回归残差和相关系数

Guangyao Zhao

2022-12-31

Contents

> 回归性能评价指标决定系数 R^2 和皮尔逊相关系数 r 有什么关系，为什么两者都可以作为评价相关性的指标，它们之间有什么内在的联系呢？

两者关系

在机器学习中经常使用回归残差（Sum squared regression, SSR）来评价回归模型的性能；而皮尔逊相关系数（Pearson correlation coefficient）经常用来评价两个变量线性相关性。

回归残差：

$$R^2=\sum_{i=1}^{n}\left(y_i-\hat{y_i}\right)^2 \tag{1}$$

皮尔逊相关系数：

$$r=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}} \tag{2}$$

那么这两者到底有什么关系？先说结论：

对于线性回归的最小二乘拟合：

$$r(x, y) = \pm\sqrt{R^2} \tag{3}$$

对于非线性拟合，也有此关系，证明见 Sec. 。

关系证明

线性回归和最小二乘

线性回归：

$$y = \beta_0 + \beta_1 x + \epsilon \tag{4}$$

其中：$\hat{y} = \beta_0 + \beta_1 x$。用最小二乘法拟合残差的平方和（Sum of square residuals, SSR）得知：

$$SSR = \sum_{i=1}^{n}(\epsilon_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \tag{5}$$

对其求偏导，并使其为零：

$$\begin{aligned} \frac{\partial SSR}{\partial \beta_0} = \sum_{i=1}^{n} 2\left(y_i - \beta_0 - \beta_1 x_i\right)(-1) = 0 \\ \frac{\partial SSR}{\partial \beta_1} = \sum_{i=1}^{n} 2\left(y_i - \beta_0 - \beta_1 x_i\right)(-x_i) = 0 \end{aligned} \tag{6}$$

则：

$$\begin{aligned} \sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right) = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) = 0 \\ \sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right) x_i = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) x_i = 0 \end{aligned} \tag{7}$$

根据上式子第一条：

$$\bar{\hat{y}} = \frac{\sum_{i=1}^{n} \hat{y}_i}{n} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y} \tag{8}$$

相关系数

$$
\begin{aligned}
\rho\left(y_i, \hat{y}_i\right) &= \frac{\operatorname{cov}\left(y_i, \hat{y}_i\right)}{\sqrt{\operatorname{var}\left(y_i\right) \operatorname{var}\left(\hat{y}_i\right)}} \\
&= \frac{\sum_{i=1}^n\left(y_i-\bar{y}\right)\left(\hat{y}_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^n\left(y_i-\bar{y}\right)^2 \sum_{i=1}^n\left(\hat{y}_i-\bar{y}\right)^2}} \\
&= \frac{\sum_{i=1}^n\left(y_i-\hat{y}_i+\hat{y}_i-\bar{y}\right)\left(\hat{y}_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^n\left(y_i-\bar{y}\right)^2 \sum_{i=1}^n\left(\hat{y}_i-\bar{y}\right)^2}} \\
&= \frac{\sum_{i=1}^n\left(y_i-\hat{y}_i\right)\left(\hat{y}_i-\bar{y}\right)+\sum_{i=1}^n\left(\hat{y}_i-\bar{y}\right)^2}{\sqrt{\sum_{i=1}^n\left(y_i-\bar{y}\right)^2 \sum_{i=1}^n\left(\hat{y}_i-\bar{y}\right)^2}} \\
&= \frac{0+\sum_{i=1}^n\left(\hat{y}_i-\bar{y}\right)^2}{\sqrt{\sum_{i=1}^n\left(y_i-\bar{y}\right)^2 \sum_{i=1}^n\left(\hat{y}_i-\bar{y}\right)^2}} \\
&= \sqrt{\frac{\sum_{i=1}^n\left(\hat{y}_i-\bar{y}\right)^2}{\sum_{i=1}^n\left(y_i-\bar{y}\right)^2}} \\
&= \sqrt{R^2}
\end{aligned} \tag{9}
$$

其中:

$$
\begin{aligned}
\sum_{i=1}^n\left(y_i-\hat{y}_i\right)\left(\hat{y}_i-\bar{y}\right) &= \sum_{i=1}^n\left(y_i-\hat{y}_i\right)\left(\beta_0+\beta_1 x_i-\bar{y}\right) \\
&= \left(\beta_0-\bar{y}\right) \sum_{i=1}^n\left(y_i-\hat{y}_i\right)+\beta_1 \sum_{i=1}^n\left(y_i-\hat{y}_i\right) x_i \\
&= 0
\end{aligned} \tag{10}
$$

