Machine Learning & Data Mining
Miniproject 1

Caltech CS/CNS/EE 155
Released February 7, 2020

---

1 Overview

In miniproject 1, you will be using 500ms aggregate "high" frequency order book data of a futures contract to predict future 1-second price changes.

The miniproject consists of two separate competitions on Kaggle, a site for data science competitions.

You are to use the training data (`train.csv`) to come up with predictions for the test data (`test.csv`). There will be a public leaderboard that will show your performance, but it only consists of approx. half of the test set (and you don't know which half). The private leaderboard ranking with the other half of the test set will only be revealed when the whole competition ends. The competition will end on Thursday, February 13th at 2:00 PM PST.

The links to the competitions, which includes the datasets and guidebooks describing the datasets, can be found here:

- https://www.kaggle.com/c/caltech-cs155-2020/

There are already several benchmarks added by the TAs in grey. You should try to beat (some) of these benchmarks, but you certainly aren't expected to. The highest benchmark should take significant effort to beat, but is still far from impossible.

Your task:

For a more detailed description of the terminology and data, please see the actual Kaggle Data tab.

The last column in `train.csv` is "`y`". Each row of data represents one snapshot in time. For the label "y", an entry of "1" means the "mid" price will go up (strictly up), and "0" means the mid price will either not move or go down. You should predict the relative probability of a "1" at each timestep of the test set.

Your task is then to predict `y` using the data in `test.csv`.

Please note that your submission should be **the probabilities** of each testing sample being `1`, instead of binary 0/1. Please follow the format in the sample submission files (`sample_submission.csv` when generating your submissions to Kaggle.

Performance metric:

The metric on which your model performance is tested is AUC, namely, the Area Under the receiver operating characteristic Curve. Note that this means your probabilities do not have to be exactly calibrated, but you can calibrate them if blending models.

2 Key Notes

- The competition ends on Thursday, February 13th at 2:00 PM PST.
- The report is due on Tuesday, February 18th at 9 PM PST, via Gradescope. See below for the report guidelines. The report should explain your process and results in a thorough manner.
- You can work in groups of up to three people. **DO NOT MAKE SUBMISSIONS UNTIL YOUR TEAM IS FINALIZED ON KAGGLE. ONCE TWO PEOPLE HAVE MADE INDIVIDUAL SUBMISSIONS THEY CANNOT MERGE AS A TEAM.**

- Your team can make up to 12 submissions a day. However, at the end, you need to select the 2 submissions that you think will perform the best on the private test sets for both competitions.
- If you have questions, please ask on Piazza! As with any Kaggle competition, it's best to get started early since you are only allowed to make 12 submissions a day.
- You can use any open-source tools and Python, using both concepts you learned in class as well as any other techniques you find online, to get the best score that you can.
- **You may collaborate fully within your miniproject team, but no collaboration is allowed between teams.**
- **You may not search for additional data related to this task; you may only train your models using the provided training set.**

3 Report Guidelines

- **Due date:** Tuesday, February 18th at 9 PM PST
- **Format:** The report should be written exactly to the length specifications given in the template. If a section of your report is too short for that section in the template, use pagebreaks as necessary. In other words, if your discussion of your approach is less than two pages, use pagebreaks to ensure that exactly two pages are used for the Approach section. If a section of your report is too long for that section, please try to be more concise. Only include code if necessary - your code should not be a significant portion of the report. We recommend a link to a GitHub repo. You should use graphs in your report, as visualization is very helpful!
- **Please submit your report in groups rather than submitting it once per student!** You can see how to submit in groups here:
  https://www.gradescope.com/help#help-center-item-student-group-members

We highly recommend that you use the LaTeX template provided to you and simply fill in the blanks. To collaborate on the report writing, we recommend using Overleaf (https://www.overleaf.com/edu/caltech), an online LaTeX editor. Caltech students can get a pro account for free using caltech.edu emails.

See our example file for guidelines. The structure is as follows:

1. **Introduction (20 points):** This section is purely for the TAs and should be brief.
   - Group members
   - Team name (needs to match your team name on Kaggle)
   - What place you got on the private leaderboard for both competitions.
   - What AUC score you got on the private leaderboard for both competitions.
   - Division of labor: Your team must ensure that each member has an equal amount of workload during the competition. If there is a noticeable discrepancy in the division of labor, team members may receive differing grades.

2. **Overview (20 points):** This section should be a concise summary of your attempts. More detailed explanations should go in the next section.

   - Models and techniques tried: What models did you try? What techniques did you use along with your models? Did you implement anything out of the ordinary?

     Descriptions should be concise, at most 1-2 sentences. Again, more details can be included in the next section. However, this section is meant to be a more general overview.
   - Work timeline: What did your timeline look like for the competition?

3. **Approach (20 points):** This section should be a more detailed explanation of how you approached the competition.

   - Data processing and manipulation: Did you manipulate the data or the features in any way? What techniques and libraries did you use to accomplish such manipulation? Please justify your methodologies.
   - Details of models and techniques: Why did you try the models and techniques that you used? What was your process of using them? What are the advantages and disadvantages of using such methods?

4. **Model Selection (20 points):** This section should outline how you chose the best models.

   - Scoring: What optimization objectives did you use, and why? How did you score your models, and why? Which models scored the best?
   - Validation and test: Did you use validation techniques? How did you test your models? What were the results of these tests, and what did the results tell you? Preferably, use computer-drawn diagram to illustrate how you split the data (if at all).

5. **Conclusion (20 points):** This section should be used to summarize the report, as well as to include any additional details.

   - Insights: Please answer the following questions
     - Among all the features in the data, which features have the most influence on the prediction target? Why? List top 10 features. (Bonus points if you can analyze whether these 10 features positively or negatively influence the prediction target.)
     - Why do we use AUC as our Kaggle competition metric? Do you think there is a better metric for this project? Why, or why not?
     - Among the machine learning methods/pipelines that your group uses, are there any methods/pipelines that are parallelizable? https://en.wikipedia.org/wiki/Parallel_algorithm If so, how can they be parallelized? If not, why not? (Conceptual descriptions are good enough. Also, you are not required to actually parallelize your codes. These questions just help promote better understanding of the methods you have learned.)
     - Did you learn anything new from this project outside of the lectures and homework?
     - Overall, what did you learn from this project?

- Challenges: What could you have done differently? What obstacles did you encounter during the process?
- Concluding remarks (optional): Anything else you'd like to mention?

4 Grading metrics

On the two competitions, you will be scored on the test set. You will see results of the public leaderboard (results of your model on half of the test set) for the duration of the competition, and the private leaderboard results (results on the other half of the test set) will be released after the deadline.

The report is worth the majority of your grade. That is, we care much more about the process and thoughts behind your results rather than the scores.