IE 582 Fall 2020 Project, Due final exam

Objective and Evaluation

Use the data provided with this project on Moodle.

One file is the **training data** that you should use for model building.

A second file contains **testing data** that will be used to evaluate your model.

The training and testing data were selected randomly from the original data set.

The aim is to build a classification model for the given data based on the methods described in the course. You are also free to use extended versions of the approaches covered in the lectures.

Performance measure:

Balanced Error Rate: Your model will be scored as your error rate on class "a" plus your error rate on class "b". This is different from the overall (weighted) error rate commonly reported. Because the classes are unbalanced, you should work to minimize the error rate on each class.

Area under the ROC curve: Although correlated with balanced error rate, area under the ROC curve gives better idea about the performance for binary classification problems.

Your submission will be evaluated based on these two measures and you are expected to maximize both. This fits well to multi-objective optimization setting and your rank will be determined based on the non-dominated sorting idea. "A solution is called nondominated, Pareto optimal, Pareto efficient or noninferior, if none of the objective functions can be improved in value without degrading some of the other objective values." (source: https://en.wikipedia.org/wiki/Multi-objective_optimization)

Prediction and Report

This project is organized as a competition (like in platforms such as www.kaggle.com). The submission system will be active next week and the details of the system will be announced. You will be able to make submissions programmatically via an application programming interface (API) using your scripts. You will be able to make five submissions every day and observe your performance on a predefined subset of the test data. The final evaluations will be performed based on your latest submission. You will be keep posted about the deadlines for final submissions towards the end of the project period. Note that 30% of your project grade will be determined by your final rank in this competition. First place will get full points (30 points) and this will decrease to a minimum of 15 points proportional to your deviation from the top performer.

You are required to submit a written report with a brief description of your final method, how you evaluated your methods, and you choose the parameter settings. You are allowed to work as a group of at most 3 members.

Your report should have the following format:

1. *Introduction:* Problem description, summary of the proposed approach, descriptive analysis of the given data.
2. *Related literature:* Summarize relevant literature if there is any
3. *Approach:* Explain your approach to this problem.
4. *Results:* Provide your results and discussion.
5. *Conclusions and Future Work:* Summarize your findings and comments regarding your approach. What are possible extensions to have a better approach?
6. *Code:* Provide the Github link for your codes at the end of your report.