Description of Assignment C

Yixiang Wang

The assignment is about to use tweet data to predict the price change of chosen stocks. It has been widely accepted that data from social media can be used to predict the economical trend and applied to event driven algorithmic trading. Though researches have suggested that social media tends to reveal information fasten than traditional media, it also has its limitations in terms of using as trading signals. Part one will analyze the difficulties of finding alpha from purely social media data. Part two discusses the approach of determine the price movement in detail. And finally part three discusses possible improvement and future works

PART ONE: Analysis

Compared to traditional media, the unique advantage of social media is the fact that many news nowadays are released on them first. It is probably the fastest possible way to publish news. It is not unusual that those traditional media will choose to release digest of news on their social media's account first. However, some limitations do exist when it comes to derive trading signals from tweets data:

1) Though valuable and useful information does exist in the tweets, most of them are irrelevant to markets. Majority use tweeter as a tool to record and share their own life, which contains very little information about economic trends (though some meaningful conclusion about macro economy can be found from the huge collections of all the tweets, it is usually not accurate enough to be used as trading signals)
2) The volume. The huge amount of tweets generated every day makes it difficult to keep up. Since the market is at least close to efficient. The alpha will usually exist for a very short of time (note that the order book itself is a very efficient mechanism of information finding). In order to match the half-life of tweets as trading alpha, big data tools are needed in practice. In the case of this assignment, Apache Spark on AWS is our big data operating system.

PART TWO: Approach

In order to develop a proper algorithm to find alphas from tweets, I began with reading some papers about applying data mining algorithms in trading. I found the following paper interesting and helpful:

https://www.cis.upenn.edu/~mkearns/papers/KearnsNevmyvakaHFTRiskBooks.pdf

This paper discussed how to apply machine learning in three trading fields:

1) Optimal execution
2) Predict the pricing movement
3) Order distribution in dark pools

The second part of the paper outlines a simple and practical method to simplify the problem:

In order to determine how to trade, the minimum information we need is in what direction will the price the move. As long as we can obtain this information faster than others, we can take the advantage of it by simply take the correct position, wait for the expected price movement happens, and close the position. Thus my algorithm will take in a data set of tweets of a given period of time, determine more positive or negative information of all the stocks in DJ30 index exist in the data set, and predict the direction of movement of the components. For each tweet, I count the number of "good" and "bad" words (as suggested by Andrew) to determine it is positive or negative.

PART THREE: Possible improvement

As discussed earlier, the signal/noise ration of tweets is very low, thus it makes sense and will probably improve the performance by incorporating other economical factors.