Final Projects

(Deadline: 8/11/2017 – no exceptions)

Requirements for the Final Project Report

You need to write a final project report in a conference paper style with related citations. You can choose any programming language while using any machine learning packages to help you achieve the goal.

In your report, you need to give a detailed explanation of each step you take to arrive at your solution. Please give a justification or explanation of the results you obtain.

The reports should also include the lessons you may have learned from doing the project. We will provide an anonymous ranking order of the accuracy each team is able to achieve.

Grading guidelines

Here is the grade distribution of each report. Please note that this is for group grading (with a maximum number of 3 students per group. All group members will get the same grade for the group contribution, but each will get their corresponding contribution grade as well. Details are given below.

- 20% for the amount of efforts from the team to tackle the problem (including the clarity of the reports). The more efforts as demonstrated in the report, the higher the grade for this portion. One way to demonstrate the efforts is by looking the number of machine learning techniques you group have explored for solving the problem. (Group contribution.)
- 10% for the overall quality of the results and the validity of your code submission. The higher the quality, the higher the grade for this portion. (Group contribution.)
- 5% for the report that includes a section clearly indicate the percentage of contribution of each team member. If not clearly specified, the team won't get this 5% grade. Please see lecture 1's notes for guidance. (Group contributions.)
- 5% for the report that includes a section to discuss how well the team worked together. If not clearly specified, the team won't get this 5% grade. Please see lecture 1's notes for guidance. (Group contributions.)
- 40% for each individual team member's contribution. If not clearly separable (either not specified or stated so), all team members will get the same grade for this portion. Each individual team member need to write a separate section about their own contribution to the portion they are responsible, demonstrating the amount of efforts they have contributed to the project, and the overall quality of your individual contribution. The higher the quality (the efforts), the higher the grade for each group member. Another consideration is the thoroughness of the approach they have considered. (Individual contributions)

- 20% for the quality of the results in comparison with other teams. The grade will be divided among the teams in 5 buckets: the top 20%-tile get the 20 points, the next 20%-tile get 16 points, etc.

Final project: Sentiment Analysis and Corn Price Movement

Background

Financial analysts, traders and market professionals globally are increasingly using Twitter to stay abreast of the market and make critical decisions.

Sentiment analysis can use natural language processing, artificial intelligence, text analysis and computational linguistics to identify the attitude of a writer with respect to a topic.

Currently Bloomberg provides real-time sentiment score of each equity company. You can find a Bloomberg terminal in one of the selected campus libraries (http://guides.library.columbia.edu/bloomberg), find an interested company, for example ***IBM US EQUITY***, type in ***FLDS***, then type ***sentiment***. This would give you a real time sentiment score of the selected company. JP Morgan also create a sentiment index of SPY 500 based on real-time tweets. You can find this index in Bloomberg by searching for ***JPUSISEN***.

Goal & Tasks

(1) In this project, you will be a hedge fund researcher working on the provided Corn tweets dataset. You're required to use one of the sentiment analysis tools (for example, Stanford CoreNLP Sentiment) to analysis each tweets' sentiment, and then use this score to build machine learning models to predict the future's Corn prices (regression) and price movement signals, Up, Down or Neutral (classification).

(2) You can play with different parameters to decide how to define a price signal is Up, Down or Neutral. For example, you can define a signal is Neutral if the future price fluctuation (either increase or decrease) is within a small range (either in the absolute value measurer or the percentage value measure).

(3) You can also play with the parameter to decide for how long into the future you would like the model to predict, for example, the next day's price or the next few day's price. You're required to report various parameters you have explored in building your machine learning models.

(4) You're also required to explore different machine learning models for both the regression problem (such as linear and nonlinear regression models) and the classification problems (such as Adaline, logistic regression, SVM, ANN etc). Please compare different ML models and their pros and cons in your analysis.

