Linear Regression

Epoch IIT Hyderabad

Ankit Saha
AI21BTECH11004

15 Jul 2022

1. Introduction

Linear regression is a machine learning algorithm that predicts real-valued labels for a given input. Estimating real-valued output is known as regression and estimating categorical output is known as classification. Linear regression is one of the easiest and most-commonly used regression algorithms.

Linear regression tries to find a linear relationship between a dependent variable and one or more independent variables (also known as regressors).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{1.1}$$

This can be written in matrix notation as

$$y = \boldsymbol{\beta}^\top \mathbf{x} \tag{1.2}$$

where $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \beta_1 & \cdots & \beta_n \end{pmatrix}^\top$ and $\mathbf{x} = \begin{pmatrix} 1 & x_1 & \cdots & x_n \end{pmatrix}^\top$

Thus, the objective of linear regression is to find the optimal parameter vector $\boldsymbol{\beta}$, i.e. finding the line (or hyperplane) that best fits the given data.

2. Mathematical Formulation

Suppose we are given m labelled samples $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)$ where $x_i \in \mathbb{R}^{n+1}, y_i \in \mathbb{R}\ \forall i \in [m]$
The predicted output matrix is given by $\hat{\mathbf{y}}^\top = \boldsymbol{\beta}^\top \mathbf{X}$ or $\hat{\mathbf{y}} = \mathbf{X}^\top \boldsymbol{\beta}$ where

$$\hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 & \hat{y}_2 & \cdots & \hat{y}_m \end{pmatrix}^\top \tag{2.1}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \beta_1 & \cdots & \beta_n \end{pmatrix}^\top \tag{2.2}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_m \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{m1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{mn} \end{pmatrix} \tag{2.3}$$

The average least squares error is often used as an evaluation metric for linear regression. It is given by

$$J(\boldsymbol{\beta}) = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 \tag{2.4}$$

$$= \frac{1}{m} \left\| \hat{\mathbf{y}} - \mathbf{y} \right\|^2 \tag{2.5}$$

$$= \frac{1}{m} \left\| \mathbf{X}^\top \boldsymbol{\beta} - \mathbf{y} \right\|^2 \tag{2.6}$$

$$= \frac{1}{m} \left(\mathbf{X}^\top \boldsymbol{\beta} - \mathbf{y} \right)^\top \left(\mathbf{X}^\top \boldsymbol{\beta} - \mathbf{y} \right) \tag{2.7}$$

Our goal is to minimize this average least squares error. The optimal parameter vector $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) \tag{2.8}$$

$$= \arg\min_{\boldsymbol{\beta}} \frac{1}{m} \left(\boldsymbol{\beta}^\top \mathbf{X} - \mathbf{y}^\top\right) \left(\mathbf{X}^\top \boldsymbol{\beta} - \mathbf{y}\right) \tag{2.9}$$

$$= \arg\min_{\boldsymbol{\beta}} \frac{1}{m} \left(\boldsymbol{\beta}^\top \mathbf{X}\mathbf{X}^\top \boldsymbol{\beta} - \boldsymbol{\beta}^\top \mathbf{X}\mathbf{y} - \mathbf{y}^\top \mathbf{X}^\top \boldsymbol{\beta} + \mathbf{y}^\top \mathbf{y}\right) \tag{2.10}$$

3. Finding the Optimal Parameters

On differentiating $J(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ (the terms are all scalars),

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left(\boldsymbol{\beta}^\top \mathbf{X}\mathbf{X}^\top \boldsymbol{\beta}\right) = \boldsymbol{\beta}^\top \left((\mathbf{X}\mathbf{X}^\top)^\top + (\mathbf{X}\mathbf{X}^\top)\right) = 2\boldsymbol{\beta}^\top \mathbf{X}\mathbf{X}^\top = 2\mathbf{X}\mathbf{X}^\top \boldsymbol{\beta} \tag{3.1}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left(\boldsymbol{\beta}^\top \mathbf{X}\mathbf{y}\right) = \mathbf{X}\mathbf{y} \tag{3.2}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left(\mathbf{y}^\top \mathbf{X}^\top \boldsymbol{\beta}\right) = \mathbf{X}\mathbf{y} \tag{3.3}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left(\mathbf{y}^\top \mathbf{y}\right) = \mathbf{0} \tag{3.4}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left(J(\boldsymbol{\beta})\right) = 2\mathbf{X}\mathbf{X}^\top \boldsymbol{\beta} - 2\mathbf{X}\mathbf{y} \tag{3.5}$$

On equating the derivative to zero, we get

$$\mathbf{X}\mathbf{X}^\top \hat{\boldsymbol{\beta}} = \mathbf{X}\mathbf{y} \tag{3.6}$$

$$\therefore \hat{\boldsymbol{\beta}} = \left(\mathbf{X}\mathbf{X}^\top\right)^{-1} \mathbf{X}\mathbf{y} \tag{3.7}$$

4. Coefficient of Determination

In some cases, another evaluation metric known as the coefficient of determination (R^2 value) is used.

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{m} (y_i - \bar{y})^2} \tag{4.1}$$

where $\bar{y}$ is the mean of all $y_i, i \in [m]$

$$\bar{y} = \frac{1}{m} \sum_{i=1}^{m} y_i \tag{4.2}$$

The coefficient of determination tells us about how much of the variation in the dependent variable is explained by the variation in the dependent variables. A high R^2 implies that the dependent variable is heavily dependent on the independent variables and our model is accurate and on the other hand, a low R^2 implies that the dependent variable is nearly independent of the dependent variables and our model is thus inaccurate.

5. Questions

i) What does the parameter β_0 represent and why is it important?
ii) Why can't we consider the sum of the differences instead of the sum of the square of the differences in the least squares method?
iii) Similarly, why do we not use the sum of the absolute values of the differences?
iv) What is the R^2 value when the number of samples, $m = 2$?
v) Can linear regression be used for classification? Explain.

6. Answers

i) β_0 represents the y-intercept of the optimal line (or hyperplane). It is important because it provides the model another degree of freedom. If we do not consider β_0, then the optimal line is constrained to pass through the origin, which may not be ideal. β_0 gives the model the freedom to start at any point it wants.
ii) Differences can be both positive and negative. If we simply consider the differences, the positive differences may cancel out the negative differences. Thus, this does not give an accurate idea of the error of the model.
iii) It is easier to solve for $\boldsymbol{\beta}$ analytically when we are using least squares error. If we consider absolute values, not only is it harder to solve because it is not differentiable, it also doesn't necessarily have a unique solution.
iv) $m = 2$ means we are given 2 points as our data set. The best fitting line will be the line passing through these 2 points as we can always find a line passing through any 2 points. Thus, $y_i = \hat{y}_i \ \forall i$ and $R^2 = 1 - 0 = 1$
v) Theoretically, linear regression can be used for classification by using a decision rule. But it is generally not used because linear regression is very sensitive to new data points. We will have to change our decision rule for every new data point added to keep the model accurate, which is not feasible in most cases.