ORIE4741/5741 Spring 2023: Projects

Haiyun He

March 7, 2023

1 Project Overview

The class project is an interesting opportunity for you to try the machine learning / data modeling techniques learnt in class to interesting real-world multivariate analysis problems. The project will comprise 30% of your final grade.

There will be three options: 1) Data analysis; 2) Algorithm development; 3) Theoretical project. It may be harder to do well for the latter two options as new algorithms or new theoretical insights ought to be made.

For ORIE 5741 students, if you choose "Data analysis" project, your project must be business-oriented and address a clear business problem. Projects will require a final (video) presentation, in addition to the other requirements for ORIE 4741.

All projects will be completed in groups of **2–3 students**.

2 Project Description

Each group can take one of three forms:

A. **Data analysis:** Formulate an important and interesting question, and show how to use big messy data to answer (or try to answer) the question. The final product will be a paper suitable to present to a subject matter expert well versed in the problem domain (but not necessarily in big messy data analysis). This project will give students experience in the kind of work that a data scientist might perform in industry, or at a government or nonprofit agency.

B. **Algorithm development:** Design a new method for analyzing big messy data. The final product will be a draft paper suitable for submission to NeurIPS, ICML, or KDD. This project will allow students to experience the kind of work that a researcher might perform in academia or in an industrial research lab.

C. **Theoretical project:** Review one technical paper from the list given in Section 9. If you would like to suggest a paper, please check with the instructor before you do so. The final product will be a paper containing a summary of the article, a review of the article from an advocate viewpoint, and a review of the article from a critic viewpoint. Additionally, students should go beyond the paper to prove something non-trivial, to simplify an existing proof, or to improve algorithmic aspects of the paper. This project will give students experience in theoretical research in machine learning and data science.

The final project report for Options A and B should be no more than 8 pages (excluding references) and for Option C, no more than 5 pages (excluding references). The final project report cannot be too much shorter than the page limit.

3 Project Guideline

Here are the class projects for the Fall 2020 term and the Fall 2021 term. Some of our favorites:

- How much is a house worth?

- Which water wells need repair?
- Which hospital should I go to?
- Can we predict basketball players' performance?
- Does Google Trends data predict stock performance?

Some algorithm development projects from Fall 2019:

- Fairness in Machine Learning
- Deep Reinforcement Learning for Combinatorial Optimization
- Fair clustering
- Learning rate in deep neural networks

— What makes a good **data analysis project**? Here are a few considerations:

- Clear outcome to predict
- Some techniques learned from class should do something interesting, e.g. linear classification, linear regression, feature engineering and etc.
- New, interesting model; not a Kaggle competition

— How might you come up with an **algorithm development project**?

- Read a collection of several (maybe 5) recent papers from a top computer science conference (eg NeurIPS, ICML, or KDD) on the same topic.
- Design some computational experiments to explore the performance of some of these methods on a collection of tasks.
- Use what you learn from these experiments to design a new method that performs better. Performance might mean many things: accuracy, speed, computational resources, interpretability, fairness, robustness to corruptions in the data, ...
- Can you prove anything about how well your method works?
- A good project would implement several methods, produce careful experiments comparing them, and would implement at least two tweaks to the published ideas and see how well they work. The end result might be recommendations for best practices when using these algorithms on a new dataset.
- A great project would produce a new algorithm that works better than the previous published state of the art on a wide variety of datasets - but this is not necessary to get a good grade on the project for this class.

— What makes a good **theoretical project**?

- In the Summary, you should articulate, in your own words, the main problem addressed in the paper, the solution approach and the main results in the paper. Try to answer the questions: "What is this paper about? What are the main ideas? What are the main results and takeaways?"
- As an Advocate, you should defend the paper by identifying the important contributions, new techniques and any other points. Try to answer the questions: "Why is this a really good paper? What are its strengths? What related problems can the results be applied to? Will this paper stand the test of time?"
- As a Critic, you should pick apart the paper by finding the weak points, any simplifying and/or unrealistic assumptions, and so on. Try to answer the questions: "Why is this not that great a paper? What are its weaknesses and limitations?"
- Additional statements proved / Proofs of existing statements simplified / Algorithms improved (to be justified either theoretically or experimentally)

4 5741 vs 4741

ORIE 5741 projects have slightly different requirements from ORIE 4741.

