(FALL 2022) KAGGLE COMPETITION

INF8245E - MACHINE LEARNING

1 Background

For this project, you will take part in a Kaggle competition based on tabular data. The goal is to design a machine learning algorithm that, given information on a particular concertgoer experience, can automatically classify the enjoyment of that concertgoer to that concert. In this classification problem, we have 4 classes. The training dataset consists of 170,000 training examples and the testing dataset contains 30,000 test examples. Each training rows contains a unique ID, 18 attributes and 1 target containing the class that needs to be predicted. You will be evaluated on the test private leaderboard mean F1-Score.

First, you need to create an account on the Kaggle website, if you haven't already. Next, you can access the competition. We expect you to be working in groups of exactly 3.

2 Team Formation

To be able to form a team, follow the instructions below:

- Each team should consist of exactly 3 members;
- Fill out this google form with your team's information by Nov 8-th at 10PM, EST;
- Register as an individual Kaggle user, enter the competition and accept the terms and conditions.
- Go to Kaggle team
- In the Invite Others section, enter your teammates' names, or team name.
- Your teammate has the option to accept your merge. The person accepting a merge is the team leader.

Note on number of submissions: The maximum amount of submissions is 2 per day for the entire team. The test data will be released after the team formation deadline which is November 8-th EST. If you make any submissions before this, you might be disqualified from the competition. All the team members will receive same marks for this competition. It is your duty to make sure everyone has contributed to the competition equally.

3 Instructions

To participate in the competition, you must provide a list of predicted outputs for the instances on the Kaggle website. To solve the problem you are encouraged to use any classification methods you can think off, presented in the course or otherwise. Looking into creative way to create new features from those provided may prove especially usefull in this competition.

Note: We suggest you to start early, allowing yourself enough time to submit multiple times and get a sense of how well you are doing.

4 Report

In addition to your methods, you must write up a report that details the pre-processing, validation, algorithmic, and optimization techniques, as well as providing your Kaggle results that we compare them with. The report should contain the following sections and elements:

- Project title;
- Team name on Kaggle, as well as the list of team members, including full names, email and matricules.
- Feature design: Describe and justify your pre-processing methods, and how you designed and selected your features.
- Algorithms: Give an overview of the learning algorithms used without going into too much detail in the class notes (e.g. SVM derivation, etc.), unless you judged necessary.
- Methodology: Include any decisions about training/validation split, distribution choice for Naive Bayes, regularization strategy, any optimization tricks, setting hyper-parameters, etc.
- Results: Present a detailed analysis of your results, including graphs and tables as appropriate. This analysis should be broader than just the Kaggle result: include a short comparison of the most important hyper- parameters and all the methods you implemented.

- Discussion: Discuss the pros/cons of your approach methodology and suggest areas of future work.

- Statement of Contributions. Briefly describe the contributions of each team member towards each of the components of the project (e.g. defining the problem, developing the methodology, coding the solution, performing the data analysis, writing the report, etc.). At the end of the Statement of Contributions, add the following statement: ”We hereby state that all the work presented in this report is that of the authors.”

- References (optional).

- Appendix (optional). Here you can include additional results, more detail of the methods, etc.

The main text of the report should not exceed 6 pages. References and appendix can be in excess of the 6 pages. You should use the ICLR format. You can find the template in this link.

5 Submission Requirements

We are expect you to follow these rules:

- You must submit the code developed during the project. The code must be well documented. The code should include a README file containing instructions on how to run the code. Submit the code as an attachment (see Submission Instructions).

- Your submission should contain your Kaggle notebook named as ”final.ipynb” which would reproduce your predictions exactly. Make sure to fix the random seeds so that the generated predictions are exactly matching your submitted prediction file.

- The prediction file must be submitted online at the Kaggle website. Please make sure your submitted result file has the correct structure and format. You should submit your result in .csv format. More information about the correct structure and format could be found in Kaggle website (go to : Overview→ Evaluation).

- You must submit a written report according to the general layout described above.

6 Submission Instructions

For this project, you will submit the report and the code to Gradescope. Make sure we can directly run your notebook in Kaggle. We should be able to run your code without making any modifications. Your group report should be submitted to Gradescope. One submission per team is sufficient for both code and report. The competition ends on Decembers 2nd and the report is expected by December 3rd.

7 Late Submission Policy

Late submission policy is the same as default policy used for the other assignments.

8 Evaluation Criteria

Marks will be attributed based on 40% for performance on the private test set in the competition and 60% for the written report. For the competition, the performance grade will be calculated as follows: The top team, according to the score on the private test set, will receive 100%. If the team doesn't cross the basic baseline, entered by the instructor, will score 0%. All other grades will be calculated according to the interpolation of the private test set scores between those two extremes. For the written report, the evaluation criteria include:

- Technical soundness of the methodology (pre-processing, feature selection, validation, algorithms, optimization).
- Technical correctness of the description of the algorithms (may be validated with the submitted code).
- Meaningful analysis of final and intermediate results.
- Originality of the approach.
- Correct insights and analysis on the link between certain attributes and the target.
- Clarity of descriptions, plots, figures, tables.
- Organization and writing. Please use a spell-checker and don't underestimate the power of a well- written report.

Do note that the grading of the report will emphasize the rationale behind the pre-processing and optimization techniques. The code should be clear enough to reflect the logic articulated in the report. We are looking for a combination of insight and clarity when grading the reports.

9 Questions and Clarifications

For additional questions, please use Piazza, or ask questions during the TAs office hours.