Course project – Data Mining 2024

Welcome everyone! This year, we're shaking things up a bit. Instead of the usual home assignments, we'll work on a data science project you can proudly showcase in your portfolio. We have borrowed some descriptions and guidelines from CS2229@Stanford to guide you through this project. But we've also added our unique spin, so follow the instructions in this document.

One crucial aspect of this project is that all the code (and the data) must be available on GitHub. This not only helps you keep track of your work, but it also counts for 10% of your overall grade. Your repository should include an informative README file, data and code in separate folders, and other relevant information such as figures and outputs. **A guide for working with GitHub can be found at https://rogerdudler.github.io/git-guide/**

https://www.codebrain.co.il/ניהול-גרסאות-קוד-git-מדריך-ל/

The first thing to do is pair up into groups of up to two students (you can decide to work alone or in a pair, but the grading criteria will be the same). **All students should be registered on the project spreadsheet (https://tinyurl.com/47bpk9ek) by 14/2. Registration (up until 14/2) will account for 5% of the project's overall grade.**

Let's talk about choosing a topic for your project. To get started, look at http://cs229.stanford.edu/projects.html for an idea of a project. I strongly recommend selecting an applicative project that can be presented using platforms like Gradio or Streamlit. For example, you might consider designing a unique EDA framework, focusing on the explainability of a model, or comparing multiple clustering algorithms graphically (and statistically). Alternatively, you could choose a more theoretical project, such as assessing what type of data is best learned using distance-based models or tree-based ones using a varied set of datasets. Once you've identified a topic of interest, use an academic search engine like http://scholar.google.com to look up existing research on relevant topics using related keywords. **Note that this project is part of the course "Advanced Topics in Machine Learning", and most of it should refer to topics learned this semester.**

Another critical aspect of designing your project is identifying one or several datasets suitable for your topic of interest. If the data requires considerable pre-processing to suit your task, or you intend to collect the data yourself, remember that this is only one part of the expected project work and can often take considerable time. We still expect a solid methodology and discussion of results, so be sure to pace your project accordingly.
Note that if you choose to use an existing dataset instead of crawling your own data or using an existing API, the highest possible project grade will be 90 (i.e., 10 points will be decreased from your final grade!).

Things to consider:

- **Preprocessed datasets**: While we don't want you to have to spend much time collecting raw data, the process of inspecting and visualizing the data, trying out different types of preprocessing, and doing error analysis is often an essential part of machine learning.
- **Replicating results**: Replicating the results in a paper can be an excellent way to learn. However, instead of replicating a paper, try using the technique on another application or analyze how each model component contributes to the final performance. In other words, your project can be partially novel but should not just duplicate previous work others did.

Project Proposals (due 29/2 at 11:59 PM)

You will select a project idea early on for the project proposal and receive feedback. Your proposal should be a PDF document that includes the project's title and the full names of all team members. The proposal should be a maximum of a single page, with a font size of 11.

Describe your project idea clearly and concisely, including any datasets, models, or techniques you plan to use. The proposal should demonstrate that you have put thought and effort into your project and clearly understand what you want to accomplish. Remember, the proposal is an opportunity to receive feedback on your project idea, so include any questions or concerns. We are here to support you in your project and want to help you succeed.

Your project proposal should include the following information:
Motivation: What problem are you tackling? Is this an application or a theoretical result?
Method: What machine learning techniques are you planning to apply or improve upon?
Intended experiments: What experiments are you planning to run? How do you plan to evaluate your algorithm? Presenting pointers to one relevant dataset and one example of prior research on the topic is a valuable (optional) addition.
Grading The project proposal is mainly intended to make sure you decide on a project topic and get feedback. As long as your proposal follows the instructions above and the project has been thought out with a reasonable plan, you should do well on the proposal.
The proposal’s grade is 10% of the project’s overall grade.

