Course Project

DA 324, IIT Guwahati

April 5, 2024

Instructions

- You can work alone or in a team of two for this project.
- Please use this link - `https://www.kaggle.com/t/7efbcac524a846f19431fb510872be57`. for joining the competition.
- Please don't make any of the course project information public.
- If you use online reading materials such as a published paper, make sure to mention them in your report.
- No use of online tools such as ChatGPT or Github Copilot etc permitted.
- Use of standard toolboxes like scipy, scikit-learn is permitted.

Task Description

You are provided three files - *attributes.csv*, *adjacency.csv*, and cluster seeds (*seed.csv*). The dataset has 11952 samples.

Part A: Use the first 10952 data samples and design a model to classify the data into 10 clusters.

- You may not use any neural network for this part.
- Use only the methods discussed in class.
- Your model must use all three files - *attributes.csv*, *adjacency.csv*, and *seed.csv*.

Part B: Use data and cluster labels from Part A to train a neural network of your choice and perform prediction on the remaining 1000 samples.

Part C: Submit the 11952 final cluster labels to the Kaggle Leaderboard. Please remember that the Kaggle competition will allow limited submissions per day.

Grading

Please submit your final report and working code in a compressed zipped folder. Students will be evaluated on the following basis -

Final Report	60 %
Kaggle Performance	20%
Code	20%

The final report should include sections on Data visualization, Model Description, HyperParameter Selection & Tuning, and Failed Attempts. Also, make sure to include a screenshot of your Kaggle leaderboard status and relevant code snippets/screenshots. You may follow any standard report template.

Your code should be readable. Please make sure we can run your code without requiring any changes. Also, make sure to comment on your codes. This is a good time to start writing production-ready codes(if not already).