期货价格预测实验报告

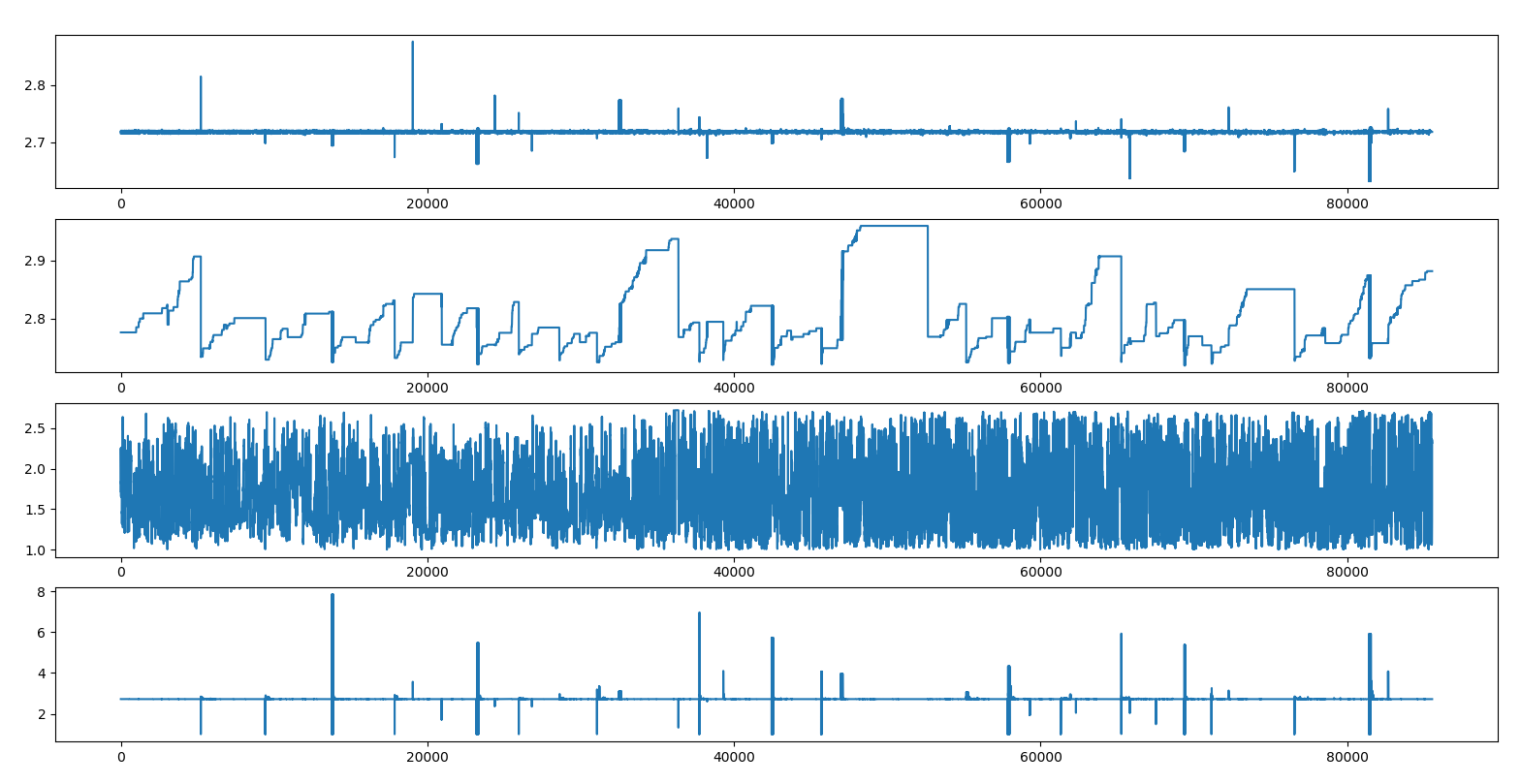
1. **分析问题**

作业要求我们利用期货的历史数据，预测未来的价格走势。分析数据后，我发现最高最低价与价格有很大的相关性，同时交易量根据日期变化有很强的周期性。

1. **数据预处理：**

我对数据进行处理，去除极端情况，选取：

1. 当前价格与前十秒价格平均值的比
2. 过去十秒最高价与最低价之比
3. 过去十秒内bid量与（ask量+bid量）之比
4. 当前交易量与前十秒平均交易量之比



