Getting started with Machine Learning

Raunak J

So far now we've learned calculating the metrics evaluation.

But what if we have metrics that calculate how well our data fits the model.

So in our case, the line fitting regression model.

So metrics that give us Goodness-of-Fit.

Don't worry, we've got that one already.

It is $\boldsymbol{R^2}$
aka **R-Squared**
aka **Co-efficient of Determination**

Formal Definition :

In statistics, the coefficient of determination, denoted R^2 or r^2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable.

Source : https://en.wikipedia.org/wiki/Coefficient_of_determination

There are many functions with which it is represented :

$$R^2 \; or \; r^2 = \frac{Explained\; Variation}{Total\; Variation}$$

But we'll look towards more easier convention where terms of ANOVA are used :

$$R^2 = 1 - \frac{SSE}{SST} \;\; or \; \frac{SSR}{SST}$$

So considering the terms, what are really SSE, SSR and SST

SSE : Sum of Squares Error

SSR : Sum of Squares Regression

SST : Sum of Squares Total

$$Sum\ of\ Squares\ Error\ (SSE) = \sum_{i=1}^{m} (y_i - \widehat{y_i})^2$$

$$Sum\ of\ Squares\ Regression\ (SSR) = \sum_{i=1}^{m} (\widehat{y_i} - \bar{\widehat{y}})^2$$

$$Sum\ of\ Squares\ Total\ (SST) = \sum_{i=1}^{m} (y_i - \bar{y})^2$$

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

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

Now let us understand with an example we learned earlier

Assume a problem where you have to predict the weight of the person given his height. Here weight is the dependent variable and height is independent variable.

Height of the person in cms (x)	Weight of the person in kgs (y)
160	72
171	76
182	77
180	83
154	76

For the given example we considered Simple Linear Regression

$$\hat{y} = \beta_0 + \beta_{1.} x_i = 41.5648 + (0.208 * x_i)$$

Height of the person in cms (x)	Weight of the person in kgs (y)	Predicted Values y-hat
160	72	74.8448
171	76	77.1328
182	77	79.4208
180	83	79.0048
154	76	73.5968

Now finding the summation and mean of y and y-hat

$$\sum y = 72 + 76 + 77 + 83 + 76 = 384 \; and \; \bar{y} = \frac{\sum y}{N} = \frac{384}{5} = 76.8$$

$$\sum \hat{y} = 74.8448 + 77.1328 + 79.4208 + 79.0048 + 73.5968 = 384 \; and \; \bar{\hat{y}} = \frac{\sum \hat{y}}{N} = \frac{384}{5} = 76.8$$

Height of the person in cms (x)	Weight of the person in kgs (y)	Predicted Values y-hat
160	72	74.8448
171	76	77.1328
182	77	79.4208
180	83	79.0048
154	76	73.5968

$$SSE = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = (72 - 74.8448)^2 + (76 - 77.1328)^2 + (77 - 79.4208)^2 + (83 - 79.0048)^2 + (76 - 73.5968)^2$$

$$\boldsymbol{SSE = 36.9733888}$$

$$SSR = \sum_{i=1}^{m} (\hat{y}_i - \bar{y})^2 = (74.8448 - 76.8)^2 + (77.1328 - 76.8)^2 + (79.4208 - 76.8)^2 + (79.0048 - 76.8)^2 + (73.5968 - 76.8)^2$$

$$\boldsymbol{SSR = 25.9237888}$$

$$SST = \sum_{i=1}^{m} (y_i - \bar{y})^2 = (72 - 76.8)^2 + (76 - 76.8)^2 + (77 - 76.8)^2 + (83 - 76.8)^2 + (76 - 76.8)^2$$

$$\mathbf{SST = 62.8}$$

Now putting the values of SSE, SSR and SST to find R^2

$$R^2 = 1 - \frac{SSE}{SST}$$

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

$$R^2 = 1 - \frac{36.9733888}{62.8} = 0.4112517707$$