Math 404: Portfolio Project

Winter 2023

The project should use several of the machine learning methods we have covered or will cover in class, and at least one method not covered in class, in original and thoughtful ways to give reliable, thoughtful, and nuanced answers to interesting questions about a large dataset.

Deep learning methods, while useful for many problems, are not the focus of this project, nor will we cover them sufficiently before the project due date. For this reason *none of the machine learning methods used in the project should be deep learning methods.*

The main focus of the project should involve a question about sequential or time-series data. That could include traditional time series like sea surface temperatures over time; covid infection, hospitalization, or death rates; electrocardiogram data; sleep patterns; Fitbit data; or birth rates over time. But it could also include things like music, language, or DNA sequences.

This project should incorporate substantially more thought and effort than just plugging pre-cleaned Kaggle data into scikit-learn and getting a few numbers out. Rather, it should represent significant original thought and deep analysis, showing a mature application of the ideas and techniques you have learned throughout this year and the rest of your ACME experience. Typically this will require approximately 20–30 hours of active work per team member. Every team member should be involved in most aspects of the project, and every team member is responsible for the final product.

The finished project should be something you are proud to show to others, including current and prospective employers. It should make people want to hire you, admit you into their graduate program, and fund your startup idea.

Structure of the Final Project Report

Your written report on the project should be written in LaTeX and should have the following main sections:

Abstract A one-paragraph abstract (tl;dr) is required. Summarize the main question and conclusions. Put this below the title but before the main text, using the commands `\begin{abstract}` and `\end{abstract}`.

Problem Statement and Motivation Give a clear statement of the problems or questions the project addresses, their context, and a compelling motivation for why they are worth studying. The problems or questions your project addresses should be original, meaningful, and reflect a deep understanding of the subject and its subtleties.

Briefly review, with proper citations, what is already known about the research questions and what techniques others have used to study these questions. Explain the scope of the project and how it fits into existing research.

Data Discuss the sources, reliability, and suitability of your data for the problems you are addressing. The dataset should be large enough and rich enough to give reliable, meaningful, and nuanced answers to the questions and problems addressed in your project. Be clear about how you handled splitting data into test, training and validation sets, and how you ensured that there was no leakage between test and training set and no other similar data problems. Be clear about how you handled missing data. Justify your choices on all these things.

Methods Your project should involve a thoughtful and original use of several of the machine learning methods and ideas we have learned in class and at least one idea or method we have not covered in class. In this section you should describe and justify your selection of models, your choices of methods, any feature engineering you did, and the ways and reasons you chose your hyperparameters or network architectures. Your discussion should demonstrate a clear understanding of the principles involved and of the strengths and weaknesses of the models and methods used.

Results Clearly and succinctly describe your results.

Analysis Give a thorough analysis and a thoughtful discussion of the results and conclusions that can be drawn. Discuss the suitability and effectiveness of the different models and methods for the problems or questions treated.

Ethical Considerations Thoughtfully analyze the ethical implications of your research questions, the data you gathered, and the analysis that was performed. Are there privacy or other implications from the collection or use of the data? Could your results and methods be misused or misunderstood? What can and should be done to prevent misuse and misunderstanding? Could your algorithms and methods result in a destructive self-fulfilling feedback loop? How could that be prevented or controlled? What other ethical implications does your work have?

Conclusion Describe the final conclusions that you draw from your computations, results, and analysis.

Additional Specifications

The following specifications can affect your score in each or all of the categories listed above.

Format and Length

The final project should be submitted on LearningSuite as a single PDF document written in LaTeX.

The main text should be **no more than 10 pages**, not including the bibliography, as measured in the default layout and font. Code and other important supplemental information may be attached in an appendix if necessary. Here are some rules for your LaTeX code:

1. The raw LaTeX should start with the following line

```
\documentclass[11pt]{article}
```

2. Do not change the margins, font, or font size. Note that Overleaf often adds something like the following to your files

```
\usepackage[letterpaper,top=2cm,bottom=2cm,left=3cm,right=3cm,
                    marginparwidth=1.75cm]{geometry}
```

That will change the margins. Delete that whole line. Don't use the geometry package at all.

3. Use the `hyperref` package to ensure your URLs are active.

4. Use the `\title` and `\author` commands, along with `\maketitle` to create the header with the title and authors' names.

5. Use the standard LaTeX bibliography tools, including the `\cite` command to cite references in the text. BibTeX works pretty well for this.

Graphics

Your report should include appropriate graphics and data visualizations to help illustrate the important findings. Graphics should communicate the important information effectively and efficiently, applying the principles taught in the various data visualization labs we have covered in ACME. They should be of high quality (300 or more dpi) and not be rescaled from the size the were when you made them. They should be attractive to look at and easy to read and understand. Points will be taken off for graphics that are pixelated or hard to read. Points will also be taken off for the use of pie charts, and double points will be taken off for 3D pie charts.

Figures should have substantial captions that clearly explain what the figure is and all the important information necessary to understand it. Axes and subplots should be clearly labeled.

Writing

Your writing should be clear, precise, and succinct, using all the key principles taught in your technical writing classes. We strongly encourage you to visit the university writing lab with an early draft of your report to get help with your writing.

Grading

Points will be given (approximately) according to the following rubric:

10 Questions: Motivation and Overview

10 Data: Quality and Discussion

20 Methods: Suitability and Justification

25 Analysis: Depth and Quality

15 Communication: Quality and Clarity

10 Graphics: Effectiveness and Quality

10 Ethics: Thoughtful consideration of all the implications

10 Knock my Socks Off: Impress me and everyone who reads it