21 | 模型性能评估（三）：从股价预测产品看回归算法常用的评估指标

2021-02-05 刘海丰

讲述：刘海丰
时长 10:01 大小 9.21M

你好，我是海丰。今天，我们借助一个股价预测产品，来学习回归算法常用的性能评估指标。

股票价格预测模型或者说算法不仅是金融界一项重要的研究课题，也经常和我们的直接经济利益相关，因此一直备受关注。

为了能够准确预测股票未来的价格，很多公司和机构不断尝试开发了很多股票价格预测的模型。但是，对于用算法来进行股票价格的预测这件事情，市场上有两种不同的声音：有的人认为算法是可以预测股票的，并且用 LSTM 算法进行了很多验证；有的人认为股票走势是随机游走的，不论用什么模型预测结果都不可能准确。

不过，这节课，我可不打算和你深入讨论股票预测是否可以用算法实现。我们只会对股票预测模型的结果进行评估，让你知道回归模型的性能评估该用什么指标，以及具体怎么做。

回归算法的评估和分类算法的评估在底层逻辑上是一致的，**都是为了找到真实标签和预测值之间的差异。只是对于分类算法来说，我们关注的是预测分类和实际分类是否相同，而对于回归算法来说，我们关注的是模型是否预测到了正确的数值。** 比如，我们预测一只股票 10 天后的价格是 10 元，在对模型进行评估的时候，你只要看 10 天后的价格和预测价格是否一致就可以了，如果不一致，再看差异有多大。

在回归算法中，常见的性能评估指标主要有 4 个，分别是 MSE（Mean Squared Error，均方误差）、RMSE（Root-mean-squared Error，均方根误差）、MAE（Mean Absolute Error，平均绝对误差 ） 和 R^2（R Squared 决定系数）。

下面，我们就借助一个预测股票的产品，来详细说说它们的原理、计算方法，以及它们是怎么对模型进行性能评估的。

如何计算 MSE、RMSE、MAE?

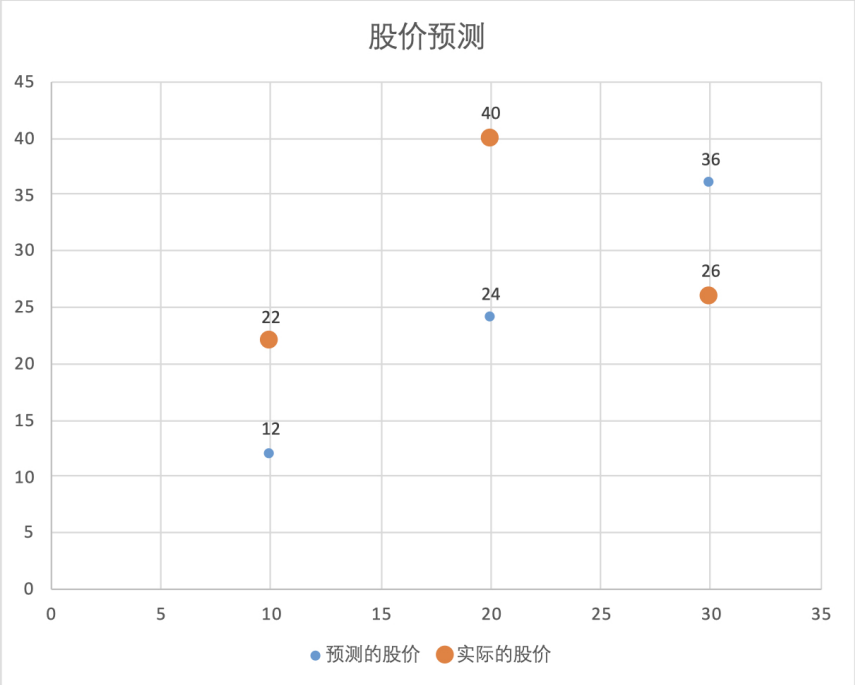
对于预测未来某一天的股票价格来说，我们能想到的最简单的方法，就是用过去一段时间它的平均价格进行预测。

假设，我们用过去 60 天的股票均价来预测这只股票第 2 天的价格。比如说，我们就用 y=1.5*x 这个最简单的算法模型进行预测，那 x 就是过去 60 天的均价，y 是我们预测的第二天的股价。这里我要补充说明一点，因为模型预测第 2 天的价格比预测第 20 天的价格更简单，为了方便理解，我们就用第 2 天举例了。

那怎么验证模型的准确性呢？我们选择三只股票，记录下它们过去 60 天的均价，以及模型预测的股价，等到第 2 天股票价格出来之后，我们再把它们和实际的股价放在一个表格中进行对比，如下图：