根据当前价格与后10秒平均价格之比，取阈值0.0004分成大幅上涨，上涨，震荡，下跌，大幅下跌五类。

可以发现，为了使数据按类别分布较为均衡，大幅上涨和大幅下跌的数据很少，数据比例约为1:30:70:30:1，因此只需三类即可。

1. **构建模型**

将以上数据使用hmmlearn构建模型，分别构建5个HMM模型。将测试集分别输入5个模型，通过维特比算法分别得出后验概率，概率最高者即为预测的类别。

输入数据的时候，以20组数据为一个序列输入，即最早会用到20秒前的数据。

HMM模型使用第三方库hmmlearn构建，在HMM中采用6个隐状态，使用高斯模型使发射矩阵变为发射函数，得到中心和标准差，可以通过连续的特征训练模型。在训练时，本来将所有类别的期货写在一个文件中执行，为了效率使用多个文件同时执行。

1. **数据测试**

在进行完整预测前，我先单独以四个数据类别为特征训练出四组HMM模型，并进行预测，结果如下：

* 1. 当前价格与前十秒价格平均值的比

准确率：[0.3 0.01428571 0.6 ]

召回率：[0.18 1. 0.2278481]

偏差值（分类结果与准确类别的差的绝对值的平均）：1.02

* 1. 过去十秒最高价与最低价之比

准确率：[0.0 0.0 1.0]

召回率：[0.0 0. 0. 23076923]

偏差值（分类结果与准确类别的差的绝对值的平均）：0.98

* 1. 过去十秒内bid量与（ask量+bid量）之比

准确率：[0.46666667 0.54285714 0.13333333]

召回率：[0.28 0.6031746 0.23529412]

偏差值（分类结果与准确类别的差的绝对值的平均）：0. 689393939

* 1. 当前交易量与前十秒平均交易量之比

准确率：[1. 0. 0.03333333]

召回率：[0.23255814 0 1. ]

偏差值（分类结果与准确类别的差的绝对值的平均）：0. 97

可以看出，第一、三类数据对分类有较大的价值，第二、四类基本没有作用。第一类数据主要分类上涨下跌的情况，第三类主要分类保持平稳的情况。我认为，历史价格是反应波动的最主要数据，因此第一类数据较为有效。同样，价格的升降也可以通过交易量反应，卖的人多自然会导致价格向下波动，买的人多则相反。因此第三类数据似乎更加有效。但是第二、四类数据的表现出乎我的意料，我原本以为过去的价格极差变化可以反应波动，后来仔细一想，不理想的原因：1.可能是这类样本变化太小，原始数据记录较为密集，因此变化小 2.这类数据似乎有明显的时间波动，受外界干扰。第四类数据则表现市场的火热程度，但是交易量大小似乎与价格变化没有直接关系，上涨、下跌皆有可能引起大规模交易。因此这两类数据表现不好也可以理解。

1. **预测**

采用第一、三类数据：当前价格与前十秒价格平均值的比、过去十秒内bid量与（ask量+bid量）之比，作为输入训练HMM模型，得到结果：

A1:

准确率：0.5, 0.6, 0.5

召回率：0.5, 0.8, 0.6

A3:

准确率：0.4, 0.3, 0.3

召回率：0.4, 0.4, 0.5

B2:

准确率：0.6, 0.5, 0.4

召回率：0.3, 0.6, 0.7

B3:

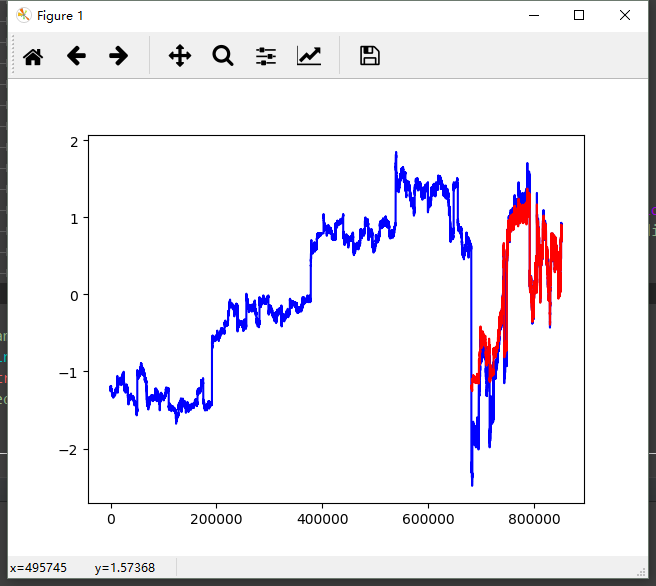
准确率：0.7, 0.4, 0.8

召回率：0.4, 0.7, 0.7

可以发现数据准确率较高。

1. 额外使用lstm预测价格

对于上述结果仍不够满意，于是采用lstm，以价格为特征输入，根据过去10秒的价格预测下一秒的价格。



（训练街与测试集分别归一化，因此图的比例不同）

可以看出虽然精确的价格没法完全预测，但是价格趋势可以很好的反应出来。

大作业作者：何凌飞 2016011256