**Linear Regression**

**What is Regression ?**

"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."

**About Algorithm**

* It is supervised learning algorithm.
* It is the algorithm based on two variables where one is used is known as an independent variable whereas another one is known for a dependent variable.
* It can be used for prediction, estimation, hypothesis testing, and modeling causal relationships.
* we try to find a linear function that predicts the response value(y) as accurately as possible as a function of the feature or independent variable(x).

**Terminologies related to this algorithm.**

* **Dependent Variable/ target variable:** The main factor in Regression analysis which we want to predict or understand is called the dependent variable.
* **Independent Variable:** The factors which affect the dependent variables or which are used to predict the values of the dependent variables.
* **Outliers:** Outlier is an observation which contains either very low value or very high value in comparison to other observed values. It should be avoided to get the unhamper result.
* **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**.

**Use cases**

* Generally used in predicting expected the future values on the basis of past data.
* Example : finance, linear regression might be used to understand the relationship between a company’s stock price and its earnings.

**Linear Regression Equation**

* Suppose Y is a dependent variable and X is an independent variable, then the population regression line is given by the equation.

Y= B0+B1X , where

B0 is a constant

B1 is the regression coefficient

**Properties of Linear Regression**

For the regression line where the regression parameters b0 and b1are defined, the following properties are applicable:

* The regression line passes through the mean of X and Y variable values.
* The regression constant b0 is equal to the y-intercept of the linear regression.
* The regression coefficient b1 is the slope of the regression line. Its value is equal to the average change in the dependent variable (Y) for a unit change in the independent variable (X)
* A regression line is used to describe the behavior of a set of data, and helps to find the logical approach to understand the relation between two variables.

**Types of Linear Regression:**

* **Simple Linear Regression:** If a single independent variable is used to predict the value of a numerical dependent variable.
* **Multiple Linear regression:** If more than one independent variable is used to predict the value of a numerical dependent variable.

**Pros and Cons of using this Algorithm.**

* **Pros**
  + - If you know the relationship between two variables, this algorithm is the best to use because of it’s less complexity to compared to other algorithm.
    - It is simple to implement and easier to interpret the output coefficients.
* **Cons**
  + - Outliers can have huge effects on the regression.
    - It assumes a linear relationship between two variable that means it assumes that there is a straight-line relationship between them.
    - It assumes independence between attributes.

**Evaluation matrix used for the performance of the regression model**

* Mean Absolute Error (MAE): It measures the average absolute difference between the predicted values and the true values. MAE is defined as the average of the absolute differences between the predicted values and the true values.

MAE = (1/n) \* Σ|yi - xi|

* Mean Squared Error (MSE): It measures the average squared difference between the predicted values and the true values. MSE is defined as the average of the squared differences between the predicted values and the true values.

MSE = (1/n) \* Σ(yi - xi)^2

* Root Mean Squared Error (RMSE): It measures the square root of the average squared difference between the predicted values and the true values. RMSE is a popular metric because it has the same unit as the target variable, making it easier to interpret.

RMSE = sqrt((1/n) \* Σ(yi - xi)^2)

**Python Libraries that used for this Algorithm**

* NumPy
* Pandas
* Seaborn
* Sklearn