



清华大学
Tsinghua University

FBDQA期末大作业 暨 AI量化模型竞赛规则说明

金融大数据与量化分析

Financial Big Data and Quantitative Analytics



01 竞赛规则概要



1.1 竞赛规则概要

- 在离线数据上进行模型训练：
 - 公开发布训练集（含验证集）
 - 包括10只（不公开）股票、79个交易日的L1 snapshot数据
 - 数据已进行规范化和隐藏处理
 - 包括5档量/价，中间价，交易量等数据（具体可参考后续数据说明）
- 在私有测试集上进行评测（数据不发布，一次性提交）
- 参赛语言为python，如有特殊需求请联系助教：dlr18@mails.tsinghua.edu.cn

1.2 竞赛规则概要

■ **预测目标：** 利用过往及当前数据预测未来中间价的移动方向

■ **输入数据：**

- 行情频率： 3秒一个数据点（也称为1个tick的snapshot）
- 每个数据点包括当前最新成交价/五档量价/过去3秒内的成交金额等数据
- 训练集中每个数据点包含5个预测标签的标注

■ 允许利用过去不超过100tick（包含当前tick）的数据，预测未来N个tick后的中间价移动方向



1.2 竞赛规则概要： mid-price

■ **预测目标：** 利用过往及当前数据预测未来中间价的移动方向

■ 中间价定义：

$$\text{midprice} = \begin{cases} \frac{\text{ask1} + \text{bid1}}{2}, & \text{if ask1} \neq 0 \text{ and bid1} \neq 0 \\ \text{ask1}, & \text{if bid1} = 0 \text{ (例如跌停)} \\ \text{bid1}, & \text{if ask1} = 0 \text{ (例如涨停)} \end{cases}$$

■ 预测时间跨度：5、10、20、40、60个tick，5个预测任务

■ 即在t时刻，分别预测t+5tick， t+10tick， t+20tick， t+40tick， t+60tick以后：

■ 最新中间价相较t时刻的中间价：下跌/不变/上涨

1.3 竞赛规则概要：移动方向

■预测目标：最新中间价的移动方向

■如何定义midprice的移动方向？

■ x = 待预测时刻的midprice – 当前时刻的midprice

$$\phi(x) = \begin{cases} 0 & x < -\alpha \\ 1 & -\alpha \leq 0 \leq \alpha \\ 2 & \alpha < x \end{cases}$$

■ 0表示下跌，1表示不变，2表示上涨

■ 对于预测窗口为5tick, 10tick, alpha取0.05%

■ 对于预测窗口20tick, 40tick, 60tick, alpha取0.1%



1.4 竞赛规则概要：评分标准

■ 模型评测指标：

■ 客观指标1（30分）：F0.5

- 对每个预测任务，将所有参赛队伍结果按排名分5档，按F0.5得分从高到低排列
- 通过正态分布将得分分布至15-30分之间
- 取5个预测任务中的最高得分

■ 客观指标2（30分）：P&L

- 对每个预测任务，将所有参赛队伍结果按排名分5档，按最终Profit高低从高到低排列
- 通过正态分布将得分分布至15-30分之间
- 取5个预测任务中的最高得分

■ 主观评价（40分）：

- 形式完整性（10分）、方案合理性（10分）、结果有效性（10分）、新颖性（10分）
- 经过路演后由评委打分得出



1.5 竞赛规则概要：奖项设置

■ 奖项设置：

■ 最高得分奖（5个）：

- 对5个预测任务，分别取F0.5最高的队伍为最高得分奖

■ 最高盈利奖（1个）：

- 计算P&L后，取5个任务中的最高的Profit得主为最高盈利奖

■ 最佳方案奖（1个）：

- 结合路演结果和方案效果，综合评选最佳方案奖1名



1.6 竞赛规则概要： F0.5

■ 基础指标：根据交易场景自定义准确率与召回率

■ 结合交易实际，上涨/下跌比不变更重要

■ 计算上涨/下跌样本的准确率与召回率

■ true positive: 即模型预测为上涨或下跌，且预测正确的样本数

■ false positive: 即模型预测为上涨或下跌，但预测错误的样本数

■ true negative: 即真实样本中标注为不变，且模型也预测为不变的样本数

■ false negative: 即真实样本中标注为上涨、下跌，但模型预测错误的样本数

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} = \frac{\text{true positive}}{\text{no.of predicted positive}}$$

$$\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} = \frac{\text{true positive}}{\text{no.of actual positive}}$$

$$\text{F1Score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$



1.6 竞赛规则概要： F0.5

■ F0.5:提高准确率的重要性

- 准确率precision和召回率recall按前文所述计算
- 然后取beta为0.5，计算F0.5的值：

$$F_{\beta} = (1 + \beta^2) \times \frac{precision \times recall}{\beta^2 \times precision + recall}$$