- For data analysis projects, projects should be more business-oriented, using techniques from the class to address a clear business problem.
- Projects will require a final (video) presentation, in addition to the other requirements for ORIE 4741.
- The project presentations should be uploaded to an online platform, e.g., YouTube, OneDrive, Dropbox and etc. (make sure to set visibility to either Public or Unlisted so that peer reviewers and graders can view it). The video link should be added to the "README.md" file on your team's public GitHub repository.

If any student in your group is taking ORIE 5741, your group will be be required to fulfill the ORIE 5741 requirements.

5 Project Timeline

All project deadlines are at 11:59pm EST.

- **Mar 5, 2023**: Submit your choice of group project here.
- **Mar 19, 2023**: Submit project proposal. For Options A and B, the proposal should include a problem statement and description of at least one data set. For Option C, the proposal should include a draft summary of the paper and an outline of the review.
- **Mar 24, 2023**: Peer reviews of project proposals due. (Review assignments will be posted later.)
- **May 09, 2023 (date of the last lecture)**: Final project reports due and project presentations due (for ORIE 5741, as prerecorded video). Add a link to your video on your project's README.
- **May 19, 2023**: Peer reviews of project reports due. (Review assignments will be posted later.)

6 Detailed Requirements

Project repository Your project team should create a public GitHub repository. Each team member should have push access to the repository. Add a file named README.md to the repository, in which you state the name of your project, list the names and NetIDs of the project members, and describe your project in a paragraph or two. Make a pull request (PR) to add a link to your repository to the list of ORIE 4741/5741 projects. (Click the link for detailed instructions.)

Project proposal The project proposal should be no more than 1 page, in PDF format. You can either write in LaTeX, markdown, Word or etc.. It should be posted on your project **Github repository** and uploaded to **"Assignment"** in Canvas as well with the filename "project_proposal". (The file extension should be .pdf)

For Options A and B, it should identify a question, and a data set that you'll use to answer the question. Justify why the problem is important, and why you think the data set will allow you to (begin to) answer the question. Stylistically, the proposal should be written as though it were a memo to your manager (at whatever kind of enterprise might care about this question: either government, nonprofit, or industry). You should justify why it's worthwhile to this enterprise for you to work on the project for a few months, and why you think you're likely to succeed.

For Option C, it should identify what the paper is about, the main results and your draft insights.

Proposal peer review What do you like about the proposal? What concerns you? Do you think you could use the results of this study? What other aspects of the question do you think the group should consider?

Submit a grade and comments on the proposal for the proposal via **Canvas peer review**. Review assignment will appear in Canvas automatically. Please give points according to the rubrics. The Canvas peer review user guide can be found here: https://community.canvaslms.com/t5/Student-Guide/How-do-I-submit-a-peer-review-to-an-assignment/ta-p/293.

Concretely, your comments should begin with a one paragraph (at least three sentence) summary of the project you're reviewing: What's it about? What data are they using/what problem do they propose? What's their objective? Then detail at least three things you like about the proposal, and three areas for improvement (at least one sentence each). Make sure to back up your subjective assessments with reasoned, detailed explanations.

Project presentation Submit a 5–10 minute video describing the problem your project sought to address, the techniques that you used to solve it, conclusions you were able to draw, and directions for future work. You might organize your presentation following the template for project reports, described below, or make it more creative!

Project final report You need to clarify each member's contributions in your final report.

— Data analysis projects: In your report, you should describe the problem, the data set, and how you tried to solve the problem. Describe the algorithms you used, the results you obtained, and discuss how confident you are in your results. Would you be willing to use them in production to change how your company or enterprise makes decisions? If not, why not?

Technically, your report should demonstrate that you tried at **least three techniques** from class on your data set, in addition to anything else you decided to do to achieve your goal. If you used techniques not discussed in class, be sure to describe how they work and provide references so that anyone reading the paper has the tools to understand it.

Your report should also include **some discussion** about whether your project might produce a Weapon of Math Destruction (as defined here) or whether fairness is an important criterion to consider when choosing a model for your application (as defined here).

