Linear Regression

Advita R.

November 22, 2024

1 Definition and Examples

Regression is when you try to find an equation that is as close as possible to all the given data points.

Example of Regression: When finding the average number of sunspots per month in a year and graphing it, we can see a pattern of it going up and down repeatedly. If we did not have the data for one of these years, but wanted to estimate the average, we could fit a curvy line to the rest of the data to estimate it.

Linear Regression is when you try to find the line of best fit (the **linear** equation that is as close as possible to all the given points).

Example of Linear Regression (or a straight line): Certain types of equations are more fitting to certain data than others. One example of a problem that linear regression can be used on is when you are comparing the experience years of a worker (x-values) and the amount of salary they get (y-value).

2 Regression General Steps

There are four steps to do Regression, including Linear Regression.

1. The first step is to get the data.
2. The second step is about choosing what type of equation would best fit the data.
3. The third step is training the model, or finding the parameters of the equation.
4. The fourth step is testing the model, or seeing if the learned model will be able to answer a question using the equation.

3 Linear Regression Steps

The data will be given as points with x (input-values) and y (output-values).

The linear equation will have have one input (x) and one output (y).

is $y = mx + b$.

The two parameters of this equation are "m" and "b" and need to be found. "m" is the slope of the equation and "b" is the y-intercept of it.

When you have more than one input (x_1 and x_2) and one output (y), the linear equation form is: $m_1 * x_1 + m_2 * x_2 + b$. "$m_1$", "$m_2$", and "$b$" would become the parameters in this situation. In books, this is normally written as: $w_0 + w_1x_1 + w_2x_2$ with "w_0" being "b", "w_1" being "m_1", and "w_2" being "m_2".

We will start with a guess of "m" and "b" and then find the error, or "e". The error is the difference between the given y and the calculated y (ie $\hat{y}$) using the "m" and the "b" We will update "m" and "b" based on this error.

$$e = \frac{1}{2}(\hat{y} - y)^2 = \frac{1}{2}(mx + b - y)^2 \tag{1}$$

The update will be found by finding the derivative of the equation and moving down the slope of the error plot.

$$\begin{aligned} \frac{\partial e}{\partial m} &= \frac{1}{2}(2)(mx + b - y)(x + 0 + 0) = (mx + b - y)x \\ \frac{\partial e}{\partial b} &= \frac{1}{2}(2)(mx + b - y)(0 + 1 + 0) = (mx + b - y) \end{aligned} \tag{2}$$

Algorithm Steps:

1. Data: Given x and y values (points).
2. Algorithm: Linear Regression
3. Training:
   (a) Pick random m and b values for the predicted line.
   (b) Calculate $\hat{y}$, or the y-value when you use the chosen m and b values.
   (c) Find the derivative using 2.
   (d) Update Parameters (m and b)
      i. Update $m = m - \eta(\frac{\partial e}{\partial m})$
      ii. Update $b = b - \eta(\frac{\partial e}{\partial b})$
   (e) Loop step b, c, and d for a certain amount of times