智信融科笔试实验报告

于雨琛　2024.03.20

Contents

1 对题目的理解与备注

题目要求为给出一个 python 文件，针对同种类型的数据文件都能够给出预测，并且输出结果为 sample id，sample value。由于每只股票的走势不同，因此需要不同的参数，也就是不能提前训练好一个模型后直接加载模型参数，而是需要在每次给出的模型上进行训练并预测。在这种情况下，我会将数据分成训练集与测试集，选择不 shuffle 从而保证数据的时间序列关系。我后续的评价指标仅考虑在测试集上的预测情况，但是在给出预测时由于要从第一个值开始预测因此会涉及到输出训练集，这部分在实际的策略构造中应当拿来作为训练集，然后保证模型参数不变每天预测下一天的值。
在检验中，我采用时间序列交叉验证的方式，分别针对不同长度的训练集进行训练并预测之后一段时间的表现，并给出一个综合的标准。
特别注意，由于给出的数据均为收益率，围绕在 0 附近波动非常剧烈，且展现出强烈随机性，很难通过神经网络来学习并进行预测，我在尝试过后选择根据给出的收益率数据逆向计算出股票的股价高开低收数据，并在训练过程中预测每天的股票收盘价，最终再将结果转换回收益率。

2 整体思路

1. 我目前拥有四个指标，分别为今天开盘价相对昨天收盘价的回报率，今天最高价相对昨天收盘价的回报率，今天最低价相对昨天收盘价的回报率，今天收盘价相对昨天收盘价的回报率。首先，我将第 0 天时的股票收盘价设为 1，根据已知的收益率逆向得到股票的高开低收价格数据，以及新的预测指标，将问题转化为预测股价。

2. 将指标确定好之后，我的下一步是使用多个模型进行尝试，包括线性回归，随机森林，LSTM。在尝试中，我发现 LSTM 的表现效果最好，因此最终模型中我只保留了 LSTM 模型，尝试了单变量预测（只使用过去的股票收盘价数据来预测未来股票收盘价）以及多变量预测（使用高开低收价格数据来预测未来股票收盘价）。

3. 在得到这些结果并分别进行评估后，我会根据 R2 选择表现较好的一个，作为我的模型的预测值。

3 模型结果

3.1 单变量 LSTM

3.1.1 模型介绍

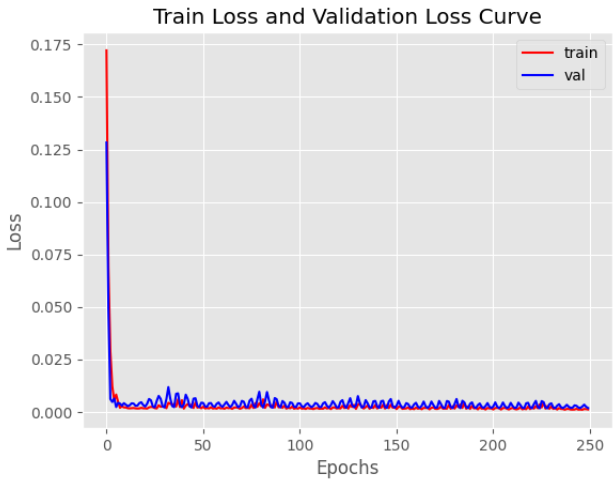
模型:

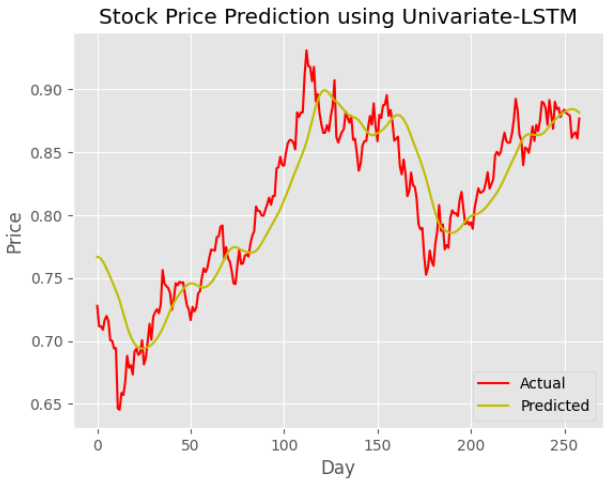
```
model = Sequential()
model.add(LSTM(1,input_shape=(X_train.shape[1],1),return_sequences=True))
for i in range(len(hl)-1):
    model.add(LSTM(hl[i],return_sequences = True))
    model.add(Dropout(0.2))
model.add(LSTM(hl[-1]))
model.add(Dense(1))
model.compile(optimizer = optimizers.Adam(learning_rate=lr),
    loss = 'mean_squared_error')
```

模型超参数:

```
timesteps = 40
hl = [40,35]
lr = 0.0005
batch_size = 64
num_epochs = 200
train_percent = 0.7
val_percent = 0.15
test_percent = 0.15
```