股票名称	过去60天平均价格	预测的股价	实际的股价
XX集团	10	12	22
YY集团	20	24	40
ZZ集团	30	36	26

为了表示它们之间的关系，我们建立一个坐标系，以过去 60 天股票均价作为 X 轴，以股票价格作为 y 轴。当我们把这三组数据放入坐标系后，每一个 X 会同时对应一个预测股价和一个实际股价，它们关系如下图所示：



这也就是说，模型每一次预测之后，我们都会得到一个真实值和预测值之间的误差，也就是同一个 X 值的情况下，蓝色点和橘色点之间的差值。那么，是不是我们把得到的是所有误差相加就可以知道这个模型预测准确情况了？

这个问题你可以先自己想一想。下面，我们直接动手来计算一下。根据刚才得到的数据，我们可以直接计算出这三只股票预测值和实际值之间的差值，分别 $22-12=10$，$40-24=16$ 和 $26-36=-10$。

这个时候，如果我们直接把这三次的误差相加，正误差和负误差就会相互抵消。为了避免正负抵消的问题，我们会对每次得到的误差求平方再相加，三次测试的误差平方和就是：$(22-12)^2+(40-24)^2+(26-36)^2=456$。

但是直接用这个数据也是不合理的。因为我们发现，只要测试样本少，即使模型的性能不是非常好，这个数值也不会太大。而且，随着样本的不断增加，即使模型的性能比较好，预测也很准确，这个数值也一定会越来越大。这对测试样本多的模型来说就非常不公平了，那我们该怎么办呢？

这个问题很好解决，我们可以求出所有测试差值的平方和，再让它除以测试样本的数量，公式为：$\frac{(22-12)^2+(40-24)^2+(26-36)^2}{3}=152$。

这就是我们用来表示当前模型性能的一个评估指标——MSE，它的公式如下：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

其中，n 代表测试样本数量，y_i 代表实际值，$\hat{y}_i$ 代表预测值。简单来说，这个指标的计算过程就是先求出所有样本真实值和预测值的差值平方，再除以样本数量。

根据这个公式的逻辑，我们可以知道，MSE 一定是一个大于等于 0，并且无穷大的数值。在对模型进行评估的时候，这个值应该是越小越好。

但是这里还有一个问题，我们在对差值取平方的时候，经常会导致差值的量纲发生变化。比如说，差值的单位是米，那我们对差值取平方，就会导致差值的量纲变成平方米。因此，为了保证量纲相同，我们可以在 MSE 的基础上，再对它求一个平方根。这其实就是 RMSE，它的计算公式是：$RMSE=\sqrt{MSE}$。

我们知道，对差值求平方是 MSE 为了防止正负差异抵消而进行的操作。事实上，要想保证每个样本的差值都是正数，除了求平方之外，我们还可以求每个差值的绝对值。这就是 MAE 了，它和 MSE 一样，也可以用来测量预测误差，它的公式如下：

$$MAE = \frac{1}{n}\sum^{n}|y_i - \hat{y}_i|$$

那么，在实际工作中，这三个指标我们该如何选择呢？**总的来说，算法工程师看得更多的是 MSE，因为 MSE 对差值取了平方，有一个数据放大的过程，更容易发现误差。但是在实际效果评估时候，我们更多地使用 MAE，相对 MSE 来说 MAE 更接近真实误差。**

除此之外，RMSE 因为经过了平方再开方的过程，会导致误差在一定程度上被放大，所以 RMSE 和 MAE 的虽然量纲相同，但是同一个模型的 RMSE 会比 MAE 要大一些。因此，如果你希望更清楚地知道模型差异就选择 RMSE，如果你想了解更真实的模型误差就选择 MAE。