— Algorithm development projects: your report should roughly follow the following outline.

- introduction: topic + question you want to answer
- important background related work papers you've read (at least 4 citations)
- your methodology: how did you try to answer the question?
- explanation and results of specific experiments you did (including a few plots or tables)
- conclusion and future work

— Theoretical projects: your report should roughly follow the following outline.

- summary of the technical paper
- review from an advocate viewpoint
- review from a critic viewpoint
- additional statements proved / Proofs of existing statements simplified / Algorithms improved (to be justified either theoretically or experimentally)
- conclusion and future work

Final peer review Instructions are the same as the previous peer review assignment. We expect your reviews to be **at least two paragraphs**, and expect that you will provide critical and useful feedback to the team you're reviewing. Think about what kind of feedback or ideas would be most useful to you, and try to give what you'd like to receive!

Submit a grade (and answer a few more specific questions about) and comments on the proposal for the project via google forms. Review assignment and link to form will be posted on our discussion forum.

7 Project Ideas

See the "Project ideas" part at this website.

8 Grading Components

- Project proposal: 10 points (Peer review: 5, TA review: 5)
- (4741) Project final report: 20 points (Peer review: 5, Instructor review: 15)
- (5741) Project presentation: 5 points (Peer review: 3, TA review: 2); Project final report: 15 points (Peer review: 5, Instructor review: 10)

9 Suggested Papers for Theoretical Projects

You *must* choose one of these papers, unless you have very good reasons to choose another paper and consulted the Instructor regarding your own choice.

1. "Boosting the margin: a new explanation for the effectiveness of voting methods", P. Bartlett, Y. Freund, W. S. Lee, and R. E. Schapire, Ann. Statistics, Vol. 26, No. 5, 1651–1686 (1998)

   Fully understand why the test error decreases long after the training error equals to 0. Suggests alternative explanations.

2. "New support vector machine algorithms", B. Schölkopf, A. J. Smola, R. C. Williamson, and P. Bartlett, Neural Computation, Vol. 12, No. 5, 1207–1245 (2000)

   Fully understand alternative SVM formulations. Compare and contrast to the standard formulation.

3. "Bayesian network classifiers", N. Friedman, D. Geiger, and M. Goldszmidt, Machine Learning, 29, 131-161, (1997)

   Compare and contrast to Naive Bayes. Suggest other structural properties of the data one could learn to improve on Tree-Augmented Networks.

4. "Learning the kernel matrix with semidefinite programming", G. R. G. Lankriet, N. Cristianini, P. Bartlett, L. El Ghaoui, M. I. Jordan, Journal of Machine Learning Research, Vol. 5, 27–72 (2004)

   Any other ways of learning kernels? Can they be framed as SDPs or other tractable convex programs?

5. "Text classification using string kernels" H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini and C. Watkins, Journal of Machine Learning Research, Vol. 2, 419–444 (2002)

   Fully understand the dynamic programming procedure therein. Any other procedures for computing reasonable similarity measures between strings?

6. "Walk-Sums and Belief Propagation in Gaussian Graphical Models", D. M. Malioutov, J. K. Johnson, A. S. Willsky, Journal of Machine Learning Research, 7(73):2031–2064, 2006.

   Fully understand the mathematical conditions for convergence of belief propagation in Gaussian graphical models.

7. "Learning with Mixtures of Trees", M. Meila and M. I. Jordan, Journal of Machine Learning Research, 1: 1–48 2000.

   Understand the algorithm and prove convergence guarantees for learning mixture of trees.

8. "Efficient Bandit Algorithms for Online Multiclass Prediction", Sham M. Kakade, S. Shalev-Shwartz and A. Tewari, International Conference on Machine Learning, 2018

   Understand the banditron algorithm and its convergence guarantees.

9. "Controlling bias in adaptive data analysis using information theory", Daniel Russo and James Zou, In Proc. of Artificial Intelligence and Statistics, pp. 1232–1240, 2016.

   Understand this alternative way of proving generalization bounds.

10. "Estimating the dimension of a model", G. Schwarz, Annals of Statistics, Vol. 6, 461–464 (1978)
    Understand the proof herein and consider other model selection criteria (e.g., MDL) and prove why they are good.