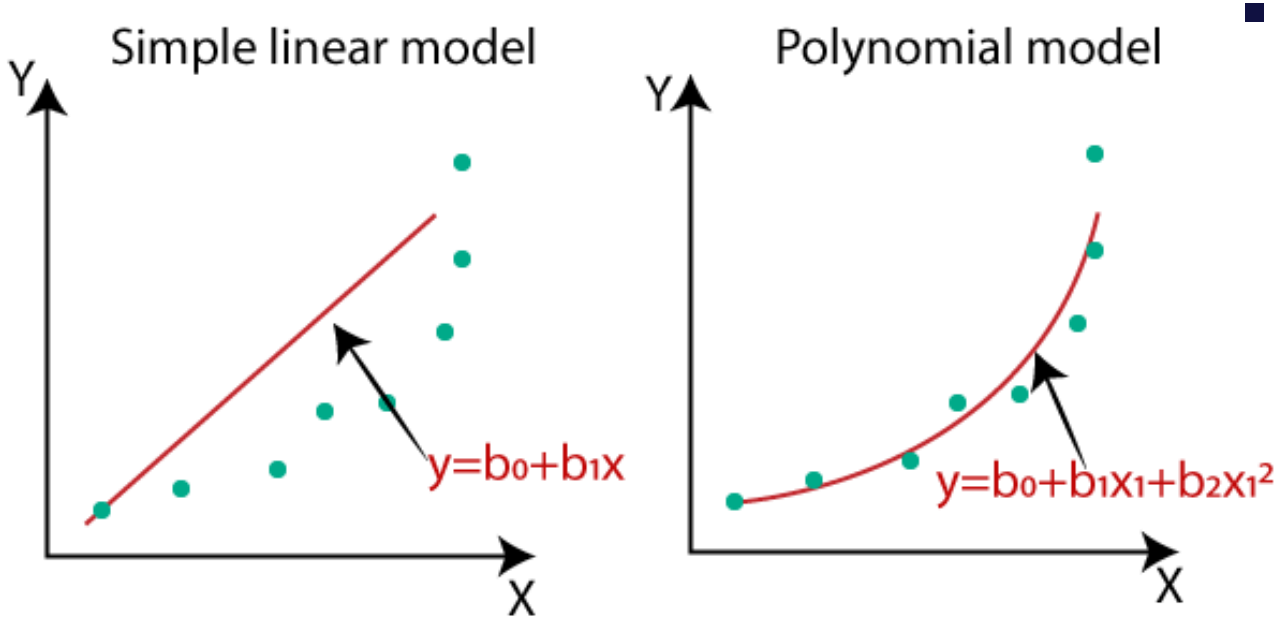
$$y = b_0 + b_1x_1 + b_2x_1^2 + b_2x_1^3 + \ldots\ldots b_nx_1^n$$

Need for Polynomial Regression

- If we apply a linear model on a linear dataset, then it provides us a good result as we have seen in Simple Linear Regression, but if we apply the same model without any modification on a non-linear dataset, then it will produce a drastic output. Due to which loss function will increase, the error rate will be high, and accuracy will be decreased.

- So for such cases, where data points are arranged in a non-linear fashion, we need the Polynomial Regression model. We can understand it in a better way using the below comparison diagram of the linear dataset and non-linear dataset.

Need for Polynomial Regression



- In the image, we have taken a dataset which is arranged non-linearly. So if we try to cover it with a linear model, then we can clearly see that it hardly covers any data point. On the other hand, a curve is suitable to cover most of the data points, which is of the Polynomial model.

- Hence, if the datasets are arranged in a non-linear fashion, then we should use the Polynomial Regression model instead of Simple Linear Regression.

Equation of the Polynomial Regression Model

Simple LR equation: $y = b_0+b_1x$ .....(a)

Multiple LR equation: $y= b_0+b_1x+ b_2x_2+ b_3x_3+....+ b_nx_n$ .....(b)

Polynomial Regression equation: $y= b_0+b_1x + b_2x^2+ b_3x^3+....+ b_nx^n$ .....(c)

- When we compare the above three equations, we can clearly see that all three equations are Polynomial equations but differ by the degree of variables. *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. So if we add a degree to our linear equations, then it will be converted into Polynomial Linear equations*.

Problem Statement

There is a Human Resource company, which is going to hire a new candidate. The candidate has told his previous salary 160K per annum, and the HR have to check whether he is telling the truth or bluff. So, to identify this, they only have a dataset of his previous company in which the salaries of the top 10 positions are mentioned with their levels. By checking the dataset available, we have found that there is a non-linear relationship between the Position levels and the salaries. Our goal is to build a **Bluffing detector regression model,** so HR can hire an honest candidate.

Steps for Polynomial Regression

- Data Pre-processing
- Build a Linear Regression model and fit it to the dataset
- Build a Polynomial Regression model and fit it to the dataset
- Visualize the result for Linear Regression and Polynomial Regression model
- Predicting the output

Regularization

- Regularization Techniques is an unavoidable and important step to improve the model prediction and reduce errors. This is also called the Shrinkage method. Which we use to add the penalty term to control the complex model to avoid overfitting by reducing the variance.

- Regularization Techniques

1. Ridge Regression (L2 Regularization)
2. Lasso Regression (L1 Regularization)