

回归问题评估指标

RMSE MAPE R2 R平方

RMSE(Root Mean Square Error, 平方根误差)

平方根误差的计算公式为:

$$RMSE = \sqrt{\frac{\sum_{i=0}^n (y_i - y'_i)^2}{n}}$$

其中 y_i 为训练数据中样本实际值, y'_i 为回归模型的输出值。使用该评价指标时如果在训练样本中存在离群值就会导致指标计算出现较大的偏差, 可能会对模型的评估产生偏差。

MAPE(Mean Absolute Percent Error, MAPE)

先看下MAPE的计算公式:

$$MAPE = median(\frac{|y_i - y'_i|}{y_i})$$

计算的方式为利用样本真实值减去模型输出值比模型真实值, 并对所有样本的计算值取中值作为评估指标, 该方法可以有效的环节离群值带来的问题, 通过 $\frac{|y_i - y'_i|}{y_i}$ 和取中值的方式都对解决离群值的干扰有帮助。

R2(R平方, 又叫决定系数)

其计算公式比较复杂, 如下所示:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$SS_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$SS_{res} = \sum_{i=1}^n (y_i - y'_i)^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

从计算公式，我们可以看到公式1和2其实求解的就是随机预测值时的均方误差，因为作为评价指标，要求这部分计算每次值是固定，因此不能通过真实的模拟随机预测值进行，而是通过随机预测值的期望值为 \bar{y} 作为替换， SS_{res} 值代表的模型的均方误差， $\frac{SS_{res}}{SS_{tot}}$ 则表示 SS_{res} 与随机预测值相比的好坏情况，该部分越大则表明模型的预测值越接近于随机预测，如果该部分大于1则表明模型预测效果比随机预测还要差，因此利用1减去该部分，则得到的值越大，表明 SS_{res} 的值越小，则模型的预测效果越好，得到的值越小，则表明模型预测值越接近于随机预测值，效果越差。我们可以看到在计算最终结果时 $\frac{SS_{res}}{SS_{tot}}$ ，这样的处理方式有利于消除离群值在指标计算时的影响。