Machine Learning & Data Mining
Miniproject 1

Caltech CS/CNS/EE 155
Released January 28^{th}, 2017

TAs: Sarthak Sahu, Andrew Kang

Overview

In miniproject 1, you will use demographic data from 2008 to predict voter turnout in both 2008 and 2012. The miniproject consists of two separate competitions on Kaggle, a site for data science competitions.

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

https://inclass.kaggle.com/c/caltech-cs-155-2017-part-1
https://inclass.kaggle.com/c/caltech-cs-155-2017-part-2

In the first competition, you are to use the 2008 training data (train_2008.csv) to come up with predictions for 2008 test data (test_2008.csv). There will be a public leaderboard that will show your performance, but it only consists of half of the test set. The private leaderboard with the other half of the test set will be revealed in class on the due date. The competition will end on Thursday, February 9th at 1:59 PM PST.

In the second competition, you will use the same model you trained from part 1 (using 2008 training data) to generate predictions for voter turnout in the 2012 election. There is no public leaderboard for the second competition. Results will also be revealed in class on Thursday.

Key Notes

- The competitions end on Thursday, February 9th at 1:59 PM PST.
- The report is due the following Monday, February 13th at 9 PM PST, via Moodle. 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, but must make submissions from a single account.
- You can make up to 5 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 5 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.

Report Guidelines

- **Due date:** Monday, February 13th at 9 PM PST
- **Format:** The report should be 5-10 pages long in single column format. 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!

We highly recommend that you use the LaTeX file provided to you and simply fill in the blanks. See our example file for guidelines. The structure is as follows:

1. **Introduction:** This section is purely for the TAs and should be brief.
   - Group members
   - Team name
   - Division of Labour: 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 labour, team members may receive differing grades.
2. **Overview:** 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:** 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?
   - Details of models and techniques: Why did you try the models and techniques that you used? What was your process of using them? What were your scores? What are the advantages and disadvantages of using such methods?
4. **Model Selection:** This section should outline how you chose the best models.
   - Scoring: How did you score your models? Which models scored the best?
   - Validation and Test: Did you use validation techniques? How were the bias and variance of your models?
5. **Conclusion:** This section should be used to summarize the report, as well as to include any last minute details.
   - Discoveries: What did you learn from this competition? Did you learn anything new outside of lecture?
   - Challenges: What could you have done differently? What obstacles did you encounter during the process?
   - Concluding remarks: Anything else you'd like to mention?
6. **Appendix (optional):** Use this section for anything else you'd like to include. Don't include this section if the above sections have covered everything.

Grading metrics

On the two competitions, you will be scored on the test sets of 2008 and 2012. For the 2008 competition, 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 leaderboards have equal weighting. For the 2012 competition, there is only a private leaderboard.

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