BLG 454E Learning From Data
Spring – 2019

-Term Project-

Autism or autism spectrum disorder (ASD) is appeared in a person as repetitive behavioral patterns and often weaken their social interactions with other people. We hypothesize that ASD affects the similarity in morphology between brain regions. For this reason, we can use these features for autistic versus normal subjects classification task.

The term project consists of two stages, the Kaggle competition and the report.

1. Kaggle Competition

Description

We have created a private class competition on Kaggle. Please click the following link for the term project competition https://www.kaggle.com/t/80529fc6241245479bcc9f83f34b7ac2

In this challenge, we ask you to apply the tools of machine learning to predict which persons have ASD.

Dataset

Each sample represents a brain graph (also called connectome). The value of each feature represents the absolute Euclidean distance between the shapes of two brain regions (or nodes in the brain graph). For instance, if two regions A and B in the brain have identical shapes, then their distance d(A,B) will be equal to zero (i.e., d(A,B) = 0). The labels are 0 and 1 referring to "normal (not autism)" and "autism" respectively.

The data has been split into two groups: training set (train.csv) and test set (test.csv). The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the “ground truth”) for each sample. Your model will be based on “features”, and class. You can also use feature engineering to create new features. The test set should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each sample. It is your job to predict these outcomes. For each sample in the test set, use the model you trained to predict whether or not they have ASD.

Goal

Your job is to predict if a person have ASD or not. For each sample in the test set, you must predict a 0 or 1 value.

Submission Process

To see the performance of your model on test data, submit your predictions of test data to Kaggle in defined format. Kaggle will calculate and rank the submission scores using the public test data throughout the competition. These scores are publicly visible on public leaderboard. After the competition end, private test data is used to calculate final model performance. Private leaderboard is not released to users until the competition has been closed. Public leaderboard is calculated with 50% of the test data. The final results will be based on the other 50%, so the final standings may be different. Therefore, train your model as general as possible to avoid overfitting on train and public part of the test data.

Scoring Metric

In Kaggle, your submission is evaluated as the percentage of persons you correctly predict, which is known as "accuracy”.

Submission File Format

You should submit a csv file with exactly 60 entries plus a header row. The file should have exactly 2 columns:

1. ID: [1,…,80] sorted in any order
2. Predicted (contains your binary predictions: 1 for autism, 0 for not-autism)

```
ID,Predicted
1,0
2,0
3,1
…
78,1
79,0
80,1
```

Submission CSVs must have a header row consisting of ID and Predicted as in the sample submission. Using different column names causes a fail in submission process. ID column must include all ID values between [1, 80] in any order (doesn’t need to be sorted 1 to 80). Predicted column must consist of 0 or 1. If you use different labels in your code like {-1, 1}, you must convert it to {0, 1} for submission file.

PS: Your submission will show an error in cases: you have extra columns (beyond ID and Predicted), extra rows, ID column doesn’t consist of integers between [1,80], Predicted column includes value other than 0 or 1.

You can download the sample submission file (sampleSubmission.csv) on the Data page.

Rules

- Every student has to create a Kaggle account
- Form a team of **2 or 3 students**
- Individual submissions are **not allowed**. In such a case, send us an email so that we assign a random teammate.
- Team members **must** be students registered of the same class
- Team names should be in the following format: StudentID1 _StudentID2_StudentID3
- Submission format is explained and a sampleSubmission file (sampleSubmission.csv) is given in the competition webpage.
- You are allowed to use only **R, Matlab,** and **Python** programming languages for the implementation.
- Academic dishonesty including cheating, plagiarism, and direct copying is unacceptable. Note that your codes and reports will be checked with the plagiarism tools!

2. Report

Prepare a report in Latex/Word using provided IEEE Conference Paper template. Your report must **not exceed** 3 pages!

The report should consist of the following sections:

1. (**15 points) Introduction:** Mention about what and why you did in this project briefly. Give your final score and rank in the competition with your Kaggle name and team name.

2. **(30 points) Datasets:** Explain your methods for data preprocessing in details.

3. **(30 points) Methods:** Explain machine learning methods used for classification task. Give all details about the methods like the algorithms used, parameter tuning, etc.

4. **(20 points) Results and Conclusions:** Explain your results. Give your Kaggle score and ranking. You can provide the score you measured with other metrics like precision, recall, etc. and plots of the related performance.

5. **(5 points) References:** The list of references cited in the report. Don't forget the citation to the related reference in the report.

3. Code

You can implement the project using one of the languages of Pyhton, Matlab, or R. Your code should start with loading train and test data and end with producing submission.csv file.

Tidy up your code as to

- run simply,
- get all necessary inputs as function parameters (train and test data, model parameters),
- produce output, i.e. the submission file (test predictions)
- have explanatory comment

You must implement the functions below:

- loadData (Xpaths) -> Xtra, Xtst
- preprocessing (Xtra) -> Xtra^new
- trainModel (Xtra or Xtra^new) -> model
- predict (model, Xtst)
- writeOutput -> submission.csv
- [and any other necessary functions]

You must provide a **How_to_run.txt** which explains how to run your code.

4. Project Overall Evaluation

For the project, you will provide a final report in IEEE conference paper format (that is given to you in both Word and Latex format). Total score of your project will be calculated as follows:

- Kaggle Result: 50%. (The public and private leaderboard scores will be averaged.)
- Report: 50%

Bonus Marks

Top five team will be rewarded with bonus marks, respectively, 50pts, 40pts, 30pts, 20pts and 10pts., according to the average of the public and private leaderboard scores.

Ninova Submission Policy

- Submit your report and code in a zip/rar file through Ninova on time.
- No late submissions will be accepted

References

To learn more about Kaggle Competition, https://www.kaggle.com/docs/competitions

Term project competition, https://www.kaggle.com/t/80529fc6241245479bcc9f83f34b7ac2

If any part is not clear, please let the teaching assistants know by email or message board on Ninova.

Araş. Gör. İsmail Bilgen, ibilgen@itu.edu.tr

Araş. Gör. Fulya Çelebi Sarıoğlu, sarioglu16@itu.edu.tr