平台简介

本平台基于Python和Jupyter Notebook，用于测试A股股票高频时序信号的因子特性。本平台经过封装，只需您指定因子计算的方式，运算会在服务器中进行，并返回按照该因子进行交易的损益详情。 本平台所使用的数据为经处理后的Level-2实时行情快照数据（即Tick数据），数据的频率为每3s一个。

微观交易介绍

真正实盘的股票交易中，交易委托分为限价委托与市价委托。市价委托进行买卖会分别以当前市场中最优卖价与最优买价成交（即卖一价与买一价），未成交的部分和限价委托均会留在订单薄（Limit order book，LOB）上。因而LOB中存在bid(买盘)与ask(卖盘)两侧。

在初始的研究阶段，为了使问题简化, 此项目中我们限定研究对象为股票指数。数据限定为为中证500指数tick级别快照数据(行情代码：000905)。只考虑价格与成交金额数据，无orderbook信息, 不考虑任何交易费用、滑点。

运行环境

1、本平台可在Python3.6与Python3.7上运行，默认使用Python3.7。

如使用3.6的环境，请将hft.pyc文件替换为api文件夹下对应版本的文件。

2、本平台的使用需要安装 jupyter notebook。

3、本平台的使用需要安装 prettytable，命令如下：

```
pip install prettytable
```

平台使用方法

在解压后的目录下打开命令行运行jupyter notebook。

```
jupyter notebook
```

运行demo文件，对client_id参数进行修改。

id数据会在微信群内发布，请在群里查看到自己id后，填在下面的client_id参数中，其他参数不做修改。

```
para = {
    "start_date": str(20190201),
    "end_date": str(20190330),
    "product_type": "index",
    "t_list": ["i000905"],
    "factor":""
}
client_id = "My_Id"
```

用于生成因子的函数为calc_factor。该函数接受参数为包含快照数据的DataFrame，返回值为因子值，其类型为Series对象，行数需与传入数据保持一样。在函数内，可以使用包括但不限于pandas和numpy等库的函数对数据进行操作。请注意，建议使用向量化的操作以提高运行的性能，减少逐行遍历等操作。本平台目前设置单次运行时间上限为5分组，超时会自动停止任务。 注意，请务必确保没有使用到任何未来数据，包含未来函数的因子是不被认可的。

calc_factor函数实例如下：

```
code_str = """
def calc_factor(df):
    index_pmr_period = np.log(df["mp"]/df["mp"].shift(20))
    index_pmr_period = index_pmr_period.fillna(0)
    index_pmr_period = index_pmr_period

    return index_pmr_period
"""
para["factor"] = code_str
result_core = factor_post(client_id, para)
```

数据介绍

指数tick快照数据格式为 pandas.Dataframe，数据格式如下

	time	InstrumentID	vol_cum	val_cum	TotalAskVolume	TotalBidVolume	ap1	bp1	av1	bv1	ap2	bp2	av2	bv2	ap3	bp3	av3	bv3	ap4	bp4	av
0	09:30:00	905.0	12960.29	7081654.722	100.0	100.0	4317.1745	4317.1745	100.0	100.0	4317.1745	4317.1745	100.0	100.0	4317.1745	4317.1745	100.0	100.0	4317.1745	4317.1745	100.
1	09:30:03	905.0	12960.29	7081654.722	100.0	100.0	4317.1745	4317.1745	100.0	100.0	4317.1745	4317.1745	100.0	100.0	4317.1745	4317.1745	100.0	100.0	4317.1745	4317.1745	100.
2	09:30:06	905.0	12960.29	7081654.722	100.0	100.0	4317.1745	4317.1745	100.0	100.0	4317.1745	4317.1745	100.0	100.0	4317.1745	4317.1745	100.0	100.0	4317.1745	4317.1745	100.
3	09:30:09	905.0	14291.00	7847503.962	100.0	100.0	4315.3795	4315.3795	100.0	100.0	4315.3795	4315.3795	100.0	100.0	4315.3795	4315.3795	100.0	100.0	4315.3795	4315.3795	100.
4	09:30:12	905.0	14291.00	7847503.962	100.0	100.0	4315.3795	4315.3795	100.0	100.0	4315.3795	4315.3795	100.0	100.0	4315.3795	4315.3795	100.0	100.0	4315.3795	4315.3795	100.

