Advanced Computational Methods: Classification Competition

Overview and Objectives

The objective of the competition is to provide you with a more structured playground for appling your newly acquired knowledge in practice together with a healthy dose of competition. The competition will take place at Kaggle In Class. Kaggle is a popular platform for machine learning competitions. Many companies use it to pose challenges where you can obtain sizable rewards or jobs. Kaggle also has many useful educational materials with numerous practical tips & tricks. Check their forums!

Groups

The computing project is carried out in groups of 3 people that you will form by yourself. By the end of the day (Friday, January 15, 23:59) you should let me know the team composition. One person from the team should send me an email with the team name and three members. You should use the same team name at Kaggle website. If some teams are not formed by the end of the day, I will assign members randomly. I will create a list of teams and members and upload it to the `competition` folder on Box.

The competition

The competition at **Kaggle** will be launched soon and you will receive invitations as soon as administrators approve the competition. Note that you should use your BGSE email accounts for registering.

At the end I opted for a dataset that can be found at UCI repository.[1] It is not the most exciting dataset out there, but I had to choose relatively clear dataset so you do not have to work a lot on feature selection/transformation, since we will not cover such topics in our classes.

You are free to use any classification method, regardless of whether it was used in class or not. Code will have to be implemented in R. There should be no code sharing between teams. To maximize the learning outcomes and to give you greater incentive to produce good code, after the competition is finished and grades delivered, the code used by each team to produce predictions (see points 3 and 4 below) will be made available to the whole class at Box.

Timeline

Interim Evaluation - Week 6, Feb 12, at 12:00

In the middle of the semester there will be an interim evaluation. You will be graded on two factors, both on scale from 0 to 10.

[1] archive.ics.uci.edu/ml/datasets/Online+News+Popularity

1. **Prediction performance** - At this date the performance of each team will be recorded according to the Kaggle In Class leaderboard. The best team will get 10 points, while other teams will get progressively smaller amounts. Grade 0 is reserved for performance at chance level or below. This will make 10% of the final competition grade.

2. **Code used for predictions** - You will have to develop an R package that implements the algorithm that produced your best predictions that were submitted on Kaggle website. This will have to be done on Github and state of the repository found on this date will be counted as a submission. This will include the procedure of training your final classifier, based on the training dataset posted on Kaggle and how it produced the predictions for the test set. Whatever manipulations you do with the initial training dataset to get to the one you finally used, it has to be done from R. I will grade several aspects. First is reproducibility, I should be able to produce the same predictions that you submitted to Kaggle if I give it the same input. Second is how well documented the package is and how easy it is to use, and third, how well the code is written. This will make 10% of the final competition grade.

Final Evaluation - Week 10, March 11, at 12:00

At the end of the semester the competition will end and the teams will be evaluated again. This will make **75%** of the competition grade. You will be graded on several factors, all of them on scale from 0 to 10 again.

1. **Prediction performance** of the solution. This part will be completed on Kaggle In Class website. This will make **35%** of the competition grade. The best team will get full 35%, while other teams will get progressively smaller amounts.

2. **Report** showing the algorithms you have considered and tried out. We are interested in accuracy of your algorithm in absolute terms, but also relative, show us it is better than the alternatives you have tried out. Here you should also provide arguments for choosing one method over another, describe how you have chosen the parameters of your classifier and discuss advantages and limitations of your classifier. The report does not have to be extensive, but it has to have the elements mentioned above. I am open-minded about the format, it can be a PDF, a web page, or something else entirely. This will make **20%** of the competition grade.

3. **Code** used for generating predictions of your winning solution on Kaggle will make another **15%** of the grade. You will add the code of your your best algorithm to the R package that you used at the interim evaluation. I will grade the same aspects as described in the interim evaluation.
4. Development of all the documents and the code related to the competition should be **under version control using Github**. This will make **10%** of the grade. It can be either a private or a public repository. At the end of the competition you will grant me access to the repository and I will check how well you have used it to organize your work, produce the reports and the code for the competition. Bare minimum will earn you 0 points. Single person usage with few commits will yield 5 points. More proper usage with all members actively collaborating will result in full points. There are couple of ways to collaborate using Github, you will have to investigate it and choose one of them.

Project Presentations - March, Time TBA

Three best teams in the final ladder of the competition will carry out a 15 minute presentation of their winning strategy.

Evaluation summary

- 10% of the course grade is based on the prediction accuracy at the interim evaluation.
- 10% of the course grade is based on reproducibility and understandability of the code at the interim evaluation.
- 35% of the course grade is based on the prediction accuracy at the final evaluation.
- 15% is for reproducibility and understandability of your code at the final evaluation.
- 10% is for development under version control.
- 20% is for the report.

Special note

Final prediction performance achieved in the classification competition will count partly in the grade for the Machine learning course.