Machine Learning Manuscript Outline

Introduction

General Context of the work

As gut microbiome field continues to grow, there will be an ever-increasing demand for reproducible machine learning methods to analyze microbiome sequence read count data and to determine association with a continuous or categorical phenotype of interest.

Narrower research area and statement of its importance

Colorectal cancer is one of the leading cause of death among cancers in the United States. Early diagnosis increases the chance of survival. However the current diagnostic methods are expensive and invasive. As a less invasive tool, numerous studies use relative abundances of the gut bacteria populations to predict disease progression.

Identification of a gap or other need for research

- Do we know what machine learning methods are available to us as microbial ecologists interested in gut microbiome?
- Each study that models colorectal cancer and microbiome association uses different machine learning pipelines to classify disease progression which negatively impacts the reproducibility, robustness and actionability of models.

Summary of appraoch to answer the research question

- We perform L2-regularized logistic regression, SVM, Decision tree, Random forest, XGBoost and Feed Forward Neural Net classification models.
- We evaluate the classification success of different machine learning methods. We also want to discuss the reproducibility, robustness, actionability, interpretibility and susceptibility to overfitting of each method.

Announcement of principal findings

- Is there a maximum threshold of prediction with all these methods?
- What gives the best AUC value?
- Does an increase in model complexity improve predictibility?
- Synthesis statement regarding modeling 16S microbiome data

Methods

Brief explanation of study design/patient sampling and 16S rRNA gene sequencing/curation

- Refer to Baxter Dataset: We sequenced the 16S rRNA genes from the stool samples of 292 patients, 120 with colorectal cancer and 172 with healthy colons. We used the relative abundances of the bacterial populations within each sample to develop classification models that detects colonic lesions using the relative abundance of gut microbiota.

Analysis of data

- Pre-proccessing of the data
- Machine Learning pipeline backbone. How do I split, train, validate and test?
- Cross-validation and hyper-parameter tuning methods for each modeling method
- Programming languages and packages/modules utilized
- Statistical methods for comparison of model performance

Results

Results of modeling in text, tables and figures

Comparisons among modeling approaches

Discussion

Interpretation of modeling results in terms of reproducibility, robustness, actionability, interpretibility and susceptibility

Consideration of possible weaknesses for each model

Consideration of possible weaknesses for our approach and chosen dataset

Relationship of results to previous literature and broader implications of this work

Prospects of future progress

Notes from Pat: For the general context, I think it's very important to hit a few things hard…

1. Most microbial communities are pretty patchy and the likelihood of a single feature that explains the differences in health is pretty small. It is likely that many biomarkers are needed to account for the patchiness as well as the context dependency of the features. So tools like linear models (e.g. metastats, lefse, wilcoxon, etc) are likely worthless.

2. The interactions between the biomarkers may be nonlinear. Obviously, the linear models will not incorporate this because they are linear.

3. ML use in microbiome literature is a bit like the wild west with lack of clarity over methods, testing, validation, etc. There is a need for guidance on how to properly implement these different methods

4. There is likely to be a tradeoff between interpretability and performance (perhaps this should go in a paragraph about ML methods that would also indicate which methods are linear/non-linear.