INFSCI 0510 Data Analysis Midterm Project

Please read the instructions carefully, you can form a group to work the midterm project. If you have any queries regarding the understanding of this sheet, please contact the TAs as soon as possible. **Report due on: May 21st.**

1 Introduction

In this midterm project, you are required to conduct research on a relevant topic for data analysis. In order to present an academic study, your report may need to explain things regarding the following suggested questions:

- What is your topic, or the details of your task?
- Why you choose the specific model in this report for the task?
- What is the theoretical demonstration of model, e.g., from input, to output and loss calculation?
- What is the general performance of your model, compared with others' in this field?
- Why you choose the specific hyperparameters, what about changing these hyperparameters or altering your model structure, how will your performance vary?
- ...

Your core functionalities should come from, but are not limited to the following learning tasks:

- Regression
- Classification
- Clustering
- Dimensionality reduction

Note that the data involved in your task can be the conventional one we have seen during lectures, or data comprises dynamics along certain dimension, e.g., time-series data.

Also note that the project can be a solo one or a group one, we expect that groups have **no more than 4 members** however we also don't want to make a strict limit. The principle is that **more group member should present higher-quality report**.

Please consider this project as:

- A place to do fun things and get professional comments for revising the project.
- A place to start some long-term research works for an actual publication in near future; e.g., this can be done easily by transferring the models or the task scenario with more influenced ones to get credits in academic conferences or journals.

2 Datasets

To support your exploration and implementation, the following platforms are recommended for accessing datasets for your task scenarios. These resources span a wide range of domains and are particularly useful.

- Kaggle (https://www.kaggle.com/datasets)
  A widely-used platform hosting a diverse collection of public datasets along with real-world competitions. It also provides kernels (notebooks) with example solutions and discussions to guide your thinking. Suitable for both beginners and advanced students.
- UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php)
  A classic and reliable source of benchmark datasets used in academic research and education. Each dataset typically includes a detailed description, attribute information, and task suggestions.
- Google Dataset Search (https://datasetsearch.research.google.com/)
  A search engine specifically designed to discover datasets across the web. It aggregates datasets from thousands of repositories and is ideal for exploring new or specific domains of interest.
- Papers with Code – Datasets (https://paperswithcode.com/datasets)
  This platform links datasets with state-of-the-art methods and code implementations, helping you align your project with current research trends. Highly useful if you aim to benchmark performance or test existing models.
- DataHub (https://datahub.io/)
  A community-driven platform that offers open datasets in clean formats (CSV, JSON) and a data preview interface. Particularly handy for quickly prototyping and testing models.
- Awesome Public Datasets (GitHub)
  (https://github.com/awesomedata/awesome-public-datasets)
  A curated list of high-quality public datasets organized by topic. It is especially useful if you are interested in domain-specific applications such as healthcare, education, finance, or the environment.

You are encouraged to explore these resources not only for datasets but also for understanding practical use cases, recent modeling techniques, and relevant evaluation strategies. Please **cite the dataset source appropriately** in your report.

3 The In-Depth Analysis

To get **excellent grade**, please consider addressing the following questions, unless not fitting your scenario:

- Besides numbers, any illustrations to demonstrate the performance?
- For complex model, how the performance vary if removing certain components from your model?
- Why the hyperparameters are set with these values?
  (Note that there is no need to explain the the learning rate and epoch number in training)
- Deficiencies of your model that leads to potential future work?
- Is your model efficient to train or use?

Feel free to design your own in-depth analysis and highlight that to get extra credits, as long as the analysis provides valuable insights of your model.

4 Writing

For the section structure in the report, we recommend the following one, note that sections with titles highlighted in blue are optional:

- Abstract

- Introduction
- Related Works
- Model
- Experiments
  - Data
  - Implementation Details
  - Performance
  - Further Studies
- Conclusion

Feel free to add subsections for improving your presentations.

As for other writing rules, we provide the **IEEE conference template** for you, the template files are either in ***MS Word*** style or ***Latex*** style, please check the Blackboard and choose one template to write your report and closely follow the instructions listed in the template file. Please note that the **page limit of your report is 5 pages, including references. Please contact us if you DO require extra pages for putting up key contents in your report.**

5 Marking

The following **marking instructions** will be applied to assess your report:

- A concise and supportive introduction to the background and the task. (10%)
- The theoretical explanation of your model. (20%)
- Data, evaluation metric(s) and general performance. (10%)
- Reproducibility based on your report. (10%)
- The in-depth analysis of your solution, see Section 3 (30%)
- The presentation of your report. Organization? Clarity? Grammar? References? ... (20%)

Along with your grade, there is also a **feedback** to explain the grade and provide instructions for future improvement, out of this course.

6 Submission

Please note that **only the group leader** is required to submit the report file for the whole group, and the grade will be broadcast to every group member.

For group leader: please rename your submission, i.e., the report (.pdf) file as:

`GroupID_GroupLeaderID_GroupLeaderName`

For example, `7_2023141520000_Sofia.pdf`