期货价格预测+实验报告

1. 分析问题

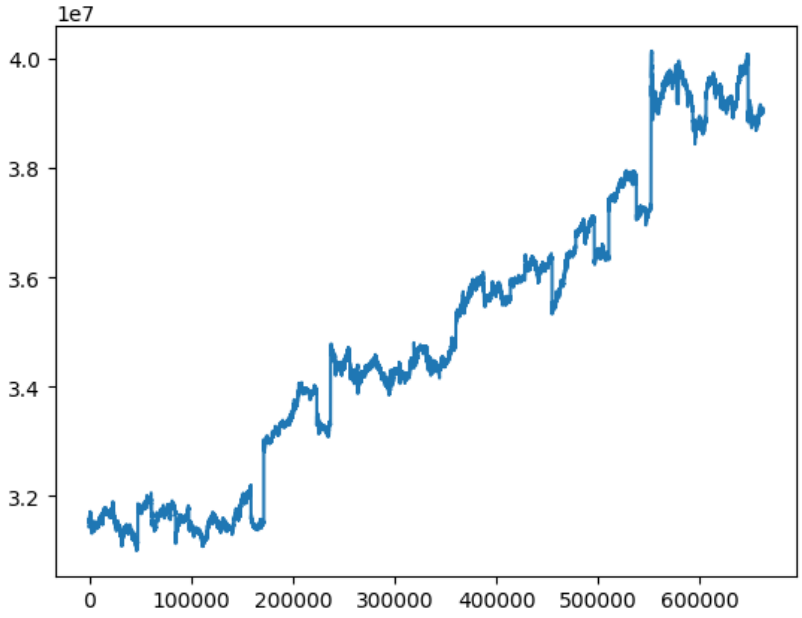
作业要求我们利用期货的历史数据，预测未来的价格走势。我在多种方法中选择了多层 lstm 进行数据预测。

2. 数据预处理：

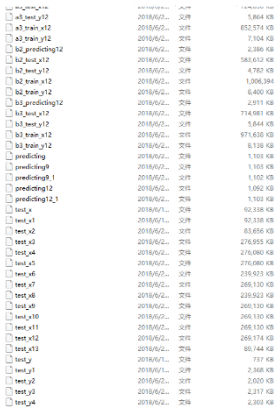
我对数据进行处理，去除极端情况，按照要求，取 k=0.3，若无交易则首项取 k*lastprice，若无没有买方或买方报价，则取另一方报价代替二者报价的均值。

$$P_{0.5}(t) = k \cdot \frac{turnover(t) - turnover(t-0.5)}{volume(t) - volume(t-0.5)} + (1-k) \cdot \frac{bidPrice1(t) + askPrice1(t)}{2}$$

除此之外，我参考上次实验的结果，增加了过去十秒内 bid 量与（ask 量+bid 量）之比这一特征。针对 a1 合约的价格绘制可得到以下图形。



做进一步预处理，按作业要求划分训练测试集。将训练集归一化，并且将测试集按训练集的参数进行标准化，以实现接受完全陌生的数据。接着我尝试了 13 种规划数据的方式，并且根据 a1 合约的结果对其进行分析，最后选择了两种进行最后的分析。一是取当前时间前 30 秒内的所有数据作为时间序列输入模型，以当前时间后 5 秒到后 25 秒内的偏差最大者作为标签；二是取当前时间前 30 秒内的所有数据作为时间序列输入模型，以当前时间后第 10 秒的价格作为标签。



3. 构建模型

将以上数据使用 TensorFlow 构建一个多层 lstm 模型，第一层为一个带 dropout 和 relu 函数激活的全连接层，接下来是两层 32 个单元的，带 dropout 的 lstm

层。最后用一个全连接层，输出数据。最后再利用训练集的参数复原归一化后的输出。

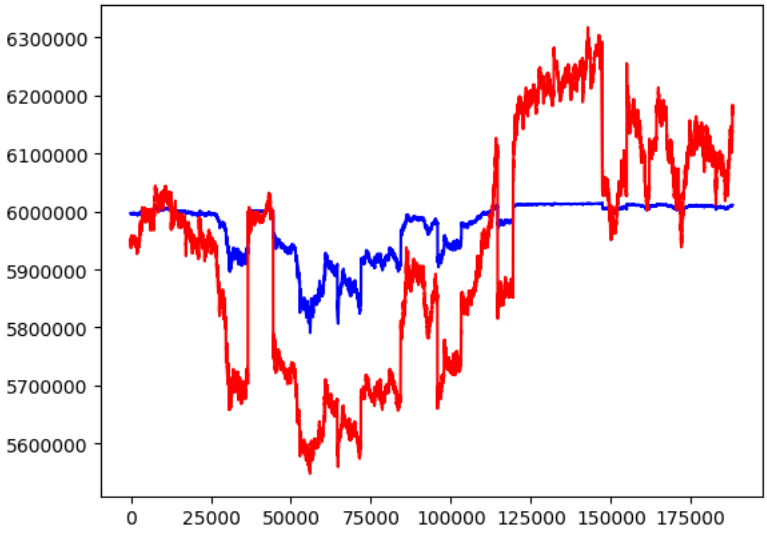
初始化采用常见的正态、常数初始化。在计算损失时采用结果与标签差的平方的平均值。使用学习率衰减，以及 Adam 优化器。

