NSYSU Practical and Innovative Analytics in Data Science

Fall 2024

Midterm Project

Goal

This course aims to prepare you to engage in a real-world project by applying skills across the entire data science pipeline: preparing, organizing, and transforming data; constructing a model; and evaluating results. The midterm project is designed to allow you to focus more on the complete data science pipeline with less emphasis on modeling. This is an opportunity for you to dive into a dataset, understand its structure, and apply your preprocessing, analysis, and interpretation skills.

Topics

The recommended subjects can be divided into two categories:

1. **Real-world project**. While AI has traditionally focused on models, the real-world data science experience of those deploying models into production often reveals that data matters more. In this type of project, select applications or datasets that interest you and explore how best to apply the preprocessing techniques you have learned in this course to address them. You can utilize data wrangling, exploratory data analysis (EDA), data cleaning, and feature engineering techniques to explore and prepare the dataset. Additionally, you can explain your model using various explainable AI techniques. Finally, deploy your model to a local or cloud server at the end of the project. Please note that your project may include some of these steps rather than all of them.

2. **Research-oriented project**. In this project, choose a topic you would like to explore. One approach is to select a topic related to data-centric AI (e.g., weak supervision). The goal is to understand the method through related papers or open-source code. Then, apply the techniques to different datasets to ensure you comprehend the method thoroughly. Alternatively, you can conduct summary research, such as comparing different implementations of the active learning framework or clarifying various explainable AI methods. For this type of research, be sure to read related papers before discussing or drawing conclusions.

.

In general, the main idea is to focus more on data preparation and model interpretation rather than solely on modeling. Choose a topic that you are passionate about or interested in! However, the project must be original, and you cannot use existing work or research for the midterm project. Nonetheless, the midterm project can build upon or extend your previous research as long as it constitutes new work you are undertaking for this class. The following databases are suggested for searching datasets:

- 政府資料開放平台
- Kaggle

- Google dataset search
- Paperwithcode

If you have other project ideas or are uncertain about how to get started, please attend my office hours or the teaching assistant's office hours. We would be happy to discuss your ideas and provide suggestions.

Example Procedure

1. **Real-world project**
   - Dataset Selection: Choose a dataset from the provided database. If you select a dataset that has been widely analyzed, strive to gain new insights from your analysis. Additionally, if the dataset is already discussed in a book, it may have been extensively explored, so consider selecting a different one.
   - Data Preparation: Perform data wrangling, exploratory data analysis (EDA), data cleaning, and feature engineering techniques to explore and prepare the dataset.
   - Model Building: Build a model; however, this is not the critical part. You may use tools like PyCaret, auto-sklearn, or opt for a widely used model for modeling purposes.
   - Model Interpretation and Deployment: Explain your model using various explainable AI techniques. At the end of the project, deploy your model to a local or cloud server.
   - Reporting: Document and report your findings and discoveries.

2. **Research-oriented Project**
   - Method Selection: Select data-centric methods that interest you.
   - Method Understanding: Understand the methods by reading related papers or reviewing the source code.
   - Application and Testing: Test the methods on different datasets and report your findings. Ensure to compare the methods on synthetic or real datasets if conducting summary research.
   - Algorithm Development: You are encouraged to develop a variant of the existing algorithm once you understand it.
   - Reporting: Document and report your findings and discoveries.

Grading Policy and deliverables (40%)

- Midterm Presentation (Scheduled at 10/28 (9:10~12:00)) (**30%**)

Each team is required to present their work to the class. The presentation will last 10 minutes, followed by an additional 3 minutes for Q&A. The grading criteria are based on:

- ✓ Clarity of the Presentation: How well the team communicates their project and findings.
- ✓ Relevance to Course Topics: The extent to which the project aligns with the topics taught in this course.
- ✓ Technical Quality of the Work: The robustness and sophistication of the methodologies and analyses used.

Additionally, each team will provide a letter grade for other teams using a grading scale from A+ to D, which will translate to 8%–15% of the final grade. The final score

will be the summation of the **grade from students (15%), TAs (10%) and me (5%)**.

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

After the presentation, all midterm write-ups will be posted online for you to review each other's work. If you prefer not to have your report posted online, please notify us one week before the final submission deadline. Your report may contain the following sections

- Abstract
- Introduction and related work
- Dataset and methods
- Experiments and results
- Discussion
- Conclusion and future work
- Acknowledgments and contributions
- Reference

Your results do not need to be positive; you can report on what you have tried and include discussions on your findings.

- ✓ Format: The report must be no longer than 10 pages, 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.
- ✓ Acknowledgments: If others have advised or assisted you with the project, fully acknowledge their contributions.
- ✓ Team Contributions: Include a section detailing each team member's contributions to the project.
- ✓ Code Submission: Code should be written in Python unless a specific package is unavailable or reproducing the required algorithm in Python is impractical. Provide a link to your code repository or upload a zip file of the code. There is no need to include data or additional libraries.

Evaluation Criteria: The midterm report will be assessed based on:

- ✓ Clarity of the Report: How well the report is written and organized.
- ✓ Relevance to Course Topics: The alignment of the project with the course material.
- ✓ Correctness of the Code: The accuracy and functionality of the submitted code.
- ✓ Technical Quality of the Work: The overall rigor and sophistication of the methodologies and analyses

The score of this part will be the summation of the **grade from TAs (5%) and me (5%).**

Reference

- https://github.com/HazyResearch/data-centric-ai
- https://github.com/daochenzha/data-centric-AI
- https://github.com/Renumics/awesome-open-data-centric-ai
- (Course) Introduction to Data-Centric AI
- Awesome production machine learning
- Awesome python data science