1.7 竞赛规则概要： P&L

■ 简单交易模拟

- 不考虑实际交易规则（实际中难以做空股票等），不考虑手续费/交易费率/可交易性（涨跌停/停牌等），采用中间价（或最优对手价）成交。
- 对于N-tick的预测模型（ $N=5,10,20,40,60$ ），
 - 每当模型预测为下跌时：
 - 做空（卖出）1个单位的股票标的，持有N个tick后平仓（买进），记录本次交易收益率
 - 每当模型预测为上涨时：
 - 做多（买进）1个单位的股票标的，持有N个tick后平仓（卖出），记录本次交易收益率
 - 累计所有交易结果，计算整体收益率






02 数据说明



2.1 文件位置

■数据集发布位置: <https://cloud.tsinghua.edu.cn/published/fbdqa-2021a-mmpchallenge/>

■文件构成:















资料库 / FBDQA-2021A-MMPChallenge		
	FBDQA2021A_MMP_Challenge.tar.gz	163.3 MB 14 天前
	FBDQA2021A_MMP_Challenge.tar.gz.md5	66 bytes 14 天前
	FBDQA2021A_MMP_Challenge_ver0.2.tar.gz	164.4 MB 7 天前
	FBDQA2021A_MMP_Challenge_ver0.2.tar.gz.md5	80 bytes 7 天前
	home.md	4.9 KB 14 天前

■请使用FBDQA2021A_MMP_CHallenge_ver0.2.tar.gz

2.2 数据说明

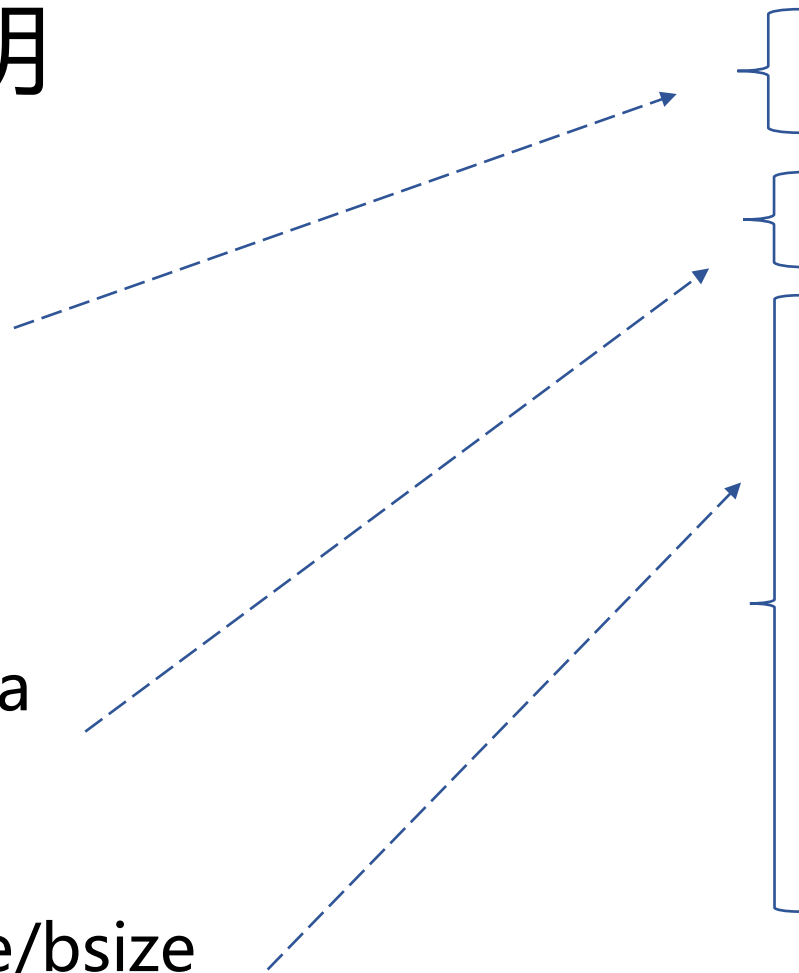
- 共79个交易日，10只标的的股票
- 所有行情文件以
 snapshot_sym<xx>_date<yy>_am/pm.csv
命名，代表第xx只证券标的在第yy天的上午/下午
的行情数据

