Computational Microbiome Analysis Workshop – Final Project Instructions

Project description

The objective of this project is to predict the presence of diseases based on gut microbiome composition using an imbalanced dataset. The dataset includes multiple cardiometabolic conditions, and the training data follows a case-control design, meaning it contains a higher proportion of sick individuals. However, only 5%-10% of the individuals in the training set will be sick, making it crucial to address class imbalance effectively. Your task is to develop a pipeline that processes data and outputs the probability that each individual has a disease when the test dataset is imbalanced, having 5%-10% cases of a certain disease.

The project will be completed in groups of 2-3 students. If you are looking for a partner, you can use the relevant forum on Moodle. For more details about the dataset, please refer to the relevant documents on Moodle. For more information regarding submission requirements, see below

Milestones

Throughout the course, there will be three milestones to ensure that progress is being made and to provide a formal opportunity for group discussions. Prior to each milestone meeting, you need to prepare a short presentation (10-20 minutes) that summarizes your progress and future plans. meetings will be scheduled via email. Please send your PowerPoint presentation at least 24 hours in advance.

Minimum progression expected and deadlines for each milestone is as follows:

	Deadline	Minimum progression expected
Milestone 1	22/5/25	Presenting preliminary exploration results and the main ideas that will be implemented in the pipeline. See below guiding questions for data exploration
milestone 2	21/8/25	Presenting preliminary results. To present preliminary results, it is recommended to start with an overview of the current pipeline. Then, show the model's performance in various of disease-control conditions and proportions, focusing on 5%-10% imbalance

Milestone 3	2/10/25	Presenting the final results, including full pipeline, validation methods, their main results and possible future directions. **Model predictions should be sent 3 days before Milestone 3 meeting.** During the meeting we will present to you the model performance on the test set.

Guiding questions for data explorations (milestone 1)

These questions are meant to serve as a starting point for data exploration, and you are not required to answer all of them (except for the naïve modeling, which will serve as a baseline in your future work). Be sure to explore additional questions that could further help you develop your pipeline.

1. What is the distribution of diseases in the metadata, and how do these categories compare in terms of sample size and demographics?
2. How do microbiome composition profiles differ between disease groups and healthy controls?
    a. Refer to alpha-diversity differences
    b. Calculate Bray-Curtis distances between all microbiome samples, plot PCoA and calculate average distance between groups.
    c. Could you detect specific bacteria that are differentially abundant in cases compared to control?
3. What are the key trends in the metabolome data across different disease states? Are certain metabolites consistently elevated or reduced?
4. How are the microbiome and metabolome data correlated? Which microbial taxa show strong associations with specific metabolites?
5. How might confounding factors (such as age and gender) influence the observed patterns in the microbiome-metabolome data? Do they correlate with both microbial diversity and metabolomic profiles
6. Naïve modeling (mandatory): Predict disease status using microbiome data with a random forest classifier and evaluate performance using AUPR. Then, assess whether incorporating metabolomic data improves prediction accuracy.

Final submission guidelines

Final submission includes: 1. Report 2. GitHub with your code 3. CSV with predictions on test set.

Report

A detailed pdf document describing the problem, the pipeline, validation and validation results and possible future directions. The paper should be 3 pages long at most with up to 4 figures (not including bibliography). It should have the following section:

- Introduction – describing the problem, and a general description of the solution's approach, pipeline and a summary of the findings.
- Methods – detailed description of the pipeline and validation.
- Results – detailed description of your findings and the validation results. You may include the results of experiments you have done even if the results were not so good, and you excluded them later.
- Discussion – summary of the results, their importance and future possible directions.
- Bibliography – we recommend using either Mendeley or Zotero to insert bibliography.

Creating figures to visualize the results, pipelines, etc. is strongly encouraged. To easily create figures in Python we suggest using either seaborn, plotly or other similar Python libraries. **The report should be sent a week after milestone 3 meeting.**

Code

Code should be submitted in GitHub. Keep the code clean, well-documented and easy to read.

Predictions on test data

Before submission, you will receive three test data files:

- A metadata file – identical to the training metadata file, except " **PatientGroup**" column, that will be excluded.
- A microbiome file and a metabolome file. These files will follow the same structure as the training set.

(output.csv) should be a csv with two columns separated by a comma. The first column should include the sample ID and the second column should include the probability that this sample

belongs to an individual with a disease. The first row of the file should be "ID,Probability ". use the following example output as reference:

Predictions should be sent three days before milestone 3 meeting. During the meeting we will present to you the model performance on the test set.

Grading

The final grade will be determined based on the model's AUPR, milestone meetings, as well as the completion and overall quality of the final project. Creativity and effort invested in the research project will be rewarded, with grades of 96-100 reserved for exceptional effort, innovative ideas, or particularly rigorous research.