ISYE 6767 Design and Implementation of Computational Finance Models Fall 2018
Professor S.J. Deng

Midterm Project 2: Data Analysis and Machine Learning

This project is to predict stock price returns in financial markets based on the application of machine learning models. The focus will be geared towards the basic steps that are required for carrying out a data analysis project and implementation of the machine learning models, followed by analysis of their advantages/disadvantages.

DATA

Collect data for the 10 stock tickers as given in the tickers.csv file from the time period 2000-01-01 to 2018-01-10 from Quandl.

CODING

Here is a description of the steps that need to be followed (it forms a general framework for other projects too);

1. Filling and normalizing the data: After observing the data you may want to get rid of the outliers and missing data by filling them (suggested reading: https://pandas.pydata.org/pandas-docs/stable/missing_data.html). Food for thought: Should you use forward fill or back fill? **Your code should have the functionality to deal with missing data.** Also many of the ML models require the data to be standardized, so **normalize all your data**.
   (Suggested reading: https://medium.com/@rrfd/standardize-or-normalize-examples-in-python-e3f174b65dfc)
2. Variables: determine the variable that is to be predicted, and the explanatory variables (or, features) to be used. The variable you want to predict here would be the returns of the closing prices and the most basic features you should be using are price moving averages. But **it is highly encouraged to use your creativity to come up with additional features.** (Suggested reading: https://www.marketwatch.com/story/use-these-market-indicators-to-predict-stock-moves-2011-02-21)
3. Choosing the model: choose any 2 machine learning models you have learned from the lectures for completing this project and choose some basic hyperparameters for running the model. You can use two of the following models: Decision Trees, Random Forests, SVMs, KNNs and Logistic Regression. The only requirement is that the model must have a **binary output**, i.e predicting whether stock price will rise or fall in the next time period.
4. Modeling training: fit your model on the training dataset, and fine tune your parameters to choose the best set of model parameters. Use 60% of the data samples for model training.
5. Test model performance with out-of-sample data: test the performance of fitted models on test datasets using evaluation metrics like accuracy, precision, ROC, loss, etc.
   (Suggested reading: https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/)
6. Use a larger universe of stocks (the stocks given in the ticker sets in Homework 4) to obtain 20 stocks which are best predicted by your model over the time period corresponding to the 40% data samples for out-of-sample testing.

ISYE 6767 Design and Implementation of Computational Finance Models Fall 2018

Professor S.J. Deng

GRADING RUBRIC

PROGRAMMING (50%)

1. (10%)- Preprocessing the data- extracting, filling, normalization and splitting of data
2. (20%)- Choice of features, model and parameter selections
3. (20%)- Training and Running the model for obtaining predictions of the data and finally obtaining out of sample evaluation metrics

REPORT (50%)

Write a report documenting your process and findings from the project. The essential points to cover are;

1. (5%) Data- How did you go about collecting the required data, and did you follow the steps highlighted above to clean the data? Describe any of the data issues which you encountered and how you resolved them.
2. (5%)- Discuss the features used and the effectiveness of the features in getting prediction results.
3. (10%) Model- A note on why you used the models you chose, why you chose the hyperparameters you chose and your starting expectations from running these models.
4. (15%) Conclusion- Discuss the effectiveness of your models. And then compare the results of the two models you used, and their advantages/disadvantages.
5. (15%) Additional Considerations;
    a) Did your model have a high accuracy? If so, can you start trading on it? If not, what can be done to make your analysis more realistic?
    b) Is your model possibly overfitted? What steps did you take to mitigate the possible overfitting? (Suggested video: https://www.youtube.com/watch?v=mfzHchd5La8)