

回归模型的几个评价指标

2018年02月26日 14:29:12 Dana-Song 阅读数 22436 更多

版权声明：本文为博主原创文章，遵循 CC 4.0 BY-SA 版权协议，转载请附上原文出处链接和本声明。
本文链接：<https://blog.csdn.net/shy19890510/article/details/79375062>

对于回归模型效果的判断指标经过了几个过程，从SSE到R-square再到Adjusted R-square, 是一个完善的过程：

SSE(误差平方和): The sum of squares due to error

R-square(决定系数): Coefficient of determination

Adjusted R-square: Degree-of-freedom adjusted coefficient of determination

下面我对以上几个名词进行详细的解释下，相信能给大家带来一定的帮助！！

一、SSE(误差平方和)

计算公式如下：

$$SSE = \sum (Y_{actual} - Y_{predict})^2$$

- 同样的数据集的情况下，SSE越小，误差越小，模型效果越好
- 缺点：

SSE数值大小本身没有意义，随着样本增加，SSE必然增加，也就是说，不同的数据集的情况下，SSE比较没有意义

二、R-square(决定系数)

$$R^2 = 1 - \frac{\sum (Y_{actual} - Y_{predict})^2}{\sum (Y_{actual} - Y_{mean})^2}$$

- **数学理解：**分母理解为原始数据的离散程度，分子为预测数据和原始数据的误差，二者相除可以消除原始数据离散程度的影响
- 其实“决定系数”是通过数据的变化来表征一个拟合的好坏。
- 理论上取值范围 $(-\infty, 1]$, 正常取值范围为 $[0, 1]$ -----实际操作中通常会选择拟合较好的曲线计算 R^2 ，因此很少出现 $-\infty$

越接近1，表明方程的变量对y的解释能力越强，这个模型对数据拟合的也较好

越接近0，表明模型拟合的越差

经验值：>0.4，拟合效果好

- 缺点：

数据集的样本越大， R^2 越大，因此，不同数据集的模型结果比较会有一些误差

三、Adjusted R-Square (校正决定系数)

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

n为样本数量，p为特征数量

- 消除了样本数量和特征数量的影响