STATS 507 – Data Science Analytics using Python

Final Project Guideline

The final project builds upon the approved proposal and aims to demonstrate your ability to execute a complete data science project using Python and Hugging Face resources. The merit of your final project will be demonstrated and evaluated through a GitHub repository which contains:

1. A well-documented executable codebase.
2. A 2-page summary report.

*An average final project should take about 40 hours.

Project Scope

- Use data analysis/science techniques taught in this class.
- Focus on practical application of Python tools.

Report Guideline

- Deadline: 04/22/25
  - The final project report should be in the format of a 2-page IEEE conference paper. [template] All formatting, including fonts, margins, columns, and citation style must adhere to IEEE guidelines. Any deviation from the IEEE template will result in point deductions.
  - Sections:

    Introduction
    - Provide background and motivation of the project.
    - State a concrete project goal.
    - Conduct a comprehensive review of existing literature, focusing on recent advancements and state-of-the-art.

    Method
    - Problem formulation: learning problem formulation (input, output), dataset description and model formulation.
    - Walk through the methodologies used to solve the problem.

    Results
    - Data pipeline or model set up.
    - Figures that present numerical simulation results.
    - Interpretation of the results.

    Conclusion

  - Tips for the final report:
    - Writing is a great tool to organize your thought. Write a draft report early and use it to lay out your project plan.
    - Start from the Method and Results sections, then the Introduction and Related Work sections. Write the Abstract last.
    - Set aside the first draft for a while and then come back with fresh perspective!
    - Revise the manuscript.

Final Submission

- Deadline: 04/22/25
- Submission Content
    - A summary report (PDF)
    - Code (link to GitHub repository in TXT)

Sample Projects:

1. **Preprocess the data and fine-tune a pretrained model on any GLUE task.**
   [Code]

   Potential Extensions:
   - Experiment with different model architectures
   - Add custom preprocessing steps
   - Implement domain-specific optimizations
   - Create novel evaluation metrics

2. **Fine-tune an image model**
   [Code]

   Potential Extensions:
   - Implement custom data augmentation
   - Add multi-task capabilities
   - Create specialized visualization tools
   - Build real-world applications

Some other working notebooks can be found here.

You might need to ensure you have the required packages installed for each notebook (use Claude or ChatGPT or go to office hours for help). You can choose to:

- Use these notebooks as starting points and build upon them
- Focus on in-depth data analysis
- Explore different model architectures
- Create entirely new implementations
- Combine multiple approaches in novel ways
- Follow UM GenAI guideline in: https://genai.umich.edu/resources/students

Remember, these examples are meant to inspire rather than limit. Feel free to explore other datasets, models, or approaches that align with your interests while demonstrating your Python data science skills.