AI 511 Machine Learning 2023

Assignment 2

Deadline: 30^{th} November 2023 11:59 pm

Marks: 20

Guidelines

- There are two questions in this assignment. You can form a team of 1-2 members.
- For Q1, besides the competition, you will submit .ipynb file and a report on LMS. Keep the name as $< TeamName >\_Q1.ipynb$. The report is expected to have all the steps taken by you: pre-processing, which models used (and why), what results did they gave, any plots, etc.
- For Q2, you will submit the .ipynb as $< TeamName >\_Q2.ipynb$
- **There will be a viva for this assignment too**.

---

1 Book Review - 15 marks

Given the review text and other information about it, you have to predict the rating associated with that review.
Check the attached excel file which contains the group details, Kaggle competition links, and WhatsApp group links.
Although the dataset is same for everyone, the competitions will be in groups. Each TA will hosting his own competition on Kaggle. You have to participate **ONLY** in the assigned competition.
This is a **classification** problem. Rating 0, 1, 2, 3, 4, and 5 are the classes. The metric to be used for leader board is **weighted F1 score**.
Each of the Kaggle competitions contain the dataset to be used. Your goal should be to try different models, parameters, etc and climb up the leader board of your group. So, your model performance matters. Final position on the private leader board (See Kaggle competition page to see what is this) will contribute to your marks. But don't worry, it is not the only thing that will bring you marks. Your efforts and understanding matter the most. **So don't get discouraged by the leader board at all**.

Note:

- For all the rows in test.csv file, you have to predict the rating (0-5). You will submit a csv file on the competition page with *review_id* as first column, and rating as second. Check the *sample_submission.csv* on the competition's page.

- You can make many submissions to the competition, the best one will be reflected on the public leader board. Although there is a **daily limit** of 15 submissions
- You can choose any 2 of your submissions to be considered for the private leader board.

Note that this is a **Natural Language Processing (NLP)** dataset, so you are expected to do extensive research on how to tackle NLP problems as this is a fairly new domain for you. To help you, here are some **hints**:

- Analyze the text data and perform appropriate preprocessing steps such as **text cleaning**, **tokenization**, or **stemming/lemmatization**.
- You can create features using basic techniques like **bag of words** or **TF-IDF**.
- The above two steps will affect your model performance a lot, so do play around with different preprocessing techniques and feature engineering hyperparameters.
- This is a multi-class classification problem. The class distribution might be skewed towards higher ratings (1 star ratings are always rare). How would you tackle class imbalance?
- Remember that you don't have test data labels. Use validation techniques to test your model before submitting on Kaggle.

Lastly, this is a fairly **large dataset**. Don't waste time by running time-consuming code again and again. Save your preprocessed data and read about saving models with pickle.

On LMS, you will submit the .ipynb that gave you the best results (the one you selected for private leader board) and the report.

2 Neural Networks - 5 marks

You have to build a neural network using **Numpy**. So you **CANNOT** use TensorFlow, PyTorch or any other library with built-in neural networks.
The dataset is uploaded on LMS along with this assignment. It is a regression task. You have to predict the 'Price' of the house.