NSYSU Statistical Learning and Data Mining–Fall 2025

Final Project

Goal

The primary objective of this course is to equip you with the ability to implement statistical learning techniques to solve real-world application challenges, as well as to provide a solid foundation for those interested in pursuing research in statistical learning or machine learning domains. The final project is designed as a stepping stone towards achieving these goals. It will serve as a platform for you to immerse yourself in a particular dataset, to thoroughly comprehend its intricacies, and to employ your skills in Exploratory Data Analysis (EDA) and predictive modeling to extract meaningful insights. This project is your chance to showcase your analytical prowess and the practical application of the concepts learned throughout the course!

Topics

The project topics for your final assignment are categorized into two primary streams:

1. **Application-oriented**. This may be the preferred choice for most. Find a dataset or a real-world issue that captures your interest, and use the statistical learning tools we've discussed in class to explore and understand it. Begin with Exploratory Data Analysis (EDA), which is like detective work to uncover what the data is telling you, often revealing patterns, anomalies, or relationships. Your dataset might be a bit untidy, so you'll likely need to clean it up and organize it before you can fully analyze it. After working through the data, your goal is to come up with some clear insights or findings about it. For instance, if you're looking at a dataset of local weather patterns, you might end up predicting rainfall amounts or identifying climate trends.
2. **Research-oriented**. If you're drawn to deep exploration, this is your route. Choose a statistical method or concept we've covered, like Generalized Linear Models (GLM), and aim to understand it deeply. You might review the original research papers or existing open-source code, and then try to recreate the algorithm in a simplified form. Once you fully understand it, you could even attempt to improve it or adapt it in a new way. Alternatively, you could compare various approaches to a common data problem, like how different algorithms perform in ensemble methods or explain the differences in tackling Multiclass, Multioutput, or Multilabel classification problems. If this is your project choice, make sure to thoroughly read and analyze related research to inform your project.

Your project should be based on something that excites you, whether it's a specific type of data, a question you want to answer, or a method you wish to master. Your work **must be original** for this class, but it can build upon what you've learned from your

past projects. It's fine if others in the class choose similar topics, as long as each person or group works independently. This way, even with the same starting point, each project can take a unique direction based on your individual analysis and perspective.

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 your own ideas for a project or feel unsure about how to begin, don't hesitate to reach out. You're welcome to drop by during the office hours of your Teaching Assistant or mine. We're here to listen to your thoughts and provide guidance. Whether you need help refining your idea or finding a starting point, we're eager to assist you in launching a project that's both exciting and rewarding.

Example Workflow

1. **Application-oriented Project**
   - ✓ Dataset Selection:
     - Choose a dataset from a suggested list or any other reputable source. Aim for less common datasets to avoid repeating analyses. If it's a popular dataset, focus on climbing the leaderboard or uncovering unique insights.
     - Avoid datasets that are extensively covered in well-known books or resources, as these have likely been thoroughly examined.
   - ✓ Data Preprocessing and EDA:
     - Clean and prepare your data for analysis.
     - Conduct exploratory data analysis using unsupervised methods or create various plots to deeply understand the characteristics of your data.
   - ✓ Model Development:
     - Build a predictive model and identify key features within the dataset.
     - Use statistical methods to argue the importance of one feature over another.
   - ✓ Model Comparison:

- Compare the performance of different models, ensuring you have selected the right parameters (for example, through cross-validation).
- Consider statistical hypothesis testing to compare model performances.
- Establish simple models like logistic or linear regression as baselines for comparison.

✓ Discovery Reporting:
- Document and report on your findings and the insights gained from your analysis.

2. **Research-oriented Project**

✓ Method Selection:
- Choose one or more statistical learning methods that pique your interest.

✓ In-depth Understanding:
- Delve into the chosen method(s) by reviewing academic papers or studying the source code.

✓ Method Development:
- Build the method from the ground up or streamline existing code to grasp its core principles.
- Apply your method to various datasets to test its efficacy.

✓ Algorithm Comparison:
- If you are summarizing, compare your algorithm's performance on synthetic or real-world datasets.
- Document the outcomes of these comparisons.

✓ Algorithm Innovation:
- Once you have a solid understanding, consider developing a variation of the algorithm.
- For example, with ridge regression, experiment with different regularization strengths for different coefficients or constrain coefficients to approach a specific value rather than zero, and work out the solution.
- Tailor your variant to different application scenarios based on your understanding and research.

Grading policy and Deliverables (30%)

1. **Final presentation** (**15%**)
- Schedule: December 17 and December 24 (9:10 AM - 12:00 PM)
- Format: Each team will present their project for 9 minutes, followed by a 3-minute Q&A session.
- Criteria: Presentations will be evaluated based on the clarity of delivery, the project's relevance to course content, and the technical sophistication of the work.
- Scoring Breakdown: The final presentation score combines Teaching Assistants' grades (7.5%), and the instructor's grade (7.5%).

2. **Final Report (15%)**
   - Deadline: December 28, 11:59 PM
   - Public Sharing: All final reports will be shared online with the class unless a team opts out one week before the deadline.
   - Report Structure: Your report should include the following sections:
     - Abstract
     - Introduction and Related Work
     - Dataset Description
     - Methods
     - Experiments and Results
     - Discussion
     - Conclusion and Future Work
     - Individual Contributions
     - References
   - Details: Note that your results may not be positive, but you can still report what you have tried so far and have some discussion. Try to follow the format described here. Your report **should not exceed 10 pages**, inclusive of appendices and figures. Use **A4 or 8.5 x 11-inch paper size with a minimum 10pt font.** Single or two-column format is acceptable, with additional pages allowed for references only.
   - Acknowledgements: If you received advice or assistance, you must fully recognize these contributions.
   - Team Contributions: **Detail each member's specific contributions to the project**.
   - Code Submission: **Submit your Python code used for generating results** (exceptions made if Python is unsuitable for the task). Provide a code link or a zip file upload, excluding data and external libraries.
   - Evaluation: The report will be assessed on clarity, course relevance, problem novelty, code correctness, and technical merit.
   - Scoring: The report's grade will be the sum of the Teaching Assistants' (7.5%) and the instructor's (7.5%) evaluations.