Supervised Learning - Linear Regression

Now that we have understood the basic terminologies needed, we are ready to learn about the key concepts behind supervised learning algorithms such as regression and classification.

Let's start with an example: Suppose you're given a dataset that contains the **area** of different houses and their market price. Your goal is to come up with an algorithm that will take the **area of the house as its input** and **return its market price as the output**.

In this example, the input variable, i.e., the area of the house is called the independent variable (**X**), the output variable, i.e., house price is called the dependent variable (**Y**), and this algorithm is an example of **supervised learning.**

In **supervised learning,** algorithms are trained using the **"labeled"** data (in this example, the dependent variable - house price, is considered a label for each house), and after training the algorithm, we can predict the output for instances where this label (house price) is not known. "Labeled data" in this context simply means that the data is already tagged with the correct output. So, in the above example, we already know the correct market price for each house, and that data is used to teach the algorithm so that it can correctly predict the price of a house if it is not known beforehand.

The reason this paradigm of machine learning is known as supervised learning is that it is similar to the process of supervision that a teacher would conduct on the test results of a student on an examination, for example, the answers that the student gives (predictions) are evaluated against the correct answers (the labels), and the difference (error) is what the student would need to minimize to score perfectly on the exam. This is exactly how supervised machine learning algorithms learn.

There are mainly 2 types of supervised learning algorithms:

. **Regression,** where the output variable is a **continuous** variable, for example, the price of a house.
. **Classification,** where the output variable is **categorical**, for example, approve the loan or not, i.e., yes or no categories.

In this lecture, we will be learning about regression algorithms, which find great use in machine learning prediction of several numerical variables, such as price, income, age, time duration, etc.

Linear Regression

Linear Regression is useful for finding the **linear relationship** between the **independent and dependent** variables of a dataset. In the previous example, the independent variable was the area of the house and the dependent variable was the market price.

This relationship is given by the linear equation: $Y = \theta_0^* + \theta_1^* X + W$

Where θ_0^* is the **constant** term in the equation, θ_1^* is the **coefficient** of the variable X, and W is the difference between the actual value Y and the predicted value $(\theta_0^* + \theta_1^* X)$.

θ_0^* and θ_1^* are called the parameters of the linear equation, while X and Y are the independent and dependent variables, respectively.

What is an error?

With given X and Y in the training data, the aim is to estimate θ_0^* and θ_1^* in such a way that the given equation **fits** the training data **the best**. The **difference** between the **actual value and the predicted value** is called the **error or residual**. Mathematically, it can be given as follows:

$$W = Y - (\theta_0^* + \theta_1^* X)$$

In order to estimate the best fit line, we need to estimate the values of θ_0^* and θ_1^* which requires minimizing the **mean squared error**. To calculate the mean squared error, we add the square of each error term and divide the sum by the total number of records:
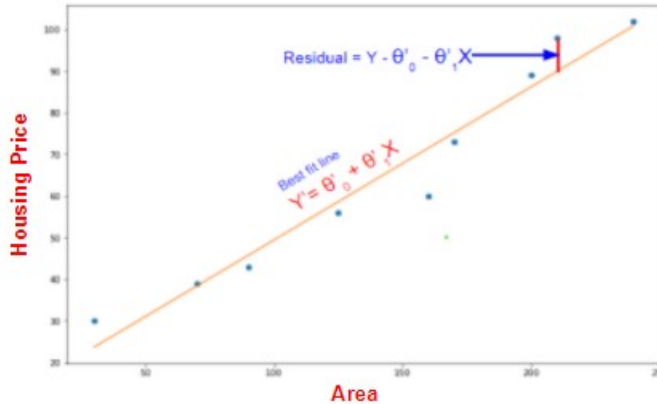
$$Mean\ Squared\ Error = \frac{1}{n}\sum_{i=1}^{n}(Y - \theta_0^* - \theta_1^*X)^2$$

The equation of the best fit line can be given as follows:

$$Y' = \theta_0' + \theta_1'X$$

Where Y' is the predicted value, $\theta_0'\ and\ \theta_1'$ are the estimated parameters.

This equation is called the **Linear Regression model**. The above explanation is demonstrated in the below picture:

Before applying the model to the unseen data, i.e., the data not observed by the model during the training phase, it is important to check its performance to make it reliable. There are a few metrics to measure the performance of a regression model.

. **R-squared:** R-squared is a useful performance metric to understand **how well** the regression model has **fitted over the training data**. For example, an R-squared of 80% reveals that the model is able to capture 80% of the variation in the dependent variable. **A higher R-squared** value indicates a **better fit for the model**.

. **Adjusted R-squared**: The adjusted R-squared is a modified version of R-squared that takes into account the **number of independent variables present in the model**. The R-squared always increases when a new independent variable is added to the model, irrespective of whether that variable adds value to the model or not. Hence, R-squared might be misleading when we have multiple independent variables and cannot identify unnecessary variables included in the model.

However, when a new variable is added, the adjusted R-squared increases if that variable adds value to the model, and decreases if it does not. Hence, adjusted R-squared is a better choice of metric than R-squared to evaluate the quality of a regression model with multiple independent variables, because adjusted R-squared only remains high when all those independent variables are required to predict the value of the dependent variable well; it decreases if there are any independent variables which don't have a significant effect on the predicted variable.

. **RMSE**: RMSE stands for **Root Mean Squared Error**. It is calculated as the square root of the mean of the squared differences between the actual output and the predictions. The lower the RMSE the better the performance of the model. Mathematically, it can be given as follows:

$$Root\ Mean\ Squared\ Error = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \theta_0^* - \theta_i^*X)^2}$$