AI 539 Project: General Guidelines

For the AI 539 final projects present a major portion of the overall class grade. All projects are based on actual real-world problems, however, in cases needed, the data files are coded in a way that they would not reveal the actual real data information.

There are two projects that the student teams need to perform. These projects are:

- Gotham Cabs
- Botanist

Students should form teams of up to **two** members, and **each team is responsible to do both projects**.

Students can access the data files and a description of each in two individual folders, as shared on Ed earlier. Below I explain the content of each folder.

Gotham Cabs The file `Project_Description.pdf` contains an overview of the data file, the input features and the response variables. The training file for this project is the CSV file `Train.csv`. The test file will be made available the day before the competition. The file `TestFileTemplate.csv` is a small test file (without the response values) that you never use, and only gives you an idea about the format of the actual test data that will be shared with you before the competition day. The actual test files would contain many more samples (rows), and are going to be larger files. The file `TestFileTemplate.csv` allows you to test your models on a small data file and make sure you do not get into any formatting or compilation issues on the competition day.

Botanist The file `Project_Description.pdf` contains an overview of the data files and the response variables. The training data are provided in image format, where the `TrainFiles` folder contains the images and the `Botanist_Training_Set.csv` contains the labels for each image in that folder. The day before the competition, you would be given a data file `TestFileToComplete.csv` containing the test image file-names, along with a folder containing the test images. You would need to complete the file `TestFileToComplete.csv` with your predicted labels and return it to the instructor. **Important Note:** The labels that need to be used to complete the file `TestFileToComplete.csv` are the integers `1,2,...,38`. If you use any other labels, the program that calculates your accuracies will fail to compile your file.

Project Structure

Each team is responsible to hand in one full report of the projects that have been addressed (**One single report for all the team members, containing the details of the two projects**). The reports need to be complete and contain the following:

- The names of the team members, plus an optional name for your team!
- An abstract: which is less than 300 words, explaining the selected projects, the motivation behind the selection and a brief overview of the project outcomes.
- Chapeter 1: This chapter would contain the description of the first project.
- Chapeter 2: This chapter would contain the description of the second project.

It is strongly suggested to keep the total length of the report below 10 pages. **Do not copy-paste your code into the report**. The code needs to be submitted as an individual file. Each chapter of you report should contain the following:

- An introduction, briefly overviewing your insight about the problem, and the techniques you have considered.
- Problem Analysis: A section that contains your technical analysis, such as feature correlations, sensitivity of the response to the features, data histograms and data visualization using methods such as PCA, t-SNE or UMAP. You need to test different models and compare them and ultimately pick the best one as the one for the day of competition. If applicable, you may also consider various model selection techniques, plot cross validation curves and justify your model selection.

  When comparing various models, you can work with a fraction of the data. For example instead of fitting your models to all the 1,000,000 samples in the Gotham problem, you may consider fitting your models to a fraction of that (anything above 10% for the Gotham data is acceptable). However, the final model that you bring to the competition, needs to be trained with the entire data.
- Concluding Remarks: A summary of your understanding and data assessment, and an overview of the models you tried and the one you picked for the competition.

Competition and Presentation

On the competition day, each team will present a summary of their project to the class (as a series of powerpoint slides). Due to the remote nature of the presentations, **only one member of each team would share their screen**, however, both members need to present. **The time allocated for each team to present is 6 minutes, make sure not to exceed this time limit**.

A day before the competition (at 3PM), the actual test files will be shared with the teams and each team needs to email their predictions to the instructor. You are given a very limited time to submit the predictions (approximately 9 hours after the test files are shared). After each team presents their work, the instructor will announce the prediction accuracy of that team.

The accuracies of the teams are ranked and each team receives a ranking for each of the two project. The teams with the best total ranking will be announced as the winners (extra bonus credits will be rewarded to the winners of each project).

Important Dates

Please read this section carefully, mark your calendar, and make sure not to miss a deadline. All the times are Pacific Time zone (Oregon time):

- **Monday, May 20, at 3PM**: deadline to finalize the team members and list them on the shared excel file, available [here]. Please do not write anything in the presentation day/time column, that section will be assigned by the instructor
- **Tuesday, June 4, 3PM:** The final test files will be released.
- **Tuesday, June 4, 11:59 PM:** Deadline to submit the test predictions.
- **Wednesday, June 5:** Presentation of the final projects on the Zoom. The presentation time is during the class time
- **Saturday, June 8, at 11.59 PM**: Deadline to submit the project reports.