在构建模型时，我试过多种模型，包括以类别向量（onehot）为输出的模型以及输出全序列的模型（即预测下一个时间序列的值），效果均不理想。

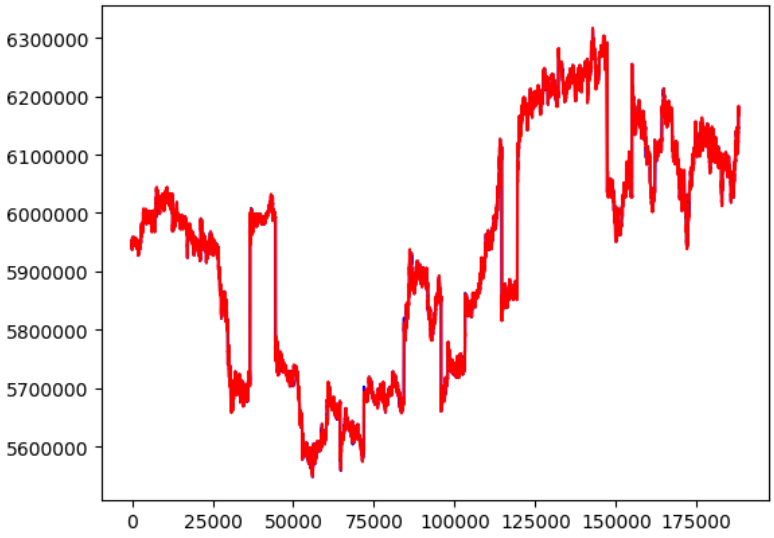
4. 数据测试

在进行完整预测前，我先单独用 a1 的数据对几个特征进行初步的测试，结果如下：

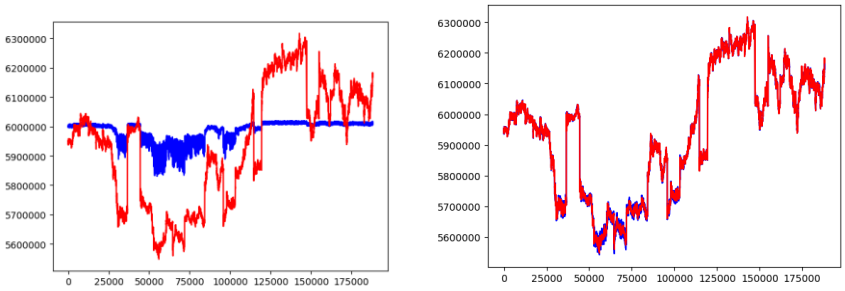
（1） 采用第一种，以当前时间后 5 秒到后 25 秒内的偏差最大者作为标签，我先只输入价格进行训练，发现得到的结果虽然很好地反映了价格变化的趋势，但是绝对偏差过大。红色为标准值，蓝色为预测值。



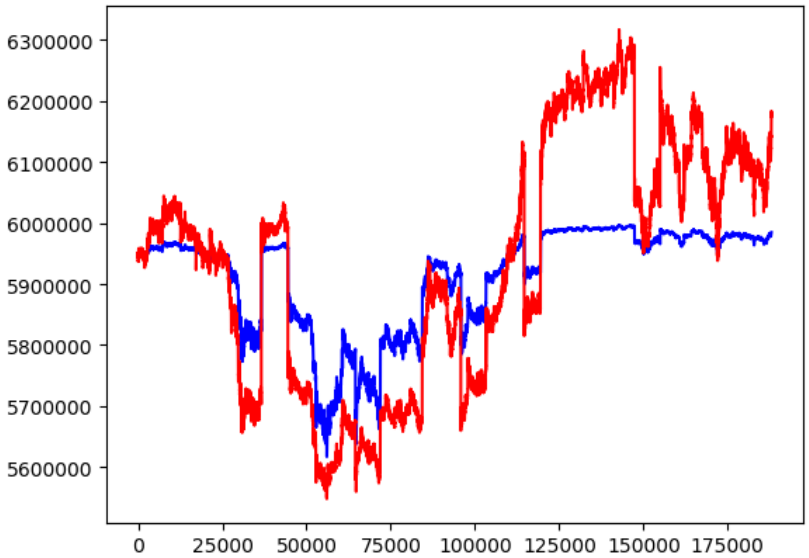
因此我将当前时间前 3 秒的标签（实际价格）的平均值除以预测价格的平均值，得到的系数再乘以当前预测的结果，得到一个较为贴近的预测序列。

