NSYSU Statistical Learning and Data Mining–Fall 2022

Final Project

Goal

This course aims to prepare you to apply statistical learning methods to application problems or to leave you well-qualified to start statistical learning or machine learning research. The final project is intended to start you in these directions. This is an opportunity for you to delve into a dataset, understand its structure, and apply your EDA and model-building skills!

Topics

The recommended subject can be divided into two categories:

1. **Application-oriented**. This should be the most common: Pick applications or datasets that interest you, and explore how best to apply the learning algorithms you have learned in this course to examine them. You can use the EDA technique first to explore the dataset. In addition, preprocessing may be required for most real-world datasets. At the end of the project, you should be able to report some insights about the dataset you choose.
2. **Research-oriented**. In this project, you can choose a topic you would like to explore. One way is to pick a statistical learning method or a family of related methods (GLM, for instance). The goal is to understand the method from the original paper or open-source code. Then you can build it from scratch or simplify the existing open-source code to ensure you understand the method. Developing a novel variant of an existing algorithm once you understand the detail is also welcome. On the other hand, you can also do summary research. For instance, you can compare the different implementations of the ensemble framework or clarify different types of classification/regression problems (Multiclass, multioutput and multilabel). For this kind of research, be sure to read related papers before discussing or giving a conclusion.

In general, you can either pick an application or dataset you are interested in or choose some learning algorithms that you want to explore more. So, 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 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. It is also okay if two teams end up working on the same application or dataset as long as they don't coordinate so as not to be biased in how they tackle the problem.

The following databases are suggested to search for datasets

政府資料開放平台
Kaggle
Google dataset search

The following packages are suggested for looking for algorithms that you are interested in or you can use when you are exploring your dataset

- https://scikit-learn.org/stable/
- https://www.statsmodels.org/stable/index.html
- http://rasbt.github.io/mlxtend/
- https://github.com/scikit-learn-contrib/scikit-learn-contrib/blob/master/README.md
- https://github.com/rapidsai/cuml

If you have other project ideas or are uncertain how to get started, please come to the office hours of TA or mine, and we'd be happy to listen to your ideas and give some suggestions.

Example Procedure

1. **Application-oriented Project**
   - Choose a dataset from the above database. If you select a dataset that has been widely analyzed, try to achieve a high leaderboard score or 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 preprocessing and EDA (Do some data analysis using the unsupervised methods or utilize different plots to understand your data)
   - Build a model and try to find important features. You can use statistical approaches to argue that one feature is more important than some other feature
   - Compare different models with proper parameter-selection techniques (i.e., cross-validation). You may employ statistical hypothesis testing to compare two or more models. (you might set up logistic regression or linear regression as baselines). Notice you can test them on both synthetic or the selected dataset
   - Report what you have discovered

2. **Research-oriented Project**
   - Select statistical learning methods that you are interested in
   - Understand the method by reading related papers or reading the source code
   - Build the method from scratch or simplify the existing code to ensure you understand the methods' essence. Test it on different datasets and report what you have discovered
   - Be sure to compare the algorithms 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. For instance, in ridge regression, you may try to give different regularization strengths to different coefficients or force the coefficient to approach a number other than 0 and derive the close-formed solution. The formulation of this variant may depend on your prior knowledge of different application scenarios.

Grading policy and Deliverables (30%)

- Final presentation (Scheduled at 12/19 and 12/26 (9:10~12:00)) (**15%**)

Each team is required to present their work to the class. The presentation time is limited to **12 minutes** with 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 that contains five characters from A+ to C (which will be translated into 6%~10% 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 (10%), TAs (2.5%) and me (2.5%)**.

- Final Report (due 12/30 at 11:59pm) (**15%**)

After the class, we will also post all the final writeups online so 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**. 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. The score of this part will be the summation of the **grade from TAs (7.5%) and me (7.5%).**