CIS 635 Project Progress Report

Jacob Morrison

Due: 7 November 2025

- **Project Name:** Packing the Bits
- **Group Member(s):** Jacob Morrison
- **Project Title:** Investigating Tissue Differences in the Human Fallopian Tube across Ovarian Cancer Subtypes

1 Project Progress Overview

1.1 Completed Tasks

To date, I have acquired the raw data, finished reading the data with Python, created the mock dataset that will be made available for grading, normalized and handled missing data, calculated descriptive statistics on the data, and calculated a dissimilarity matrix. I have also started working on putting together plots for the descriptive statistics and the dissimilarity matrix, though those will need more work to improve their aesthetic and utility.

1.2 Data

In order to save space on the GitHub repository, the mock data files have not been uploaded. Instead, the Python script I used to generate the data is available within the `data` directory on GitHub: `https://github.com/GVSU-CIS635/projects-packingthebits/tree/main/data`. The script requires Pandas and a BED file as input (see Line 9 of the script for instructions on how to download the BED file). A seed is set for the random number generator, so the data generated by the user is the same as what I generated (at least to first approximation for the `random` library in Python).

2 Challenges

2.1 Encountered Difficulties

The first difficulty I encountered was how to provide the mock data. When creating this dataset, I wasn't sure how to easily provide the files that were generated. While there weren't many files, their size was non-trivial and would require them to be hosted somewhere, whether on GitHub through a version release of the project or on Dropbox. The second difficulty I faced was the dissimilarity metric to choose. There are plenty of good options to choose from (Manhattan distance, Euclidean distance, Minkowski distance, etc.). However, none of these guarantee a normalized value between 0 and 1, which is something I was hoping to have for an easier comparison between different samples.

2.2 Challenge Solutions

To solve the issue with the mock data, I decided to create a Python script that generated fake data using a Monte-carlo algorithm. As mentioned in Section 1.2, by fixing the seed for the random number generator, I'm able to reasonably assume a repeatable dataset will be generated across multiple runs of the script. Thus, the user can run the script themselves to create the dataset I worked with on their own computer without having to download the data.

With respect to my choice of dissimilarity metrics, I chose the cosine similarity metric, rather than one of the standard distance metrics. In order to turn the similarity metric into a dissimilarity metric, I used the following equation:

$$1 - \frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| \times ||\vec{y}||} \tag{1}$$

3 Collaboration

As a group of one, I don't meet with anyone on a regular basis. Instead, I try to set aside a day each week to work on my project in order to make meaningful, consistent progress towards completing the project.

4 Next Steps

My next steps are to complete the steps outlined in Weeks 12–15 of Table 1. As mentioned in Section 3, I work on the project each week and am right on schedule, so I plan to complete each task on the week it is listed through the end of the semester. One potential challenge that may come up is setting up and running the machine learning model. Depending on the complexity of the model, it may take a while to create a trained model with which I am happy, which may cause this goal to bleed into the final week or two and impact the amount of time I have to create the final report and presentation.

5 Timeline

As a reminder of the proposed timeline of the project, see Table 1.

Week	Goals	Completed
7	Finalize and submit proposal	Yes
8	Read data with Python, Create mock dataset	Yes
9	Normalize, handle missing data	Yes
10	Calculate descriptive statistics	Yes
11	Calculate dissimilarity matrix	Yes
12	Perform clustering analysis	No
13	Set up and run ML models	No
14	Write up final report	No
15	Turn in	No

Table 1: Project timeline and the status of each step from the midterm until the end of class.