(5) Suppose each team is given $10,000 cash as your initial fund, please calculate an annual return of investment based on your predicated signals (assuming there is no transaction cost for simplicity) by using a very simple all long (buying) or all short (sell) strategy. For example, if your ML model predicts that the corn price would go up tomorrow (assuming you're only looking into the next day's trading), and the corn price today is P_t , then your position would be all long (i.e., buying), i.e., by buying all the shares with your available cash (such as at $\frac{10{,}000}{Pt}$ assuming your initial cash is $10,000. Should you decide to sell all of your shares at the next price point, your profits would be $(P_{t+1}\text{-}P_t)\times\frac{10{,}000}{Pt}$. Similarly, if your model predicts the price would go down,

then you'd take a short position and sell all of your positions at today's priceP_t. If your model predicts the price would be neutral, you will do nothing to your positions.

Annualized return calculation can be found in this link: http://www.investopedia.com/terms/a/annual-return.asp. Please denote your annualized return result in %.

(6) Explore the time varying effect. Something popular 10 days ago in social media might have an impact on today's price movement, so it's a good idea to explore time varying effects. You can achieve this by create a new variable, for example, 10 days' sentiment moving average, and add it to your predicting model.

Data

(1) Twitter dataset, this dataset contains 25,695 tweets, from 2008 to Jun 2017. ***Date*** is in GMC time zone. Those data are collected using web-scrapping technique from twitter website.

(2) Corn price dataset, this dataset contains data from 2008-05-15 to 2017-06-14. The price you'll use is the ***Close price.*** For the classification problem, please convert this price to the Up, Down or Neutral signals based on your choice of threshold parameters for defining them. To do this, you'll first calculate the relative price change, and then depending on that value, you will convert the price signal to Up / Down or Neutral.

Some suggestions:

(1) Twitter data is noisy, you'll need to clean the ***text*** before you put it into sentiment analysis.

(2) Tweets in 2008 and 2009 may be very sparse, which means in some certain days, you do not have tweets related to corn. This is because Twitter start from 2006. For those initial years, you may drop them or figure out a method to impute those missing values.

(3) Stanford CoreNLP can be downloaded at https://stanfordnlp.github.io/CoreNLP/. You'll need to install Java and Python/R wrapper to use it.

Here is a live demo to play this tool, http://nlp.stanford.edu:8080/sentiment/rntnDemo.html

(4) Steps you might take in this project: Cleaning data; Choose sentiment analysis tool; Connect sentiment analysis result with corn price (build model); Predict corn price based on your model; Evaluate your model performance.

Deliverables

Your final results should include at least two parts:

(1) All machine learning techniques you have explored, and the FINAL recommended one with justification and explain why you recommend that one. Even if you are using some existing sentiment analysis tools, you can discuss the advantages of the tools you choose and the machine learning models behind it.

Note that, the most accurate one may not necessarily be the best one – in practice, many factors may determine which machine learning techniques get chosen for deployment. Some considerations include

the training complexity, the deployment complexity, and the difficulty of explaining the classification results with their features.

(2) The quality metric for each machine learning technique you have explored. We define the quality metric as follows:

- Annualized return (in percent) based on your models.
- A percentage number for the training data (out of 284807 data points). Denoted this as P1
- AUPRC of training (denoted as AUPRC TRAIN) and testing set (denoted as AUPRC TEST). See this link for an example:

https://stats.stackexchange.com/questions/10501/calculating-aupr-in-r

In your final report, please clearly indicate your achieved AUPRC for each machine learning technique you have explored. Please round the results to 4 decimal points if needed. Please put all the results into a table at the beginning of your report with the following format:

ML Technique explored	P1	Annualized return in training	Annualized return in testing	AUPRC in training	AUPRC TEST in testing	Corresponding Section # as discussed in the report
Method 1						
Method 2…						

Miscellaneous

Please feel free to use any machine learning package implementation or modify the existing implementation to achieve your own goals. The goal is to make yourself comfortable to explore various machine learning techniques to solve a specific problem.