BERNARD BARUCH COLLEGE

MTH 4330 INTRO TO ML (FALL 2024)

TAKE HOME FINAL - DRAFT

1. Disclaimer

This is a Draft. Please take a look at it and let me know in class if you have questions about it. I will then update it with the questions/answers we will discuss in class.

2. Description of the project

The objective of this Final is to simulate the revision process that a paper need to normally undergo when submitted to a peer-reviewed journal. You can find the paper, recently appeared on the arXiv at the link https://arxiv.org/pdf/2402.17194, regarding stock trend forecast using random forest. There are no right or wrong anwers, I will pay particular attention to the way you justify your procedures and how clearly you describe them. The output should not have more than 10 pages on a pdf file. The file should not be just an export of a Jupyter notebook and it should look like a professional scientific paper. The style should be concise, describe what is the goal of each section, what you intend to do to achieve the goal and why you want to achieve the goal. Every figure should be referenced in the text, have a title, lables on the axis and a short caption with a description of what is displayed in the figure. Each metric or feature that you are using for the model should be clearly defined mathematically (exceptions are allowed for quantities we defined in class like the least squared loss function).

The project is divided in 3 parts. The goal of the first part is to fill out the missing details from the paper. The paper is not really convincing from many point of views: other papers are cited without references in the text, there is not a clear definition of the features that are used and not a reference on how the finetuning of the model at hand (in this case random forest and SVD among others) is performed. In this first part you should clearly define the indicators that are mentioned in the Table 1 of the paper (a good, starting point is investopedia https://www.investopedia.com/top-7-technical-analysis-tools-4773275) and summerize what the goal of the paper is and describe the procedure that the authors intend to use to achieve the goal.

The second part regards checking the results presented in the paper. Can they be reproduced? Can you reproduce the figures appearing in the paper if you train the model on a similar dataset (e.g. regarding a stock they are using with roughly the same time period)? This part of the project is hard because you are essentially filling out the gaps of a procedure we are not even sure is working. You can focus to reproduce only the results that are relative to the use of random forest. You should use python and you are allowed to use the scikit library and the xgboost library (https://xgboost.readthedocs.io/en/stable/) . Your code should be included in the submission on a separte file(s). Contrary to what is done in the paper you are reviewing, you should clearly describe any procedure (e.g. how you are tuning the model or training it) you are using and motivate your choice.

The third part can be used for two purposes: in case your results are far from the ones presented in the paper you can discuss the reason for this discrepancy and propose further checks highlighting why you think the results presented in the paper are not reliable. On the other hand, if you find the paper reliable, can you improve its results?

Useful questions you can ask while working on each of the 3 parts and some additional hints are listed in the following subsections.

Part one.

- What is the goal of the algorithm presented in the paper?
- Which ML algorithm am I going to use (e.g. random forest) to verify the results of the paper?
- How am I going to test the outcome of these algorithms? Keep in mind what we said in class on using cross validation on times series.
- What is the formula to be used in the code to extract the label and the features described in the paper?

Part two.

- Once you specified the missing steps in part one and wrote a first draft of the code you are going to use to reproduce the results in the paper, make sure to use your code on a simpler setting (e.g. the two moon data set or a simple synthetic time series you create playing the role of the Oracle). The goal is to see if your own code is reliable when used as a classifier before testing it on the data described in the paper. You should dedicate a section in your write up to this end.
- Set up a pipe line to dowload the data. You can do that using the yahoo finance python library (`https://pypi.org/project/yfinance/`).
- The authors of the paper do not seem to specifiy the training set except for saying that consists in 7000 trading days. How does your analysis change as a function of the time period chosed for these 7000 training days?
- The result of this section should be a python program (possibly a jupyter notebook) with clear comments that I can run on my own laptop. You can break up the code to multiple files. I suggest referring each file to a section of the paper you are trying to reproduce the result of (the file should start with "This code is used to reproduce the results in section XX).

Part three.

- The code you are using for this part of the project should be contained in a separate python file called Part3.
- Remeber you are acting as a reviewer of a paper in a scientific setting. If you think that this paper is a scam you can not just reject the paper saying that you were not able to reproduce the results. You should find a way to convince the editor that the results presented can not possibly be true.
- If the results in the paper seem admissible, but you were not able to reproduce them with the same quality is there something you can do that is not described in the paper in order to improbe these results?