可能使用的字段介绍

time：tick数据的时间戳。指数数据已经过处理，其间隔为3秒。

mp、**wp**和**index**：指数的最新价格。

val：指数成份股在这个tick的总成交金额。

val_cum： val的累计加和。

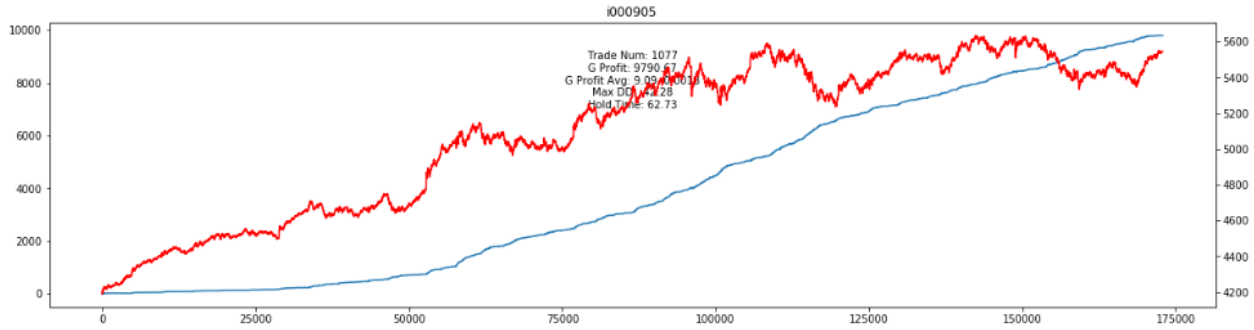
vol与vol_cum为指数成分股的成交量数据，不做使用。此外，由于是指数行情，ap、bp、av和bv等订单簿信息可以忽略。

特别注意：在附件中附有一份样例数据，见sample_data.csv。建议在上传因子之前，使用样例数据对因子函数进行测试，确保因子值能正确计算，且无缺失值与inf。

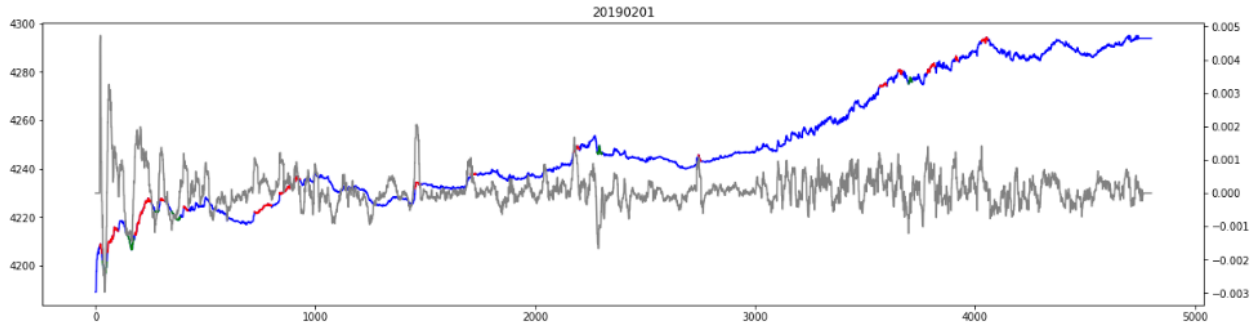
结果与评价

正常回测后会出现如下图表

std	mean	y_mid_10_corr	y_mid_10_mean	y_mid_30_corr	y_mid_30_mean	y_mid_60_corr	y_mid_60_mean
0.000784	2.9e-05	0.488197	0.00051	0.384206	0.000925	0.229733	0.000835



num	profit	hold_time	profit_total	win_rate
1077.0	0.001725	62.733519	1.857462	0.827298



① 第一个表为因子的一些时序统计值，依次为因子的标准差，均值，y_mid_10_corr代表因子值与10个tick之后的收益率的相关系数，y_mid_10_mean代表因子值大于0.85分位的值的未来10tick收益的平均值（考虑多空），依次类推。