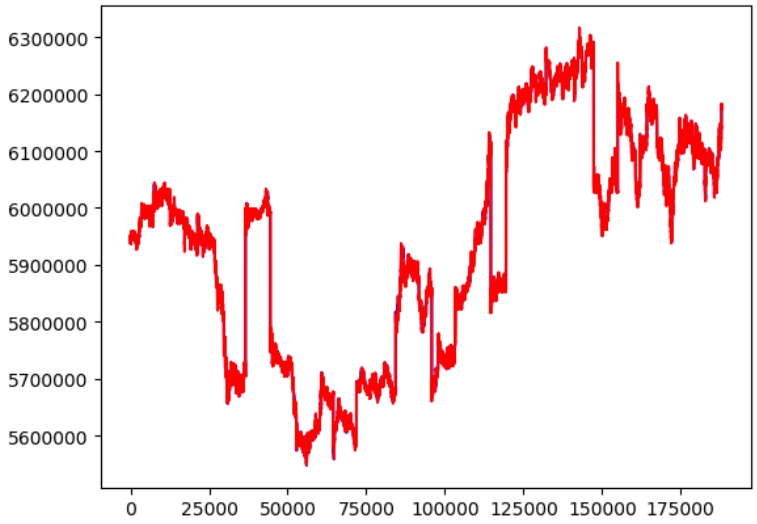


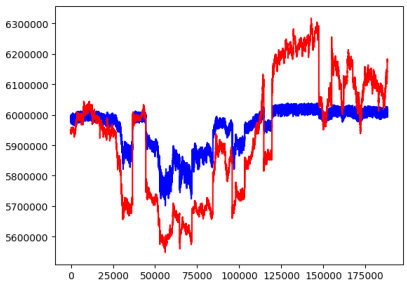
接着将第二项特征即叫价量的比值输入进模型，发现效果并不好，出现了很多噪声。

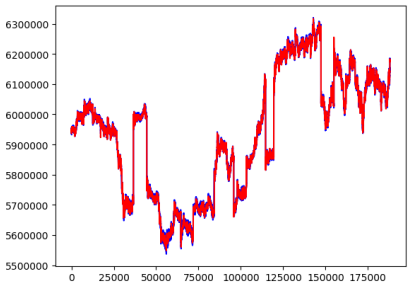


(2) 对第二种，以当前时间后第 10 秒的价格作为标签，我也进行了相同的操作:









可以用肉眼看出第二种比第一种有较小的提升。

(3) 将结果进行分类，有

a. 第一种标签

i. 未加入第二项特征

总类别[6178 175694 6108]（分别为下跌 0.15%，稳定，上涨）

正确类别[3785 173286 3557]

总分类[4969 178226 4785]

精确率[0.76172268 0.97228238 0.74336468]

召回率[0.61265782 0.98629435 0.58235102]

ii. 加入第二项特征

总类别[6178 175694 6108]

正确类别[3671 161749 3552]

总分类[10583 166755 10642]

精确率[0.34687707 0.96997991 0.33377185]

召回率[0.59420524 0.92062905 0.58153242]

b. 第二种标签

i. 未加入第二项特征

总类别[5003 177690 5287]（预处理方法不同，数据量不同）

正确类别[3313 175260 3526]

总分类[4505 178680 4795]

精确率[0.73540511 0.98085964 0.73534932]
召回率[0.66220268 0.9863245 0.66691886]

ii. 加入第二项特征

总类别[5003 177690 5287]（预处理方法不同，数据量不同）
正确类别[3121 159051   3213]
总分类[12153 162972 12855]
精确率[0.25680902 0.97594065 0.24994166]
召回率[0.6238257 0.89510383 0.60771704]

