IT1244: Artificial Intelligence: Technology and Impact
Semester 1, 2023/2024

Course Project

This course project helps you to apply the AI/ML concepts that you have learnt in this course into real problems. The course project should be done in groups of 4 students. For team project, the members are responsible for dividing up the work equally and making sure that each member contributes to the project.

Revision history of this project document:

- Sep 25, 2023 – Initial release

1 Project Submission

Project is to be submitted to Coursemology Project submission folder (one submission per team).

1.1 Due dates

- Project proposal submission[1] – **5th October, 11:59 pm**
- Final project submission – **5th November, 11:59 pm**
- Project presentations are scheduled to commence from **6th November**.

1.2 Deliverables

The following are the deliverables for this project:

(a) **Code folder:** Implement your code in Colab or in Jupyter notebook, and write proper comments and documentations for your code. This folder should contains your colab/jupyter notebook and any additional code files that you use for your project.

(b) **Dataset folder:** If you use any other external data in your project, then add this data into this folder.

(c) **Readme.pdf:** Write a step-by-step instruction on how to run the project. The grader will follow this steps to run your code.

(d) **Project Report (pdf):** Write a clear, succinct 4 page report about your project. For more details, see the subsequent sections.

All the above should be put into a folder and the folder should be archived to a zip file. Submit only the zip file to the Coursemology project submission folder. The following naming format should be used: “IT1244_Team<number>_Project.zip”. Example: “IT1244_Team5_Project.zip”. Team numbers will be assigned later.

1.3 Presentation Slides

Project presentation slides should be submitted to Coursemology after your presentation with the same naming format. Example, “IT1244_Team5_Presentation.pptx”

[1] https://forms.office.com/r/NE69g7Q4cC

2 Grading

The maximum score for this project is 100 marks. The grading rubrics, report formats and presentation details are given below:

2.1 Project Main Component – (50%)

2.1.1 Feature Extraction and Visualization

- Are you applying any method for feature selection?
- Have you visualized the features that you are using with respect to the ouputs that you need to predict?
- How are you handling rare/missing values?
- What kind of algorithms are you using for effective feature representation?

2.1.2 Related works

- For your implementation, which articles/research papers did you take ideas from? (we are expecting 2-3 such explorations to get some idea)
- What kind of step did you take to overcome the limitations existing in your studied works?

NOTE: Make sure to give the article/research paper references in your submitted colab/jupyter notebook somewhere at the very start for our convenience. Also mention your approach for overcoming their limitation in 2-3 sentences as comment.

2.1.3 Machine Learning Algorithm Details

- Are the AI/ML solutions adopted appropriate for the problem?
- Is the AI/ML solution implemented correctly?
- What are the original elements done in the project?
- Code documentation: are code comments provided?
- Attributions: are appropriate references/acknowledgement/citations given for the codes that are taken somewhere else?

2.1.4 Experiments and Results

- Did you implement multiple models (baselines and best)?
- Did you perform training, cross validation and testing of your model properly?
- Did you use reasonable performance metrics?
- Did you compare with any previously done work?

2.2 Project Report – (30%)

The project report should be strictly following the AAAI format[2] and should not exceed 4 pages (does not include references). The report should be precise and explain your contributions clearly. The following organization of your report is recommended:

[2] https://www.aaai.org/Publications/Templates/AuthorKit23.zip

(a) **Introduction –** Describe the problem that you want to solve using AI/ML, why it is important, and what AI/ML techniques that you plan to use and why. Investigate the works (2-3 recent works) that has been done with respect to the problem, what methods have been used to solve (if any) and what are their drawbacks. You can ignore the abstract section and start the report with introduction.

(b) **Dataset –** Explain the dataset that you are using, what are the issues with the dataset and what analysis and processing that you did to the dataset (includes visualization and plots), etc. In addition, if you use external data to help with your project, explain the reasoning behind their use and how you used them.

(c) **Methods –** Explain your technical approach in solving the problem that you stated in the Introduction. You will formulate the problem (mathematically if needed), explain why you chose this particular AI/ML technique. If you are using multiple methods you can explain each one of them and can justify them, if you are using AI/ML solutions that are not taught in this course, please specify them. Explain your method using figures, tables and flow-charts to explain your solution. If your method solves some limitations of previous works, then mention that as well.