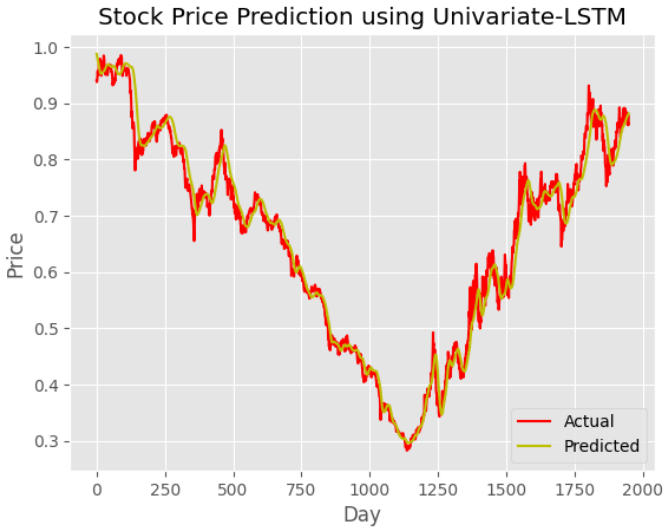
3.1.2 训练结果





验证集结果: RMSE = 0.037, MSE = 0.0014, R-Squared Score = 0.8211
Accuracy = 0.4595, Cumulative Strategy = 0.1750
分析: 从损失函数和预测结果来看可以看出预测的趋势较为准确, R2-score 为 0.82 表示预测能够较好的拟合。Accuracy 为将股价转为回报率的 return, 发现准确率并不够高, 表明这个预测的趋势粒度还无法做到每天的准确预测, 但是能捕捉到股价的整体走势。
cumulative strategy 为我使用预测数据构造的投资策略, 如果我的预测下一天股价高于上一天, 那么我就买入 1 单位股票, 否则卖出 1 单位股票, 根据实际股价数据求和, 来判断这一段时间区间

内我的总收益，若为正表示能够从这个预测中获得正面收益，结果为正符合预期，即模型在捕捉股价整体趋势上是成功的。



全集上的股价预测值与实际值

3.1.3 交叉验证结果

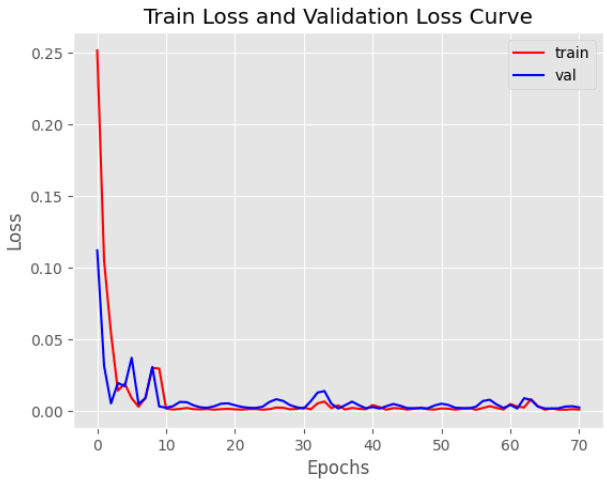
在交叉验证中，我采用时间序列的交叉验证，第一次选用前 600 个数据训练，使用之后 200 个数据进行验证和再之后的 200 个作为测试集；之后选用前 800 个训练，以此类推，直到达到全集。Average result: MSE = 0.0034, RMSE = 0.05815, R-Squared Score = 0.5659, Accuracy = 0.4731, Cumulative Strategy = 0.0817

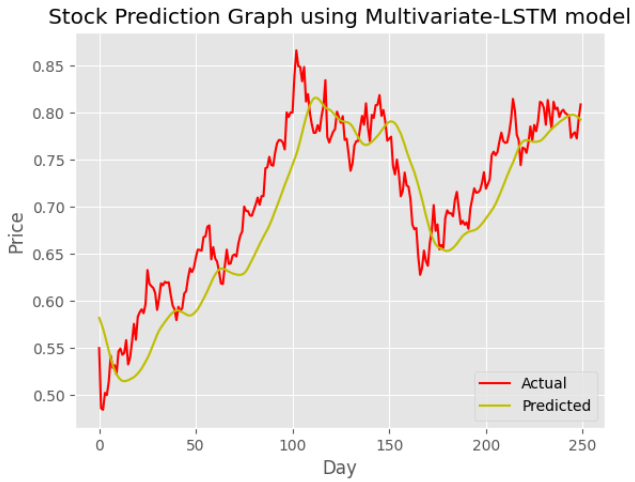
3.2 多变量 LSTM

3.2.1 模型介绍

模型与上述单变量基本一致，输入变为使用股票的高开低收价格四个数据来作为输入，预测下一天的股票收盘价。

3.2.2 训练结果





验证集结果：MSE = 0.0017, RMSE = 0.0417, R-Squared Score = 0.7644
Accuracy = 0.464, Cumulative Strategy = 0.0854

3.2.3 交叉验证结果

MSE = 0.0019, RMSE = 0.04328, R-Squared Score = 0.6295, Accuracy = 0.475, Cumulative Strategy = 0.1879

4 备注

1. 最初我试图在原始的回报率数据上进行训练，但由于回报率的波动性太大，且很难寻找到一个整体的趋势性，(例如即使在股票持续上涨的大趋势下回报率仍然可能是正负持续波动的)，因此很难训练出一个准确的模型。因此，我决定将回报率转化为股价，并对股价进行训练拟合，表现效果显著提高。

2. 起初我采用了包括线性回归，随机森林等简单训练模型，并增加了包括指数平滑、rsi 等因子指标，但在这些模型下很难有很好的预测表现，特别是因为不同股票的走势是不同的，类似线性回归这种模型即使在本题中有较好的表现也很难有很好的泛化能力，因此最终我选用了 LSTM 模型。

3. 实际应用中每次的输出都是 out of sample 预测值，但是在本题中由于我首先需要数据进行训练，所以靠前的输出结果实际是 in sample 预测值（实际情况中应当以历史数据作为训练集，每天更新模型)。