上海交通大学 & Binquant 高频量化材料1

高频交易收益预测研究

问题背景：高频交易一般指金融领域高频率的自动化交易。高频交易有如下特点：指令完全由电脑发送、系统由专用的软硬件组成、交易指令直接发送至交易所、高速处理大量订单的自动化交易。利用人眼无法检测到的市场条件，高频交易可以在几分之一秒内发现盈利潜力，因此具有较大的研究价值。除此之外，高频交易对整体市场而言也具有好处。高频交易使得买卖价差显著降低，导致市场更有效率；而且，高频交易创造了高流动性，从而缓解了市场碎片化；再者，高频交易由于订单数量大，可以协助价格发现和价格形成过程。

限价订单簿与逐笔交易是证券交易的重要组成部分。它体现了市场上证券的即时供需情况和市场中的买卖信息，其中隐含了投资者向市场传递的不同预期，促使投资者根据市场预期调整自己的交易策略，推动证券价格向新的方向移动。市场交易的信息通过（待）成交量和价格的方式体现，从而影响证券价格形成和发现。

常见的订单簿会显示即时买卖各五个价格，即同一时间可以看到5个买盘价格和5个卖盘价格，分别为：买一、买二、买三、买四、买五；卖一、卖二、卖三、卖四、卖五；其中未成交的最低卖价就是卖一，次低卖价是卖二，未成交的最高买价就是买一，次高买价是买二，其余类推。每个买盘价格和卖盘价格下都有对应的待成交量。下单成交先基于价格优先后基于时间优先排序。例如某证券的最新报价：

- 卖二 (ask price 2) 10.02元
- 卖一 (ask price 2) 10.01元
- 买一 (bid price 1) 10.00元
- 买二 (bid price 2) 9.99元

考虑买入的情况，当报价高于卖一，即高于10.01元的任何价位，就可即时成交，成交价是10.01元。如果报价是9.99元，那么就得排在其他报价9.99元买入的投资者之后，等到即价格跌至9.99元且队列前方的所有报价被成交完，之前下的单才有机会成交。

数据和任务：

Filegetter库提供了Binance与OKEx主流币种期货合约从2022年3月1日之后的逐笔交易与委托信息（Trades Quotes），时间粒度不固定，一般在100ms以下。附件数据中各字段说明如下：

请根据业务背景知识和附件中的数据信息，建立数学模型，对收益进行预测。

表 1: Orderbook 字段说明

ts	时间
bp1, bp2, bp3, bp4, bp5	买方一到五档价格
sp1, sp2, sp3, sp4, sp5	卖方一到五档价格
bv1, bv2, bv3, bv4, bv5	买方一到五档待成交量
sv1, sv2, sv3, sv4, sv5	卖方一到五档待成交量

表 2: Trades字段说明

ts	时间
p	成交价格
v	成交量
sign(v)	主动成交方向

问题1： 请使用t时刻构建因子，对股指期货t到$t+15$秒的收益进行建模预测。你可以使用一档的待成交量的买卖压差即bv1-sv1作为因子，也可以自行构造其他因子。收益的计算方式为买一和卖一均值的对数，在$t+2$时刻和t时刻的差值：

$$\ln\left(\frac{\text{bp1}+\text{sp1}}{2}\right)\Big|_{t+15} - \ln\left(\frac{\text{bp1}+\text{sp1}}{2}\right)\Big|_{t}$$

请使用已有数据，合理划分测试集和训练集，训练你的模型，并给出模型的误差分析。误差分析至少需包括：1. 模型在测试集上的预测效果（以Pearson Corr描述，定义见公式）；2. 测试集中预测收益在99.9%以上和0.1%以下对应的真实收益的样本均值、样本标准差等样本分布信息。

Pearson R：

$$R = \frac{Cov(f_i, y_i)}{\sigma_{f_i}\sigma_{y_i}} \tag{1}$$

其中y_i为真实值，f_i为预测值, $\bar{y}$是样本均值，$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$。

问题2： 在问题1的基础上，研究除了一档待成交量（或其他你选择的因子）外，多档待成交量对收益预测的影响。进一步引入多档待成交量，改进你的模型。对比改进前后的模型预测效果，并测试当预测收益的时间长度不同时，pearson R的变化规律。

问题3： 时间序列动量效应指证券若在历史一段时间内获得较好（差）收益则未来其仍能继续获得显著高（低）收益的现象。时间序列反转效应是指证券若在历史一段时间内获得较好（差）收益则未来其收益会显著变差（好）。投资者可以依据这一规律对股票的未来收益进行预测，构建动量和反转策略。构建策略的关键之一是探索如何选取历史时间窗口的长短。尝试构建日内

（即因子的构建和预测所用的数据均为同一天）动量/反转因子，在t时刻用$t-k$到t时段的数据预测当日t到$t+15$秒的收益。探索时间窗口的长短即参数k的取值对预测效果的影响。并分析动量因子和反转因子的最优时间窗口长短是否有所不同。

问题4：请进一步改进你的预测模型。例如，你可以从订单薄随时间的变化、主动买入卖出交易量等信息中提取别的因子，或者更高阶次的因子组合。对比改进前后的模型预测效果。并对模型进行解释和讨论。

数据详见相关notebook操作