Bioinformatics Research Project Log

Chandrashekar CR
Supervisor: Dr. Eran Elhaik
Lund University

March 31, 2025

Contents

March 31, 2025

Tasks for the Day

- Understand the mGPS algorithm from the R code and implement the preprocessing steps.
- Set up the working environment and install all required libraries.
- Begin tracking the project.

Notes and Observations

- I am currently using the `ai_env` environment from my previous projects and accessing resources on the bioinformatics server.
- I am still struggling to understand the R code, but I expect to progress faster once I grasp the underlying logic.

April 1, 2025

Tasks for the Day

- Understand and implement Recursive Feature Elimination followed by the XGBoost machine learning algorithm with the correct hierarchical steps.
- Initialize Git for version control and push code to GitHub.
- Gain a deeper understanding of cross-validation.

Notes and Observations

- I have understood the general workflow and added comments to the MetaSUB preprocessing script in R.
- I learned about cross-validation and its importance, though I have not yet implemented it.

Pending Tasks

- Git tracking for the project has not been set up yet.
- Preprocessing steps are still in progress.

April 2, 2025

Tasks for the Day

- Ask Vignesh how to access the LUNARC server.
- Reimplement the XGBoost machine learning algorithm with the correct hierarchical steps from the previous day.

Notes and Observations

- I initialized Git for the project and started tracking all files except data files and research papers.
- The `metasub_global_git.csv` file contains all the Geographically Informative Taxa (GITs) required for the XGBoost model.
- I realized that I do not need to recreate the exact model; my primary objective is to understand how the input data for XGBoost is preprocessed.
- The key questions to answer: What is the shape of the input data? What are we predicting?

Pending Tasks

- Gain access to the LUNARC server.

April 3, 2025

Tasks for the Day

- Implement basic neural network architectures in PyTorch. The dataset contains n data points, but hyperparameter tuning has shown that 200 GITs are sufficient for accurate predictions.
- Integrate Ray parallel processing into the hyperparameter optimization process, reducing the search time (currently estimated at 4-5 hours).
- Preprocess the data into numerical format by converting categorical variables (continents and cities) using one-hot encoding or label encoding. One-hot encoding seems to be the safer option.
- Implement stratified K-fold cross-validation before proceeding with a simple neural network model.

Notes and Observations

- I successfully logged into the LUNARC server's login node, but I still need to figure out how to access and use the GPUs for training neural networks.
- Authentication and login are complete, but I still need assistance in understanding the following:
  - Where my allocated storage is located.
  - How to submit jobs using SBATCH.
  - I need to learn the basics of working on HPC?
- After performing Recursive Feature Elimination (RFE), the final dataset was created with a shape of 4070×204. Out of the 204 columns, 200 represent the selected features, while the remaining 4 correspond to the target variables.
- Each row in the dataset contains 200 features representing the relative sequence abundance (RSA) of microorganisms. The 4 target columns include the continent, city, latitude, and longitude of the sample collection site.
- The dataset comprises samples collected from 40 unique cities across 7 continents.
- Categorical variables (continent and city) were encoded using `sklearn`'s `LabelEncoder`, while latitude and longitude were standardized using `StandardScaler`.
- Initially, stratified cross-validation was considered for training the neural network. However, this approach was temporarily replaced with `train_test_split` for initial model development. Once the basic neural network is functional, stratified cross-validation will be reconsidered for improved model performance.

Neural Network Architecture

- The initial model will be a simple feedforward neural network, designed to follow a hierarchical approach similar to the previous study using XGBoost.
- The next step will involve experimenting with single-layer architectures before exploring more complex models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).