Course Project – Artificial Intelligence

Introduction

The purpose of the course project is to provide students an opportunity to implement numerous **Machine Learning Models** on a real-world data set. **This is a group project with a maximum team size of Three (3) students. Also, COMP 5600 students can only partner with another COMP 5600 student, while COMP 6600 students can only partner with COMP 6600 students.**

Grading Criteria: Your project (25% of your total grade) will be graded primarily based on the following weights:

- Project Proposal Report - 5%
- Project Progress Report - 10%
- Project Final Report - 10%

Topic (Dataset) Selection

Each team will select a multi-class classification problem (at least 3 or more classes) from the Huggingface Dataset Hub. https://huggingface.co/datasets

a. The dataset can be of the following four types.
    i. Audio
    ii. VISION
    iii. TEXT
    iv. Tabular
b. If you are a 5600 group, the data set must have the following properties:
    i. More than 5K training samples
    ii. More than 1K testing samples.
    iii. More than 6 Attributes (excluding the class)
c. If you are a 6600 group, the data set must have the following properties:
    i. More than 100K training samples
    ii. More than 10K testing samples.
    iii. More than 2 classes
    iv. More than 20 Attributes (excluding the class)

A great tutorial on how to use Huggingface Datasets can be found here:
https://huggingface.co/docs/datasets/index

Some domains of interest where multi-class classification may be applied are given below:

- Recommendation Systems
- Image Classification
- Search & Information Retrieval
- Social Network Analysis
- Sentiment/Emotion Classification
- Health Document Classification
- Legal Document Classification
- Spam filtering
- Customer behavior prediction
- Ad click-through rate prediction.
- Song Genre Classification
- Product categorization
- Malware classification
- Anomaly detection problems such as fraud detection

Methods to Implement

Once you pick a multi-class classification problem, e.g., image classification, text classification, etc., you will solve it by using the following Machine Learning Algorithms.

1. If you are a 5600 group, you will implement:
    a. Logistic Regression
    b. Multi-layer Perceptrons (MLP)
2. If you are a 6600 group, you will implement:
    a. Decision Tree
    b. Naive Bayes Classifiers
    c. Logistic Regression
    d. Support Vector Machine
    e. Multi-layer Perceptrons (MLP)

You can use Huggingface/SKLearn library functions for your project.

Project Proposal Report

You must write a two-page proposal before you begin your project in-depth. These will be submitted via Canvas. In the proposal, you should (1) address the following questions, (2) include the names and email addresses of all the team members, and (3) identify the coordinator of the project, who would take the primary responsibility of coordinating the work of all team members; the coordinator is also our primary contact for providing feedback about the project.

As long as these questions are addressed, the proposal does not have to be very long. A couple of sentences for each question would be sufficient. You should focus on the following in your proposal:

- Define clearly what dataset will you work on. (What is the primary application domain? What are the inputs? What are the outputs?)
- How Machine Learning techniques can be applied to solve this problem?
- What are the potential challenges you may face in this project?
- What is your dataset? How will you create/build your dataset?
- How do you plan to build it? Identify what technologies you plan on leveraging to implement your software. This may be programming languages, supporting libraries, etc.
- What hyper-parameters will be involved in your classifier? How would you fine-tune these hyper-parameters? (optional for 5600 students)
- How will you demonstrate the usefulness and correctness of your classifier?
- Provide a rough timeline to show when you expect to finish what. List a couple of milestones if possible (they can be tentative).

Project Progress Report

You must write a three-page project progress report. These will be submitted via Canvas. In the report, you should (1) address the following questions, (2) include the names and email addresses of all the team members, and (3) identify the coordinator of the project in case the team has more than one member, who would take the primary responsibility of coordinating the work of all team members; the coordinator is also our primary contact for providing feedback about the project.

In this progress report, you should focus on the following in your proposal:

- Provide detailed statistics on your dataset, including the number of training examples, testing examples, attributes, etc.
- What evaluation metrics you have chosen to evaluate the learning method and why?
- Preliminary results of at least one (for 5600 students) / two (for 6600 students) machine learning method.
- Create a random baseline for your task and compare your preliminary results against the baseline.
- Summarize the difficulties you have faced so far.
- Your plans to overcome the challenges you have already faced?
- Training loss curves for your preliminary results? (optional for 5600 students)

Project Final Report

At the end of the semester, during the final exam week, every project team must submit a project report. The main artifact to deliver here is the codebase itself, which will make up 40% of the grade for your report.

You, however, must also submit a written report. This report should serve two purposes: (1) it should describe the ML algorithm implemented to solve the task in sufficient mathematical detail, and (2) it should provide documentation in the form of a tutorial that describes how to use your software. Also, clearly describe your evaluation metric and the formulas used to compute that metric. Based on the results, summarize your work, conclude, and discuss how you think the work can be further improved/extended.

The project report should be:

- 8 pages minimum with 11 pt. font, if you are a 6600 group.
- 5 pages minimum with 11 pt. font, if you are a 5600 group.

In your documentation, you should provide enough details that a new user could utilize your software. There is no page limit here, but you should be as detailed as is necessary but not overly verbose. Your documentation should be practical. (No, your code is not documentation.) We should be able to follow your tutorial and be able to use your software.

Because of the nature of the final deliverable for such a project, a strong emphasis will be placed on your actual software itself. You will submit your code base.

Report Grading. We will focus on five main areas: (1) **[10%]** Problem description (clarity, completeness); (2) **[10%]** ML theory and implementation (clarity, correctness); (3) **[10%]** Software documentation (clarity, completeness, instruction to run learning algorithms); (4) **[30%]** Results of multiple methods and discussion of findings/observations including Hyper-parameter tuning. The hyper-parameter tuning part is optional for 5600 students. (5) **[40%]** Code Base.

Each project group needs to submit a single report. **Please make sure to include, for each member of your team, a couple of sentences to describe what he/she did exactly for the project.**