NSYSU Practical and Innovative Analytics in Data Science - Spring 2023

Midterm Project

Goal

This course aims to prepare you to engage in a real-world project requiring you to apply skills from the entire data science pipeline: preparing, organizing, and transforming data, constructing a model, and evaluating results. The midterm project is intended to let you focus more on the entire data science pipeline with less on modeling. This is an opportunity for you to jump into a dataset, and understand its structure, then apply your preprocessing, analysis, and deployment skills.

Topics

The recommended subject can be divided into two categories:

1. **Real-world project**. While AI has been focused on models, the real-world data science experience of those who put models into production is that the data often matters more. In this kind of project, pick applications or datasets that interest you and explore how best to apply the preprocessing you have learned in this course to solve them. You can use data wrangling, EDA, data cleaning, and feature engineering techniques to explore and prepare the dataset. In addition, you can explain your model using different explainable AI techniques. Finally, you can deploy your model to the local or cloud server at the end of the project.
2. **Research-oriented**. In this project, you can choose a topic you would like to explore. One way is to pick a topic that is related to data-centric AI (weak supervision, for instance). The goal is to understand the method from the related papers or open-source code. Then you can apply the techniques to different datasets to ensure you understand the method. On the other hand, you can also do summary research. For instance, you can compare the different implementations of the active learning framework or clarify different explainable AI methods. For this kind of research, be sure to read related papers before discussing or giving a conclusion.

.

In general, the main idea here is to spend more time on data preparation and model deployment rather than modeling. Try to pick something that you are passionate about or one that you are interested in! However, the project has to be new, and you cannot use existing work or research for the midterm project. Nonetheless, the midterm project can be built upon your previous research as long as the midterm project is new work you are doing for this class. The following databases are suggested for searching datasets:

- 政府資料開放平台
- Kaggle

- Google dataset search
- UC Irvine Machine Learning Repository
- Amazon's AWS datasets

If you have other project ideas or are uncertain about how to get started, please come to my office hour or the TA hour, and we would be happy to listen to your ideas and give some suggestions.

Example Procedure

1. **Real-world project**
   - Choose a dataset from the above database. If you select a dataset that has been widely analyzed, try to gain new insight into the analysis. In addition, if the dataset is already discussed in a book, then many people have touched it already, and you should try to pick something else
   - Perform data wrangling, EDA, data cleaning, and feature engineering technique to explore and prepare the dataset
   - Build a model, but this is not the critical part, so you can try to use PyCaret, auto-sklearn, or just choose a widely used model for inference purposes.
   - Try to explain your model using different explainable AI techniques. At the end of the project, you can deploy your model to the local server or the cloud server
   - You can find useful tools here
   - Report what you have discovered

2. **Research-oriented Project**
   - Select data-centric methods that you are interested in
   - Understand the method by reading related papers or reading the source code
   - Test it on different datasets and report what you have discovered
   - Be sure to compare the methods on synthetic or the real dataset if you are doing summary work and report what you have discovered
   - You are encouraged to develop a variant of the existing algorithm once you understand it
   - You can find useful tools here
   - Report what you have discovered

Grading Policy and deliverables (40%)

- Midterm Presentation (Scheduled at 4/17 (9:10~12:00)) (**30%**)

Each team is required to present their work to the class. The presentation time is **10 minutes**, with an additional **3 minutes** for Q & A. The grading score will be based on the clarity of the presentation, the relevance of the project to topics taught in this course, and the technical quality of the work. Each team will also be given a grading list containing six characters from A+ to D (which will be translated into 13%~20% in the final grading). You need to provide a letter grade for other teams.

The final score will be the summation of the **grade from students (20%), TAs (5%) me (5%)**.

- Midterm Report (due 4/23 at 11:59pm) (**10%**)

After the class, we will also post all the midterm writeups online so that you can read about each other's work. If you do not want your report to be posted online, please tell us a week before the final submission deadline. Your report may contain the following sections

  - Abstract
  - Introduction and related work
  - Dataset
  - Methods
  - Experiments and results
  - Discussion
  - Conclusion and future work
  - Contributions
  - Reference

Note that your results may not be positive, but you can still report what you have tried so far and have some discussion.

The midterm project report can be at most **10 pages** long (including appendices and figures). The paper size is **standard A4** or 8.5 x 11 inches and the font size must be **greater than or equal to 10pt**. You are free to use single-column or two-column layouts and we will allow for extra pages containing only references. If someone else had advised or helped you on this project, your report must fully acknowledge their contributions. **Please include a section that describes what each team member worked on and contributed to the project.** You are requested to hand in the code to reproduce your result and the code should be written in **Python** unless you can't find the desired package in Python or reproducing the required algorithm in Python is tedious. **Please include a link to your code or upload the zip file of the code** for your midterm project. You do not have to include the data or additional libraries.

The midterm report will be judged based on the clarity of the report, the relevance of the project to topics taught in this course, the novelty of the problem, the correctness of your code and the technical quality of the work**.** The score of this part will be the summation of the **grade from TAs (5%) and me (5%).**

Reference

- Data-Centric AI
- Introduction to Data-Centric AI
- Awesome production machine learning
- Awesome python data science