② 第二个图为因子的pnl（蓝色）和 index的价格图（红色），pnl的计算方式如下：

阈值定义： 对全部因子值(记为f)求绝对值后取分位数，定义0.85分位数为开仓阈值,记为a；定义0.15分位数为平仓阈值，记为b。

开仓： $f>a$时开多头，$f<-a$的时候开空头。

平仓： 仓位为多头，当$f<-b$时平仓。 仓位为空头，当$f>b$时平仓。

③ 第三个表为pnl的统计信息

num：总交易次数

profit：单笔盈利（百分比）

hold_time：单笔持仓时间（单位为tick）

profit_total：总利润（百分比）

win_rate：胜率（盈利交易次数占总交易次数的比例）

④ 第四个图为回测的第一天的进出场点位图

蓝色为价格曲线，其中标红的地方是因子做多的位置，标绿的地方是因子做空的位置

灰色的曲线为因子的值

注意事项

1、提交结果后，如果在运行过程中停止运算并且重新运行，那么服务器会停止上一次的计算，并重新开始执行你新的计算请求。

2、目前本工具还处于beta阶段，请大家在使用过程中遇到问题或有任何需求及时在群里沟通，沟通能力同样是面试过程中我们需要考察的能力之一。

3、如果运行后服务器异常，可能是带宽问题，请重新提交运行。

特别注意：本平台仅用于实习生选拔过程中的网测部分，请勿将数据与代码文件分享至互联网

研究方向提示

1、初始阶段，可以使用常见的价量指标进行测试来熟悉平台。确保能正确计算出因子值，并对因子中的缺失值、inf等进行处理。指数相比个股，其价格变化十分平滑，因而对应的因子也应当是一个连续的变量。指数的行情也较少出现单边上涨的情形，因而建议因子能呈现以0位均值的分布，且总体上平稳，因子值大于0表示当前偏向看多，因子值小于0表示当前偏向看空。

2、对因子计算的想法并无任何限制，可以来自对市场的理解与直观想法、数据的统计规律，也可以来自从数据中挖掘出的模式。（受限于数据带宽，目前开放的数据有限，有需求的同学可以在群里提出。） 这里我们给出一个例子，提供一些思考的角度，以期能带来一些启发。

样例 我们首先从观察市场入手。使用同花顺(windows版)辅助观察，输入代码399905进入中证500行情界面，从K线图选中一日的行情双击进入小窗。使用历史回放功能可以逐tick回放当日行情。

通过观察行情走势，我们发现指数的价格变化往往是存在一定"惯性"的，即经常在一段时间内连续上涨。因而产生想法：捕捉上涨趋势的开端做多来实现盈利。（做空同理）

为了实现这一想法，我们需要找到恰当的因子，满足在上涨趋势的开端因子值大。于是，我们定义计算短期的涨幅的因子如下：

$$f = \frac{\text{当前价格}}{n\text{个}tick\text{之前的价格}}$$

得到因子的定义，我们需要先考察因子所刻画的与我们的想法是否一致。这里，f值越大意味着过去n个tick内涨幅越大。而我们的想法是捕捉开端。所以选取的n不能过大，不然会错过趋势。我们先使用平台对参数的尝试，比如可以取$n = 5$。

观察结果中的第四个图可以发现，虽然涨幅的开端能捕捉到，但因为计算的回看周期很短，因子捕捉到了许多震荡，而非上涨的趋势。尝试对n的取值进行改进，取$n = 20$。测试结果可以发现，入场的次数变少了，但抓趋势更为准确。 因子上还可以再有改进，比如这个因子的波动和时序上的分布仍不够理想。可以尝试对价格取移动平均来平滑信号，但代价是信号变得滞后。也可以对因子值使用tanh函数，来改变信号的分布。

对短期涨幅的刻画上也可以有其他途径，比如这里的区间涨幅并未考虑区间内的变动。用来刻画的指标可以参考研、学术论文等。研报关于量价因子的分析可以给大家拓展一些思路，学术论文关于时间序列动量的研究也有参考意义。（常见的来源包括但不限于萝卜投研、SSRN、seeking alpha等。）