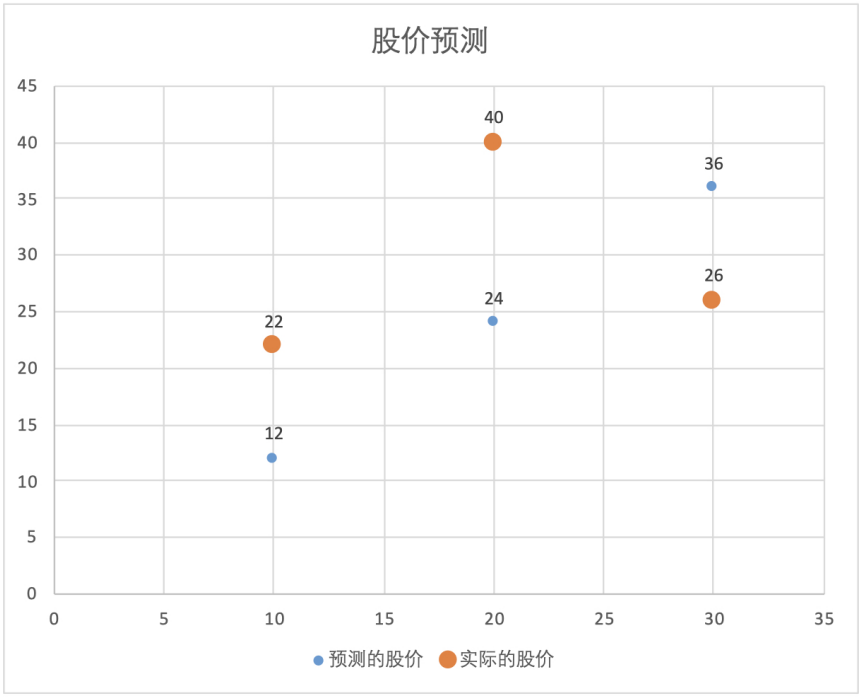
如何计算 R^2？

刚才这 3 个指标的范围都是 0 到正无穷，它们的数值越小代表模型效果越好。但是，当我们想要用一个模型来预测不同场景问题的时候，就会存在不同的预测场景有不同的取值范围。比如，预测股票价格的取值范围就是从几元到几百元；预测房价的取值范围就是几十万到几千万元，而预测身高就又变成了 1 到 2 米左右的数值，可读性非常差。

我们想到，分类算法评估标准的准确率都是 0~1 之间的概率值，非常直观，那么回归模型有没有这样的衡量标准呢？这个时候，我们就可以用 R^2 来进行评估。

R^2 的计算公式是为 $R^2 = \frac{(TSS-RSS)}{TSS}$。其中，TSS 代表总离差平方和，RSS 代表残差平方和。从公式来看，它的分子是模型的预测误差，分母是平均数，所以它的值一般都在 0-1 的范围内，并且它越靠近 1，说明模型预测得越准确。

那么 R^2 具体怎么计算呢？接下来，我就借助刚才股票预测的数据，来试着计算一下。



极客时间

在这个例子中，TSS 就等于每个样本点实际值和实际值均值之间的差值平方和，RSS 就等于每个样本点实际值和预测值之间的差值平方和。具体的计算过程可以分成 4 步：

1. 求出实际值的均值：$\frac{22+40+26}{3} = 29.3$
2. 求出每个实际值和实际值均值之间的差值平方和 TSS：$TSS = （22 - 29.3）^2 + （40 - 29.3）^2 + （26 - 29.3）^2$
3. 求出每个实际值和预测值之间的差值平方和 RSS：$RSS = （22 - 12）^2 + （40 - 24）^2 + （26 - 36）^2$
4. 把 TSS 和 RSS 带入 R^2 公式：$R^2 = \frac{(TSS-RSS)}{TSS}$

小结

这节课，我们讲了回归模型中 4 个非常重要的评估指标。

1. 均方误差 MSE，它的应用最广泛，用来判断预测值和实际值之间误差的指标。它的范围是 0 到正无穷，数值越小代表模型性能越好。
2. 均方根误差 RMSE，它是由 MSE 开根号得到的，也是用来判断预测值和实际值之间误差的指标。它的范围也是 0 到正无穷，数值越小代表模型性能越好。

3. 平均绝对误差 MAE，它的计算过程和 MSE 类似，但是它不对差值求平方，而是直接取绝对值。同样的，它的数值越小代表模型性能越好。
4. 决定系数 R^2，它是实际结果与模型构建的预测值之间的相关系数的平方，决定系数值越高，代表模型效果越好，它的范围一般为 0 到 1。

在使用这几个指标的时候，我们可以参考这 3 点：

1. MAE 相对于 MSE 来说更接近真实误差，所以在评估模型性能的时候，我们会优先选择 MAE；
2. 想要更清楚地知道模型误差就选择 MSE，想要知道更真实的模型误差就选择 MAE；
3. 当我们想要用一个模型来解决不同问题的时候，选择 R^2 可以横向比较这个模型在哪个问题上表现更好。

思考题：

我们今天说了，R^2 的范围一般是 0 到 1，但也可能是负值。那你觉得在哪些情况下，R^2 会是负值？为什么呢？

期待在留言区看到你的思考和答案，我们下节课见！

提建议

上一篇 20 | 模型性能评估（二）：从信用评分产品看什么是KS、AUC?

下一篇 22 | 模型稳定性评估：如何用PSI来评估信用评分产品的稳定性?

精选留言 (1)

写留言

悠悠
2021-02-10

当rss比tss大的时候， 回出现负值，rss越大，说明预测越不准