HALİL YAMAN
4/8/20

Linear Regression

Simple Linear Regression

This method is very simple but it doesn't mean that it is not important. Ones we understand linear regression very well, we can interpret and build other complex methods as well. Linear regression is used for predicting Y value based on giving X values. It assumes that there are approximately linear relationship between X and Y.

This relationship is illustrated as $Y \approx \beta_0 + \beta_1 X$

Our goal is to predict the best values for $\beta_0 \; and \; \beta_1$. Once we find the correct values, we can predict the Y value.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$\hat{y}$ indicates the prediction of Y based on X = x.

Estimating The Coefficients

Actually, in the linear notation, β_0 is an intercept, and β_1 is the slope of the linear regression line. Our result line must be as close as possible to data points. There are several ways of measuring *closeness*. The most common approach involves minimizing the least square criterion.

Let's provide a very easy perspective. What is our error when we predict a value by using linear regression?

It is very easy, right? The answer is "actual value minus predicted value". This is our residual for this prediction, and illustrated as follows.

$$e_i = y_i - \hat{y}_i.$$

We define the residual sum of squares (RSS) as follows.

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

$$RSS = \left(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \left(y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2\right)^2 + \cdots + \left(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right)^2$$

The least-squares approach chooses $\hat{\beta}_0\ and\ \hat{\beta}_1$ to minimize the RSS.

Using some calculus we can get the equations for coefficients.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Where $\bar{y}\ and\ \bar{x}$ are the sample means.

Evaluating the Accuracy of the Coefficient Estimates

We get an estimation of coefficients by using the formulas above. Unfortunately, those estimates are not the true $\hat{\beta}_0\ and\ \hat{\beta}_1$values. However, if we repeat this process on the huge amount of different datasets, and calculate the average of estimations, then we get the exact values of $\hat{\beta}_0\ and\ \hat{\beta}_1$.

Well then, how do we know the difference between the exact values and estimations?
There are two other formulas:

$$SE(\hat{\beta}_0)^2 = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Using the formulas above, we calculate the standard errors of the coefficients.

σ is the standard deviation. It is equal to the square root of variance.
Generally, σ is not known but we can calculate the standard deviation from the residual sum of squares.

This estimation is known as residual standard error (RSE). The formula is

$$RSE = \sqrt{RSS/(n-2)}$$

The true value of $\hat{\beta}_1$ is in following interval with 95% of chance. $\hat{\beta}_1 \pm 2 * SE(\hat{\beta}_1)$