二次回归的决定系数和皮尔逊相关系数

二次回归决定系数: $\hat{y}=\beta_0+\beta_1 x_1+\beta_2 x_2$, 最小二乘法的残差平方和:

$$
SSR=\sum_{i=1}^n\left(e_i\right)^2=\sum_{i=1}^n\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^n\left(y_i-\beta_0-\beta_1 x_i-\beta_2 x_i^2\right)^2 \tag{11}
$$

对 SSR 求参数的偏导, 令偏导数等于零, 可得最优参数:

$$
\begin{aligned}
\frac{\partial SSR}{\partial \beta_0} &= \sum_{i=1}^n 2\left(y_i-\beta_0-\beta_1 x_i-\beta_2 x_i^2\right)(-1)=0 \\
\frac{\partial SSR}{\partial \beta_1} &= \sum_{i=1}^n 2\left(y_i-\beta_0-\beta_1 x_i-\beta_2 x_i^2\right)\left(-x_i\right)=0 \\
\frac{\partial SSR}{\partial \beta_2} &= \sum_{i=1}^n 2\left(y_i-\beta_0-\beta_1 x_i-\beta_2 x_i^2\right)\left(-x_i^2\right)=0
\end{aligned} \tag{12}
$$

可以得到：

$$\begin{array}{c}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i-\beta_2 x_i^2\right)=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)=0 \\ \sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i-\beta_2 x_i^2\right) x_i=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right) x_i=0 \\ \sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i-\beta_2 x_i^2\right) x_i^2=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right) x_i^2=0\end{array} \tag{13}$$

根据上面式子可以得到：

$$\bar{\hat{y}}=\bar{y} \tag{14}$$

根据相关系数的公式：

$$\begin{aligned}\rho(y, \hat{y}) & =\frac{\operatorname{cov}\left(y_i, \hat{y}_i\right)}{\sqrt{\operatorname{var}\left(y_i\right) \operatorname{var}\left(\hat{y}_i\right)}} \\ & =\frac{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\left(\hat{y}_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 \sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}} \\ & =\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i+\hat{y}_i-\bar{y}\right)\left(\hat{y}_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 \sum_{i=1}^{nb}\left(\hat{y}_i-\bar{y}\right)^2}} \\ & =\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)\left(\hat{y}_i-\bar{y}\right)+\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}{\sqrt{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 \sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}} \\ & =\frac{0+\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}{\sqrt{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 \sum_{i=1}^{nb}\left(\hat{y}_i-\bar{y}\right)^2}} \\ & =\sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}} \\ & =\sqrt{R^2}\end{aligned} \tag{15}$$

其中：

$$\begin{aligned}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)\left(\hat{y}_i-\bar{y}\right) & =\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)\left(\beta_0+\beta_1 x_i+\beta_2 x_i^2-\bar{y}\right) \\ & =\left(\beta_0-\bar{y}\right) \sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)+\beta_1 \sum_{i=1}^{n}\left(y_i-\hat{y}_i\right) x_i+\beta_2 \sum_{i=1}^{n}\left(y_i-\hat{y}_i\right) x_i^2 \\ & =0\end{aligned} \tag{16}$$

可见，只要通过最小二乘法拟合，就能得到 $r(x, y)=\pm\sqrt{R^2}$，进而推广到多项式。任意非线性函数可以由泰勒拟合为多项式，所以进而可以说任意函数都有 $r(x, y)=\pm\sqrt{R^2}$。

易混淆的公式

- Sum square error, SSE: $\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$
- Sum square regression, SSM: $\sum_{i=1}^{n}\left(y_i-\bar{\hat{y}}_i\right)^2$
- Sum square total, SST: $\sum_{i=1}^{n}\left(y_i-\bar{y}_i\right)^2$

其中: $SST = SSE + SSR$, 证明如下:

$$
\begin{aligned}
SST &= \sum_{i=1}^{n}\left(y_i-\bar{y}_i\right)^2 \\
&= \sum_{i=1}^{n}\left(\left(y_i-\hat{y}_i\right)+\left(\hat{y}_i-\bar{y}_i\right)\right)^2 \\
&= \sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2+\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}_i\right)^2+\underbrace{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)\left(\hat{y}_i-\bar{y}\right)}_{value=0}
\end{aligned}
\tag{17}
$$

其中 `value=0` 一项参考 Eq. 10。