C-M-005: Machine Learning Lab I '21-22

Lab. 5 (Project) (Given November 20, 2021; Due November 29, 2021)

Your answers must be entered in LMS (and Kaggle) by midnight of the day it is due. You must submit the code as well as a self-contained PDF which has the approach, an explanation of the implementation, the output as well as anything else asked by the question. Marks devoted to this Project are indicated in the "Syllabus" sheet that was provided the first day of class.

Your goal in this project is to produce a classifier to achieve the best possible classification (from amongst the set of classifiers that we have covered in ML1).

The problem is a classification problem which distinguishes between a signal process which produces supersymmetric particles and a background process which does not. The first 8 features are kinematic properties measured by the particle detectors in the accelerator. The last ten features are functions of the first 8 features; these are high-level features derived by physicists to help discriminate between the two classes. Class 0 represents the background and class 1 represents the signal.

You will post your results in Kaggle and submit your final code in LMS. So you will be able to see your performance relative to the performance of your peers. You may submit multiple times but only the alst submitted entry will be counted. The Kaggle link is,

`https://www.kaggle.com/c/plaksha22-cm005-project/`

Some salient points appear below.

1. The submitted code must be able to reproduce results obtained from the Kaggle submission. No restrictions on the amount of time taken for hyperparameter tuning, but once the best hyperparameters have been obtained, training the final model (or combination of models) using those obtained hyperparameter values should not exceed an hour. When documenting the code, ensure to include a final section listing the models and best hyperparameters found, along with the random seed used. This would be used to reproduce and validate the submission on Kaggle.

2. To ensure that everyone uses the same set of tools to solve the problem, there is a restriction in the libraries that can be used for the contest.

The following is the list of libraries allowed. You are free to use any functionality available in these libraries, Numpy, Scipy, Pandas, Matplotlib/Seaborn, Scikit Learn, Keras and TensorFlow

3. Ensure that you use 5-fold cross validation.

4. Make sure to have your name on the leaderboard exactly the same as that on Brightspace (your full name) to ensure that the submissions will be reflected correctly while grading. This can be changed in the "Team" tab on Kaggle.

5. Kaggle allows you to select a particular model amongst all your submissions as the final one to be considered. If not selected, the model that performed best on the public leaderboard will be automatically chosen. In case any of you have experimented with libraries not stated in the rules, make sure to select the right submission that follows all the rules to ensure the score to be considered valid.

6. Make use of the 20 submissions available per day to keep iterating on your models, checking their performance and improving them. The best submission will automatically be considered as the final submission, so no harm in submitting early and often on Kaggle.

7. After 11:59PM on the deadline day, the official leaderboards will freeze. You could still try new models and submit to see how it would have performed, however these submissions post deadline will not get saved on the leaderboard by Kaggle.

8. We would recommend you to not only do modelling, but also work on documenting the code for submission on Brightspace. Only the .ipynb notebook needs to be submitted on Brightspace. The documentation can be done within the jupyter notebook itself. It must contain brief details regarding the following, Exploratory Data Analysis, Pre-processing, Feature Engineering(newly derived or modified features) (not mandatory, only needed if it has been done), Modelling approaches taken, Final Model with hyperparameters

9. Discussions on the Kaggle website often have rich insights on different approaches to solve ML problems, so it will be advantageous to check

out some previous competitions and related discussions. Here is another resource (`https://www.coursera.org/learn/competitive-data-science`) that might be helpful.