DATA SCIENCE PROJECT

Here you can understand Mathematics & Calculations involved in linear regresion problem in easyway

Non-Critical & easy to implement

LINEAR REGRESSION:

Linear regression is a simple supervised machine learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (features). In the case of simple linear regression, we have only one independent variable, while multiple linear regression deals with more than one independent variable.

The simple linear regression model is represented as:

$$y = \beta_0 + \beta_1 * x$$

Where:

- y is the dependent variable (Exam Score in our case).
- x is the independent variable (Hours Studied in our case).
- β_0 is the intercept (y-intercept) of the regression line.
- β_1 is the slope of the regression line.

The goal of linear regression is to find the best values for β_0 and β_1 that minimize the difference between the predicted values and the actual values of the dependent variable.

MATHEMATICS BEHIND THE MODEL :

The linear regression model aims to find the best-fit line that minimizes the sum of squared differences between the actual target values and the predicted values. This process is typically achieved using the Ordinary Least Squares (OLS) method.

The cost function (also known as the loss function) for the linear regression model is the Mean Squared Error (MSE). It is defined as the average of the squared differences between the actual target values (y_i) and the predicted values ($\hat{y}_i$) for all data points in the training set:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where n is the number of data points.

The coefficients β_0 and β_1 are estimated using the following formulas:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1\bar{x}$$

Where:

- $\bar{x}$ is the mean of the independent variable (Hours Studied).
- $\bar{y}$ is the mean of the dependent variable (Exam Score).

R - SQUARED:

1. R-squared (Coefficient of Determination):
   The R-squared (or coefficient of determination) is a metric used to evaluate the goodness of fit of the linear regression model. It represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

R-squared is calculated using the formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

A value of R^2 close to 1 indicates that the model explains a large proportion of the variance in the data, while a value close to 0 suggests that the model does not provide a good fit.

In our code example, we used scikit-learn to perform the linear regression. The library takes care of all the underlying mathematics and calculations, including the least squares optimization to find the coefficients and the R-squared calculation.

Understanding the mathematical concepts behind linear regression helps in grasping the intuition of how the model works and how the parameters are estimated to make predictions.

Thank you

Follow me on LinkedIn & GitHub