BWT-Data Science Task 14

Bisma Fajar

July 2024

1 Linear Regression in Machine learning

Linear regression is also a type of machine-learning algorithm more specifically a supervised machine-learning algorithm that learns from the labelled datasets and maps the data points to the most optimized linear functions. which can be used for prediction on new datasets. **Labeled data** means the dataset whose respective target value is already known. **Supervised learning has two types:**

- **Classification:** It predicts the class of the dataset based on the independent input variable. Class is the categorical or discrete values. like the image of an animal is a cat or dog?
- **Regression:** It predicts the continuous output variables based on the independent input variable. like the prediction of house prices based on different parameters like house age, distance from the main road, location, area, etc.

1.1 What is Linear Regression?

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.

When there is only one independent feature, it is known as **Simple Linear Regression**, and when there are more than one feature, it is known as **Multiple Linear Regression**.

Similarly, when there is only one dependent variable, it is considered **Univariate Linear Regression**, while when there are more than one dependent variables, it is known as **Multivariate Regression**.

1.2 Types of Linear Regression

1.2.1 Simple Linear Regression

This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

where:

- y is the dependent variable,
- β_0 is the y-intercept,
- β_1 is the slope of the regression line,
- x is the independent variable,
- ϵ is the error term.

1.2.2 Multiple Linear Regression

This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

$$\mathbf{y} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{x_1} + \beta_\mathbf{2}\mathbf{x_2} + \cdots + \beta_\mathbf{k}\mathbf{x_k} + \epsilon \tag{2}$$

where:

- $\mathbf{y}$ is the dependent variable,
- $\mathbf{x}_1, x_2, \ldots, x_k aretheindependentvariables, \beta_\mathbf{0}$ **is the intercept,**
- $\beta_\mathbf{1}, \beta_\mathbf{2}, \ldots, \beta_\mathbf{k}$ **are the coefficients of the independent variables,**
- ϵ **is the error term.**

1.3 What is the best Fit Line?

In the context of linear regression, the best fit line is a straight line that best represents the data points on a scatter plot. It minimizes the sum of the squared differences (residuals) between the observed values and the values predicted by the line.

1.4 Linear Regression Equation

The equation for a simple linear regression line is given by:

$$\mathbf{y} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{x} + \epsilon \tag{3}$$

where:

- $\mathbf{y}$ is the dependent variable,
- $\mathbf{x}$ is the independent variable,
- $\beta_\mathbf{0}$ is the intercept,
- $\beta_\mathbf{1}$ is the slope,
- ϵ is the error term.

1.5 Objective of Linear Regression

The objective of linear regression is to find the values of β_0 and β_1 that minimize the sum of squared residuals. The residual for each data point is the difference between the observed value and the predicted value:

$$Residual = y_i - (\beta_0 + \beta_1 x_i) \tag{4}$$

1.6 Sum of Squared Residuals

The sum of squared residuals (SSR) is given by:

$$SSR = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 \tag{5}$$

1.7 Least Squares Method

The least squares method is used to estimate the parameters β_0 and β_1 by minimizing the SSR. The estimates are given by:

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} \tag{7}$$

where:

- $\hat{\beta_1}$ is the estimated slope,
- $\hat{\beta_0}$ is the estimated intercept,
- $\bar{x}$ is the mean of the independent variable,
- $\bar{y}$ is the mean of the dependent variable.

1.8 Summary

Linear regression, a supervised machine learning algorithm, models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. It involves finding the coefficients that minimize the sum of squared residuals using methods like least squares. Simple linear regression uses one independent variable, whereas multiple linear regression uses several. The best fit line is the line that minimizes the differences between observed and predicted values, capturing the linear relationship between the variables.