NSYSU Data Science Capstone Project - Spring 2022

Final 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 final project is intended to let you put more attention on the entire data science pipeline with less focus 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. **Data-Centric AI project**. While AI has been pretty focused on models, the real-world data science experience of those who put models into production is that the data often matters more. This project aims to consolidate this experience: Data-Centric AI embodies a particular point of view around how this progress can happen: by focusing on making it easier for practitioners to understand, program, and iterate on datasets instead of spending time on models. We recommend you to use the Roman MNIST Dataset, which is available here. But you can also work on other real-world datasets from different databases. The main idea of this project is that you should try to fix your model and work on improving the quality of the dataset so that the performance can be improved.

2. **Real-world project**. 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 technique to explore and prepare the dataset. In addition, you can try to explain your model using different explainable AI techniques. You can deploy your model to the local server or the cloud server at the end of the project.

. 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 the one that you are interested in! However, the project **has to be new**, and you cannot use existing work or research for the final project. Nonetheless, the final project can be built upon your previous research as long as the final project is new work you are doing for this class. The following databases are suggested to search for 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, and I would be happy to listen to your ideas and give some suggestions.

Example Procedure

1. **Data-Centric AI project**
   - Choose a dataset from the above database or choose the default dataset, which is Roman MNIST. The Roman MNIST dataset contains 2,067 training data, 813 validation data, and 50 testing data that belong to 10 different Classes.
   - Your task is to optimize model performance by improving the dataset and making training and validation splits. You can try fixing incorrect labels, adding data for side case tuning, applying data augmentation techniques, or using any other method to improve the data. For more information about the data-centric AI and the tools available, please refer to here.
   - The training notebook for Roman MNIST can be accessed in the course website (the model should be held fixed). If you choose other datasets, you can follow the structure of the notebook to perform your task.
   - The baseline will achieve around 68% validation accuracy and 54% testing accuracy in the Roman MNIST dataset. You should try to optimize model performance by improving the dataset.
   - Report what you have discovered.

2. **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.

Grading Policy and deliverables (40%)

- Final presentation (Scheduled at 6/7 (9:10~12:00) and 6/14 (10:10~12:00)) (**16%**)

Each team is required to present their work to the class. The presentation time is limited to **18 minutes**, with additional **4 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 that contains five characters from A+ to E (which will be translated into 11.2%~16%

in the final grading). You need to provide a letter grade for other teams. Finally, the **slide should be written in English**, and it is encouraging to present in English. However, if you found it challenging to present in English, you can still use Chinese.

The final score will be the summation of the **grade from students (16%) and me (4%)**.

- Final Report (due 6/14 at 11:59pm) (**20%**)

After the class, we will also post all the final 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 final 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**. In addition, you should use **English** to write the report. 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 final project. You do not have to include the data or additional libraries.

The final 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**.**

Reference

- Data-Centric AI
- Awesome production machine learning
- Roman MNIST Dataset