codes > FBDQA_MMP_Challenge > FBDQA2021A_MMP_Challenge > data

名称	修改日期	类型
 snapshot_sym0_date0_am.csv	2021/11/16 14:34	Microsoft Excel ...
 snapshot_sym0_date0_pm.csv	2021/11/16 14:34	Microsoft Excel ...
 snapshot_sym0_date1_am.csv	2021/11/16 14:34	Microsoft Excel ...
 snapshot_sym0_date1_pm.csv	2021/11/16 14:34	Microsoft Excel ...
 snapshot_sym0_date2_am.csv	2021/11/16 14:34	Microsoft Excel ...
 snapshot_sym0_date2_pm.csv	2021/11/16 14:34	Microsoft Excel ...
 snapshot_sym0_date3_am.csv	2021/11/16 14:34	Microsoft Excel ...
 snapshot_sym0_date3_pm.csv	2021/11/16 14:34	Microsoft Excel ...
 snapshot_sym0_date4_am.csv	2021/11/16 14:34	Microsoft Excel ...
 snapshot_sym0_date4_pm.csv	2021/11/16 14:34	Microsoft Excel ...
 snapshot_sym0_date5_am.csv	2021/11/16 14:34	Microsoft Excel ...
 snapshot_sym0_date5_pm.csv	2021/11/16 14:34	Microsoft Excel ...
 snapshot_sym0_date6_am.csv	2021/11/16 14:34	Microsoft Excel ...
 snapshot_sym0_date6_pm.csv	2021/11/16 14:34	Microsoft Excel ...

2.2 数据说明

- 信息戳
 - 日期
 - 时间
 - 股票代码
- 成交数据
 - amount_delta
 - n_close
- 量价数据
 - ask/bid/asksize/bsize



字段	含义	说明
date	日期	sequential标号：既保留跨标的的可比性，也隐去实际时间
time	时间戳	保留实际时间戳，3s一档行情
sym	标的(仅序号)	
close	最新价/收盘价	以涨跌幅表示
amount_delta	成交量变化	从上个tick到当前tick发生的成交金额
n_midprice	中间价	标准化后的中间价，以涨跌幅表示
n_bid1	买一价	标准化后的买一价，以下类似
n_bsize1	买一量	
n_bid2	买二价	
n_bsize2	买二量	
n_bid3	买三价	
n_bsize3	买三量	
n_bid4	买四价	
n_bsize4	买四量	
n_bid5	买五价	
n_bsize5	买五量	
n_ask1	卖一价	
n_asksize1	卖一量	
n_ask2	卖二价	
n_asksize2	卖二量	
n_ask3	卖三价	
n_asksize3	卖三量	
n_ask4	卖四价	
n_asksize4	卖四量	
n_ask5	卖五价	
n_asksize5	卖五量	
label5	5tick价格移动方向	当前tick中间价相对于5tick之前的移动方向，0为下跌，1为不变，2为上涨
label10	10tick价格移动方向	
label20	20tick价格移动方向	
label40	40tick价格移动方向	
label60	60tick价格移动方向	

2.3 字段说明

字段	含义	说明
date	日期	去掉实际日期，但是保留序号，取值范围0-79，保留跨标的的可比性
time	时间戳	保留实际时间戳，3s一档行情
sym	股票标的	仅序号，10只股票分别为0-9
n_close	最新成交价	无量纲，以涨跌幅表示，-1 ~ 1，即跌100%到涨100%（实际取值范围更小）
amount_delta	成交量变化	从上个tick到当前tick发生的成交金额，单位元
n_midprice	中间价	无量纲，以涨跌幅表示
n_bid1-n_bid5	买一价-买五价	无量纲，以涨跌幅表示
n_ask1-n_ask5	卖一价-卖五价	无量纲，以涨跌幅表示
n_bsize1-n_bsize5	买一量-买五量	无量纲，以换手率表示
n_asisize1-n_asisize5	卖一量-卖五量	无量纲，以换手率表示



03 测试方法



3.1 测试方法

- 不提供公榜测试
- 测试方式为交互式调用
 - 参赛者需要提交源代码及模型，模型文件不超过2GB，FP32精度
- 代码需提供一个Predictor类，对一组数据点进行一次预测
 - 类应至少具有__init__方法和predict方法：
 - `def predict(x:dataframe)->List[int]`
 - 评测程序会多次调用predict，得到单个预测并汇总计算P&L



3.2 predict方法

- 考虑到同学可能会需要多个时间步（不超过100tick）的数据用于单次预测，单次会输入形如[100,26]的数据：
 - 100行为过去100个tick的数据
 - 每行长度为26，与发布数据相同(去除标注项)
- 函数应返回5个int类型预测标签（标签为0/1/2）
- 测试程序样例已发布，具体pnl算法将于近期公布

3.3 调用方式

- 为了公平性与数据保密性，评测程序会打乱测试点输入顺序
 - 单次输入的100个tick为连续数据点
 - 不同次输入的数据无先后关系
 - date: 日期会置为0，无意义
 - sym: 可能的范围0-10（可能有不来自于10只训练集股票的数据，仅作泛化性的额外参考，不列入最终积分）
 - time: time会保留实际交易时间，以备参赛模型会根据时段不同做推理



清华大学
Tsinghua University

THANKS