(4) 确实与图表现相同。从数据可以发现，采用 0.15%作为阈值分析会造成数据集中在中间的类别。采用 0.055%作为分类，有

a. 第一种标签

i. 未加入第二项特征

总类别[44689 98147 45144]
正确类别[30237 86255 30602]
总分类[36322 114840 36818]
精确率[0.83247068 0.75108847 0.83116954]
召回率[0.67660946 0.87883481 0.67787524]

ii. 加入第二项特征

总类别[44689 98147 45144]
正确类别[28499 68097 28490]
总分类[45335 97260 45385]
精确率[0.6286313 0.70015423 0.62774044]
召回率[0.63771845 0.69382661 0.63109162]

b. 第二种标签

i. 未加入第二项特征

总类别[42055 102976 42949]
正确类别[29491 90403 30108]
总分类[35977 115545 36458]
精确率[0.81971815 0.78240512 0.82582698]
召回率[0.70124837 0.87790359 0.70101749]

ii. 加入第二项特征

总类别[42055 102976 42949]
正确类别[25950 56242 26238]
总分类[51701 84276 52003]
精确率[0.50192453 0.66735488 0.50454781]
召回率[0.6170491 0.5461661 0.61091061]

可以看出，在当前阈值下，b.i.有较好的结果。

在 0.15%阈值下，四种合约类别（a1,a3,b2,b3）分别占：

[6540, 598327, 6417] [0.01069879 0.97880363 0.01049758]
[8200, 567051, 7490] [0.01407143 0.97307552 0.01285305]
[1881, 683275, 1780] [0.00273825 0.99467054 0.00259122]
[1976, 657941, 1776] [0.00298628 0.9943297 0.00268402]

在 0.055%阈值下，分类比例为：

[84717, 439753, 86814] [0.13858861 0.7193923 0.14201909]
[90001, 399217, 93523] [0.15444426 0.68506764 0.16048811]
[33278, 617664, 35994] [0.04844411 0.899158   0.05239789]
[33239, 592476, 35978] [0.05023327 0.89539409 0.05437265]

5. 预测

（5） 采用当前时间后第 10 秒的价格作为特征标签，得到结果：

a. A1

i. 0.15%

总类别[5003 177690 5287]
正确类别[3313 175260   3526]
总分类[4505 178680   4795]
精确率[0.73540511 0.98085964 0.73534932]
召回率[0.66220268 0.9863245 0.66691886]
上下均值【0.735377215 0. 66456077】

ii. 0.055%

总类别[42055 102976 42949]
正确类别[29491 90403 30108]
总分类[35977 115545 36458]
精确率[0.81971815 0.78240512 0.82582698]
召回率[0.70124837 0.87790359 0.70101749]
上下均值【0. 82277256 0. 70113293】

b. A3

i. 0.15%

总类别[6373 496059   6548]
正确类别[4140 494011   4104]
总分类[5151 498619   5210]
精确率[0.80372743 0.99075847 0.78771593]
召回率[0.64961557 0.99587146 0.62675626]
上下均值【0.79572168 0.638185915】

ii. 0.055%

总类别[79315 349219 80446]
正确类别[57242 331569 58645]
总分类[66184 375034 67762]
精确率[0.86489182 0.88410384 0.86545557]
召回率[0.7217046 0.94945865 0.72899833]
上下均值【0.865173695 0.725351465】

c. B2

i. 0.15%

总类别[737 405354   889]
正确类别[535 404953   681]
总分类[721 405337   922]
精确率[0.74202497 0.99905264 0.73861171]
召回率[0.72591588 0.99901074 0.76602925]

上下均值【0.74031834 0.745972565】

ii. 0.055%

总类别[15821 375866 15293]
正确类别[10973 368683 10353]
总分类[14652 378381 13947]
精确率[0.748908 0.97436975 0.74231017]
召回率[0.69357183 0.98088947 0.67697639]
上下均值【0.745609085 0.68527411】

d. B3

i. 0.15%

总类别[1636 494729 1615]
正确类别[1135 493855 1141]
总分类[1561 494808 1611]
精确率[0.72709801 0.998074 0.70825574]
召回率[0.69376528 0.99823338 0.70650155]
上下均值【0.717676875 0.700133415】

ii. 0.055%

总类别[29241 439228 29511]
正确类别[21829 424193 21820]
总分类[29331 439186 29463]
精确率[0.74422965 0.96586184 0.74058989]
召回率[0.7465203 0.96576949 0.73938531]
上下均值【0.74240977 0.742952805】

可以发现数据较为准确，特别是中间稳定的类别，精确率和召回率都较高，上下两类也 不错。

取最优的结果:

0.15%：p=0.843648525, r=0.801066985

0.055%：p=0.846673192, r=0.808639447

大作业作者：何凌飞 2016011256