Project Instruction

The file is only designed as reference for final project of 2021.M2.BigDataAnalysis

Introduction

In the project, you are required to work with high-frequency data to predict one-day forward return. A scenario is assumed so that the result can be evaluated out-of-sample

Scenario

As a strategy team, you daily work is to build models to **predict one-day forward return** (assumes you trade on a daily basis). On day T, you are asked to predict day $T+1$ return, basically, you can choose one definition of return from the following setups:

1. T+1 BOD to T+2 BOD
2. T+1 vwap to T+2 vwap
3. T+1 BOD to T+1 EOD
4. [Extra credits] use T+1 10am close to T+2 10am close

Among all four setups, the 4th setup is the only situation which you can use data on day T+1(but be sure it is prior to 10am, excluded)

You In-sample test period will be: 20190101 - 20190331

Data

Transaction data

Transaction data is stored on HDFS, the dictionary is as follows (Note: if you don't include transaction data in your trial, points will be deducted)

File location: hdfs://hadoop-namenode:9082/tickData/tickData

字段	含义	数据类型	说明
SecurityID	证券代码	STRING	
TradeTime	成交时间	NUMBER	20151123091630O2 精确到百分之一秒
TradePrice	成交价格	NUMBER (3)	
TradeQty	成交量	NUMBER	
TradeAmount	成交金额	NUMBER (3)	
BuyNo	买方订单号	NUMBER	
SellNo	卖方订单号	NUMBER	

Wind RDF

Please refer to official document or visit wds.wind.com.cn in PHBS

Performance Evaluation

Note: Code can be a mixture of Java & MATLAB or Java & Python

Data Preprocessing

In data preprocessing part, you are required to realized the following functions:

1. [Base point 1, Optional, but points will be deducted if not included] Process transaction data and generate some base signal(can use reference attached, though they are not able to make profits in our setups)
2. [Base Point 2] Process wind rdf data, use any data you like
3. [Extra credits, Optional, but point will be deducted in backtest and performance part if not included] Generate a stock universe on a daily basis, clarify the reason of such choice in presentation

Model

Do whatever you like, you can use your final predicted return as final signal directly.

Backtest

There will be a basic score in backtest part, in backtest section, you are required to realized the following functions:

Note: code in this part will not be reviewed by TA (but you need to submit code), you should include your efforts in your presentation slides!

Suppose the whole stock universe (i.e. the whole A share market) is S, you are asked to realize the following tasks:

1. Every day, in your defined stock universe S', which is a subset of S and is time-variant, calculate the **cumsum** net value of 10 groups of portfolios (according to their value rank, a line chart) as well as their final cumsum net value(a bar chart)
2. Plot your signal's Rank IC(defined in the next section) time series in a MA5(5 days rolling mean) manner as well as a histogram plot with text annotating mean and variance
3. [Optional] Plot the stability of your signal, which is defined as the ρ of AR(1) process, plot in a MA5 manner
4. [Optional] Plot your signal's correlation with well-known risk factors, here, it means the log(market cap), plot in a MA5 manner

In-sample & Out-of-sample performance

The factor will be evaluated from 3 aspects:

1. Rank IC, suppose on a cross section, your signal in the stock universe is $X_t \in R^{N\times 1}$, the return label is $R_t \in R^{N\times 1}$, the rank IC is calculated as $\rho = 1 - \frac{6\sum_{i=1}^{N}(Rank(X_{i,t}) - Rank(R_{i,t}))^2}{N(N^2-1)}$
2. Standard deviation of IC, or ICIR = mean(IC)/std(IC)
3. Turnover, given an ideal weight matrix $W \in R^{T\times N}$, the each row is $w_t \in R^{1\times N}$, the mean turnover is defined as

$$\frac{\sum_{t=2}^{T}\sum_{i=1}^{N}|w_{i,t}-w_{i,t-1}|}{T-1}$$

Note: Since it is a class project, you are not required to perform well in all 3 aspects to obtain a high performance score, e.g. an outstanding performance in ICIR (which is criteria 1 & 2) is sufficient to obtain all points in performance section.

Code Submission Guideline

1. All code should be submitted, exclude the model part, you can only submitted trained computing graph. Your codes should be zipped into **ONE zip file**, with name **2021.M2.BigDataAnalysis.GroupX.code**
2. You must separate your data preprocessing progress and your analysis progress. A document should be attached on how data is preprocessed, all preprocessed data must be stored either on local machine or HDFS and should be directly accessed by your signal generation process. **Note: user should be able to gain new results of data preprocess through changing start data and end date.**
3. You should include a file named with `main_signal` to access all file preprocessed. the `main_signal` should generate a signal file called `signal_{start_date}_{end_date}.csv` of the following format

tradeDate	ticker	signal	group	weight
str, YYYYmmdd(or matlab datenum)	str, XXXXXX.SH/.SZ	float	int	float