CS690A: Assignment - Predicting the Drug Resistance in *Mycobacterium tuberculosis*

Hamim Zafar
hamim@iitk.ac.in

Indian Institute of Technology Kanpur — September 22, 2020

Introduction

Multidrug-resistant tuberculosis (MDR-TB) poses a significant public health challenge globally. The conventional approach of culture based antimicrobial susceptibility testing requires very long time (weeks to months) to determine the resistance/susceptibility of a certain TB isolate. As an alternative, molecular diagnostic tests (GeneXpert MTB/RIF, Hain line probe assay (LPA), etc.) are used but such tests rely on a small number of genomic positions and do not profile the rare gene mutations of the targeted loci (plural of locus - genomic position). Whole genome sequencing (WGS) of the TB isolate captures both known and rare mutations that may contribute to the drug resistance. From whole genome sequences mutations are inferred and the mutations are used as features for classifying an isolate as susceptible or resistant. The goal of this assignment is to come up with a machine learning/statistical model that improves the predictive performance of WGS for a number of first and second-line drugs.

Info: To learn more details on the problem, read the following papers

- Yang, Yang, Katherine E. Niehaus, Timothy M. Walker, Zamin Iqbal, A. Sarah Walker, Daniel J. Wilson, Tim EA Peto et al. "Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data." Bioinformatics 34, no. 10 (2018): 1666-1671.
- Deelder, Wouter, Sofia Christakoudi, Jody Phelan, Ernest Diez Benavente, Susana Campino, Ruth McNerney, Luigi Palla, and Taane Gregory Clark. "Machine learning predicts accurately Mycobacterium tuberculosis drug resistance from whole genome sequencing data." Frontiers in Genetics 10 (2019): 922.

1 Dataset Description

The assignment contains two training and two test datasets as described in the following

1.1 Dataset 1

The first dataset contains a training data consisting of 3393 MTB isolates and a test data consisting of 1000 MTB isolates from which the mutations have been profiled. Each data point contains the features which can be considered as resistance predictors. The features used for prediction consist of two groups. In the first group, each mutation is considered a predictor and its status is binary (either present (1) or absent (0)). The second group contains 'derived' categories that groups the rarer mutations by gene locus. These features are also binary indicating the presence or absence of one of the mutations in the group. There are total 222 features. Last 56 are in the second category. For the isolates with missing data for a specific feature, we use a status of -1.

The labels for the training data include the resistance status for eleven drugs: first-line drugs (rifampicin, isoniazid, pyrazinamide, and ethambutol); streptomycin; second-line injectable drugs (capreomycin, amikacin, and kanamycin); and fluoroquinolones (ciprofloxacin, moxifloxacin, and ofloxacin).

The labels are classified as resistant (1), susceptible (0), or not available (-1). For the test data, you need to predict the resistance status of the isolate for each of these 11 drugs. The associated files are as follows

- `X_trainData_1.csv` : training data points
- `Y_trainData_1.csv` : labels for the training data points corresponding to 11 drugs
- `X_testData_1.csv` : test data points

1.2 Dataset 2

The second dataset contains a training data consisting of 900 MTB isolates and a test data consisting of 200 MTB isolates. Each data point has 89 features which are binary and indicates the presence (1) or absence (0) of a mutation. For the training data, the labels include the resistance status for rifampicin. The labels are classified as resistant (1) and susceptible (0). For the test data, you need to predict the resistance status of the isolate for rifampicin. The associated files are as follows

- `XY_trainData_2.csv` : training data points, the features and the labels are in the same file, the last column contains the label
- `X_testData_2.csv` : test data points

2 Tasks

You need to come up with a machine/statistical learning model to predict the resistance status of the TB isolates from the mutation-based features derived from WGS. You will train your model on the training dataset and then submit the prediction on the test data for evaluation. The leader board for the challenge will be maintained based on the performance on the test data. Consider the points below when preparing your solution.

- Dataset 1 has labels for 11 drugs, so this problem can be posed as a multi-task classification problem. You can try to develop a single model which can simultaneously perform all the classification tasks. You can also go for a single-task classification.
- Notice that there is a class imbalance. You need to come up with appropriate strategy to deal with this problem while developing your solution.
- In addition to trying out different ML models, you should also experiment with the feature set. The given feature set consists of 222 and 89 features (for datasets 1 and 2 respectively) each of which measures the mutation status of a particular locus. You can also consider pair of mutations as a feature, mutual exclusivity of two mutations, etc.
- You can consult the research papers mentioned above for directions regarding ML models. However, you should not just use the models used in the papers. Marks will be deducted if you restrict yourself only to models reported in the papers.
- Try to be as creative as possible in solving the problem. It involves a research question and if the solutions you submit are better than the state-of-the-art, we will write a research paper in which you will be a co-author.

For each dataset, you will perform the prediction on the test data and the predictions need to be submitted in csv file format. The format of the csv file should be as follows

Listing 1: Format of csv file for dataset 1

```
id,RIF,INH,PZA,EMB,STR,CIP,CAP,AMK,MOXI,OFLX,KAN
1,0,0,1,0,0,1,0,0,1,1,1
2,0,0,0,0,0,1,1,0,0,0,0
3,1,1,0,1,1,0,0,0,0,0,1
```

Listing 2: Format of csv file for dataset 2

```
id,rifampicin
1,0
2,1
3,0
4,1
```

Notice: In case we require a change in the format of the csv file, we will notify you. Keep an eye on the announcements.

Kaggle Leader board: You can submit the csv files multiple times and check your performance on the test data. We are trying to setup a Kaggle competition for this assignment and once that is done the link will be shared. In case the Kaggle competition platform cannot be used due to data privacy issue, we will come up with an alternate arrangement. Keep an eye on the Piazza and hello.iitk course website for an announcement regarding the submission of the csv files.

3 Deliverables

The deliverables for the assignment are the following

1. Prediction on the test data for datasets 1 and 2. These results will be evaluated and the leader board will be maintained based on the scores in evaluation
2. Runnable code (in Jupyter Notebook) for the ML models you have developed
3. Scripts for running your code to generate the predictions on test data. TAs will run these scripts to reproduce the csv files you submit for the challenge
4. A short-writeup describing the steps taken to solve the challenge. Describe in brief the models you have used, any extra feature set you have developed, training process, training accuracy, etc. The writeup should also contain a section describing the contribution of each member in the team. The writeup should mention the names and roll numbers of the team members.

For submission, all the deliverables should be zipped in a single file and the zip file should be named as `Group_i_CS690_MDRTB_assignment.zip`, `i` should be replaced by your group number. Also, each file in the zip folder should start with the phrase `Group_i_` (`i` replaced by your group number). The file should be emailed to the instructor with the associated TA copied in the email. You can find the allotted TA for your group from the excelsheet (`https://docs.google.com/spreadsheets/d/1q0CYMpIusIgupx3B_cn2NbHHNZd5_WPbvCAlhPHHjdc/edit?usp=sharing`) that contains the group info. The subject line of the email should mention the group number, [CS690A] and the phrase 'MDRTB assignment'.

4 Submission Deadline

October 15^{th} 11:59 PM.