Bioinformatics Research Project Log

Chandrashekar CR
Supervisor: Dr. Eran Elhaik
Lund University

April 10, 2025

Contents

March 31, 2025

Tasks for the Day

- Understood the mGPS algorithm from the R code and implemented the preprocessing steps.
- Set up the working environment and installed all required libraries.
- Began tracking the project.

Notes and Observations

- Utilizing the `ai_env` environment from previous projects and accessing resources on the bioinformatics server.
- Initial challenges in understanding the R code are anticipated to decrease with further engagement.

April 1, 2025

Tasks for the Day

- Understood and implemented Recursive Feature Elimination followed by the XGBoost machine learning algorithm with the correct hierarchical steps.
- Initialized Git for version control and pushed code to GitHub.
- Gained a deeper understanding of cross-validation principles.

Notes and Observations

- Achieved a general understanding of the workflow and added comments to the MetaSUB preprocessing script in R.
- Acquired knowledge regarding the importance of cross-validation, although implementation is pending.

April 2, 2025

Tasks for the Day

- Acquired information from Vignesh regarding access to the LUNARC server.
- Determined that reimplementing the exact XGBoost model is unnecessary; the focus is on understanding the input data preprocessing.

Notes and Observations

- Git repository initialized for the project, tracking all files except data and research papers.
- The `metasub_global_git.csv` file contains the Geographically Informative Taxa (GITs) required for the XGBoost model.
- The primary objective is to comprehend the preprocessing of input data for XGBoost.
- Key questions identified: What is the shape of the input data? What are the prediction targets?

April 3, 2025

Tasks for the Day

- Implemented basic neural network architectures in PyTorch. Hyperparameter tuning indicates that 200 GITs are sufficient for accurate predictions, despite the dataset containing n data points.
- Integrated Ray parallel processing to optimize hyperparameters, aiming to reduce the estimated 4-5 hour search time.
- Preprocessed data into numerical format by converting categorical variables (continents and cities) using one-hot encoding.
- Deferred the implementation of stratified K-fold cross-validation for later.

Notes and Observations

- Successfully logged into the LUNARC server's login node, but GPU access and utilization for neural network training require further investigation.
- Authentication and login to LUNARC are complete; however, assistance is needed to understand:
  - The location of allocated storage.
  - The process of submitting jobs using SBATCH.
  - The fundamentals of working on High-Performance Computing (HPC).
- Following Recursive Feature Elimination (RFE), the final dataset has a shape of 4070×204, with 200 features and 4 target variables.
- Each data point comprises 200 features (GITs) representing the relative sequence abundance (RSA) of microorganisms. The 4 target variables are continent, city, latitude, and longitude.
- The dataset includes samples 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`.
- Initial consideration of stratified cross-validation was temporarily replaced with `train_test_split` for initial model development. Stratified cross-validation will be revisited for enhanced model performance.

Neural Network Architecture

- The initial model is a simple feedforward neural network, inspired by the hierarchical structure of the previous XGBoost study.
- The first version includes an input layer (200 nodes), two hidden layers (400 nodes each), a smaller hidden layer (2 nodes), and an output layer (7 nodes for the 7 continents).
- The plan is to initially train this model to predict the continent. Subsequently, the predicted continent probabilities will be concatenated with the original 200 features as input for a second neural network (similar architecture) to predict the city.
- The optimal handling of latitude and longitude values within the neural network remains an open question.
- Long training times due to CPU-based computation on the bioinformatics server are a significant challenge. GPU access on the LUNARC cluster is required.

April 4, 2025

Tasks for the Day

- Finalized the presentation for the weekly lab meeting.
- Focused on obtaining GPU access and understanding HPC architecture.
- Initiated preprocessing for marine and soil datasets, aiming for modular script design applicable to all datasets.

Notes and Observations

Explored various neural network implementations and gained initial understanding of HPC principles.

April 7, 2025

Tasks for the Day

- Developed a functional neural network to predict all target variables (continent, city, latitude, and longitude).
- Completed preprocessing steps for marine and soil datasets.
- Began modularizing code for efficient execution on the HPC.

Notes and Observations

- Developed a working script for MetaSUB data processing, with pending error handling and file format validation.
- Created a working script to extract relevant features using Recursive Feature Elimination with a Random Forest base model.
- Initiated work on the HPC, understanding basic operations and starting to modularize scripts for GPU compute nodes.

April 8, 2025

Notes and Observations

- Started building multiple neural network models.
- Began learning batch scripting using SLURM.

April 10, 2025

Notes and Observations

- Exploring alternative scaling methods for latitude and longitude, including conversion to radians and trigonometric transformation into a two-dimensional space.

April 11, 2025

Tasks for the day

- These are some of my thoughts. The currrent approach is always making a new neural network model from scratch for each dataset. A final model should be made that utilizes all the iterations done and must work for all type of datasets regardless of the layers of predictions.
- For example, the MetaSUB dataset contains informaiton on the continent, city, latitude and longitude. Whereas the marine dataset contain information only the sea, latitude and longitude. There should be a way that can handle these cases instead of definfing a newtwork from scratch.
- Finish the logic for the latitude and longitude neural network model.
- Compare the three models with metasub dataset on accuracy, confusion matrix, plot on world map.