(d) **Results & Discussions –** You will explain how you have evaluated the solution – how many experiments you have run, what performance metrics you have used to evaluate your model, how did you fine-tune your performance, etc. You will also report the results in tables, charts and figures. You should also list out your findings after running your experiments – explaining with evidence on why a particular model is performing poorly or well.

2.3 Presentation – (20%)

You will be given 20 min to present your project and 5 min for Q/A. More details on presentation schedules will be released later. All the members of the team are required to be present during the presentation. In the final slide of your presentation, you will include the breakdown of the percentage of work completed by each group member.

2.3.1 Quality of presentation

- Is the presentation presented within the allocated time?
- Does the presentation flow well?
- Is the presentation paced appropriately?
- Did the speakers speak in an appropriate tone?
- Do the speakers explain well instead of just reading text in the slides?

2.3.2 Clarity/Understanding of the members

- Do the speakers express the ideas clearly to the audience?
- Do the speakers motivate the importance of the project to the audience?
- Does each member have the technical knowledge of the project?

2.3.3 Presentation Materials

- How are the materials in the slides? Are they neat, colorful or visually creative?
- Have appropriate figures/graphics been used to support the main ideas?

3 Datasets

Here, we provide some datasets that are curated by IT1244 project coordinators. We divide the datasets into several categories to make it easier for you to browse and select. You may perform supervised (e.g., classification, regression) and unsupervised (e.g., clustering) learning, or combinations of them on any of this dataset. You can download the datasets from this link[3].

To assist you in choosing a dataset, we have included an estimated difficulty level for each task, categorized as either easy, medium, or challenging. Naturally, the higher the difficulty, the more learning opportunities, the higher the room for improvements, and the more we will appreciate your work. You know that the entire project carries 25% of the total grade for IT1244. If you choose to work with medium or challenging dataset, then you will be awarded a bonus of up to 2% marks (i.e., up to 8 marks for 100) if you fail to receive the full 25% grade of the project, i.e., the maximum marks that you can get for this project is capped to 25% of your final grade.

3.1 Tabular Datasets

3.1.1 Housing Price Dataset [Easy]

Resale transacted prices in Singapore from 1990 to present, managed by the Housing and Development Board (HDB).

3.1.2 Breast Cancer Dataset [Easy]

Dataset containing features that are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass and the diagnosis.

3.1.3 Disaster Dataset [Easy]

Dataset containing details of the passengers of a certain ship disaster.

3.1.4 Fraud in Electricity and Gas Consumption Dataset [Medium]

Fraudulent activities data from an electricity and gas company; containing information about client data and billing history from early 2000.

3.2 Textual Datasets

3.2.1 Twitter Sentiment [Medium]

A sentiment analysis dataset based on Twitter tweets in the format of csv.

3.2.2 Sentence Sentiment [Medium]

A sentence-based sentiment analysis dataset in the format of csv.

3.2.3 Movie Review [Challenging]

A large-scale sentiment analysis dataset based on public movie reviews in the format of raw text.

[3] `https://tinyurl.com/2wt7eywy`

3.2.4 DNA Binding Protein [Challenging]

A large protein sequence dataset in the form of raw fasta file.

3.3 Table-Text Dataset

3.3.1 Financial Table-Text Question Answering Dataset [Challenging]

A large-scale table-text question answering (QA) dataset, where the tables and text are selected from real-world financial reports, and the questions are generated manually by humans with rich financial knowledge.

3.4 Time-series Datasets

3.4.1 Stock Market Dataset [Challenging]

Multiyear stock market data on a few hundred companies in US.

3.5 Image Datasets

3.5.1 Clothing Dataset [Medium]

Dataset containing clothing image and its associated category (e.g., shirt, trouser, shoes).

3.5.2 Brain Tumor Dataset [Challenging]

Dataset containing brain MRI image and its prognosis.

3.6 Audio Datasets

3.6.1 Spoken Digit Dataset [Challenging]

A speech dataset containing spoken digit in English from several speakers in a total of a few thousands recording.

3.6.2 Cats and Dogs Dataset [Medium]

A cat and dog sound dataset containing a few hundred audio recordings of these animal sounds.