Final Writeup (due 20/03 at 11:59 PM)

We understand the dedication and hard work students put into their projects, which is why we take the time to review and thoroughly comprehend every write-up submitted. In the spirit of sharing knowledge and achievements, we will proudly post all final write-ups online for everyone to read and learn from. If you prefer to keep your write-up private, please notify us at least a week before the final submission deadline.

As for the format, we ask that final project write-ups adhere to a maximum of seven pages, including appendices and figures. You may add additional pages for references only. If your project involved collaboration with others or the guidance of another professor, please acknowledge their contributions in your write-up, following the Stanford report guidelines (http://cs229.stanford.edu/final-report-guidelines.pdf). Additionally, we require a presentation summarizing the project's progress from day one to the write-up submission.

We value teamwork and fairness, so please include a section detailing each team member's contributions in your write-up. **We want to see cooperation but also independent work.** If any concerns arise about your team's collaboration, please reach out to us via email. We may consider team contributions and evaluations when assigning project grades.

Lastly, please provide a link to a GitHub repository containing your final project's code. Remember to include the data and make a requirements.txt file listing the libraries used. The final report's grade will be based on its clarity, relevance to topics covered in ML and DM classes, the novelty of the problem, and the technical quality and significance of the work. **The write-up's grade is 40% of the overall project grade, with the presentation accounting for 10%.**

Oral defense (TBD, around 4/24)

Each team will have to defend their project, demonstrating knowledge in every aspect of the project. You will have to show that the code on GitHub performs exactly as was presented in the final writeup. Among others, each team member will be asked a few questions about the project and code and will be graded individually. **It is your responsibility to log in from a computer with a functional IDE and can run some of your code if asked.**

Frequently asked questions:

1. **Should the final project use only methods taught in the class?**
   We don't restrict you to only using methods/topics/problems taught in class. You can always consult me if you need clarification on any method or problem statement.

2. **Is it okay to use a dataset that is not public?**
   We don't mind you using a dataset that is not public as long as you have the required permissions to use it. Note that we are asking you to share the dataset as part of your final report.

3. **Can I use datasets from Kaggle?**
   Using datasets from Kaggle and other repositories (such as UCI) is **not allowed**!!

4. **Can this project be combined with that of another class?**
   No**.** Note that the project should not be related to your final project as well.

5. **What are acceptable team sizes, and how does grading differ as a function of the team size?**
   We recommend teams of 2 students, while a single-student team is also acceptable. In particular, we expect the team to submit a completed, so keep in mind that all projects require you to spend a decent minimum effort towards gathering data and setting up the infrastructure to reach some form of result.
   YOU ARE NOT ALLOWED TO WORK IN GROUPS OF THREE (OR MORE) STUDENTS

6. **Can I change groups?**
   Generally not. If there is a justified reason why you can't longer be part of your team, please let me know as soon as possible.

7. **Do I have to be on campus to submit the final report?**
   No, the final report will be submitted via Moodle.

8. **What is the late-day policy for a group project?**
   Each group is allocated **three** late days to divide between the different submissions (i.e., proposal and final writeup). Note that late submission applies to all group members; use these days wisely.

9. **Can we use Machine Learning libraries such as scikit-learn, or are we expected to implement them from scratch?**
   You can use any library and API for the project. Write what libraries you used in the final report and the requirements file on your repo.

10. **Can we use a public repository for version control?**
    It would be best if you used GitHub, and I must be able to access your code once you submit your reports. After the class ends, many students may want to make their work public to point to for interviews or post a blog on their project, which is acceptable.

11. **What if two teams end up working on the same project?**
    When registering on the Google spreadsheet, review the projects your friends chose and verify that no one chose the same project!

12. **Will we be provided any cloud computing resource credit?**
    Exceptional projects will be allowed to use the computer infrastructure in my lab. I encourage you to check out Google Collab (https://colab.research.google.com) for free GPU resources.

13. **Are we required to use Python for the project?**
    YES

14. **Can you repeat how the grade is divided?**
    Registration – 5%
    Project proposal – 10%
    Crawling and API – 10%
    Final writeup – 40%
    Presentation – 10%
    GitHub – 10%
    Oral defense -15%