Drug Resistance Prediction Models Summary

Belina Jang

September 29, 2024

Data Preparation First, I load XX (predictor 1246 x 228; mutation information of each sample(patient)'s HIV virus) and YY (outcome 1246 x 5; resistance information of 5 drugs in NRTI class). The data YY contains 5 columns for each drug and value (cell) represents resistance to the drug. I first defined the cutoff to determine drug resistance. I assigned 1 when the patient is not resistant (value smaller than cutoff) to the drug and 0 when the patient is resistant to the drug (value bigger than cutoff). After merging XX and YY (now binary), each row in the data represents the values of 228 mutation variables (predictors) values and 5 response variables of resistance information for 5 different drugs.

Training the Model The aim is to train models to predict the response (1: non-resistance, 0: resistance) for each of 5 drugs using 228 mutation variables (predictors). For simplicity, I will explain the process for building and comparing logistic regression, LDA and KNN ($K = 2, \ldots, 10$). I trained the models separately for each drug, assuming outcome variables (YY) are independent. I divided the data into training sets and test sets while making sure both sets have a similar proportion of resistance and non-resistance data. In order to achieve this, I used a 5-fold stratified cross-validation method, sampling the index for stratified sampling by randomly assigning values between 1 to 5 to the resistance and non-resistance data groups separately. Since I used 5 folds, around 20% of each data was used for testing, and 80% of each was used for training. Then I used the train data to train models.

Testing the Model To compare and estimate the performance of our models, I used the test data and made predictions then calculated misclassification rate, precision, recall and F1 score for each model. I repeated these training and testing processes 50 times to assess the uncertainty of the models.

Comparing the Models I first determined the best KNN model among $K = 2, \ldots, 10$ with the highest F1 score. Then I compared LDA, logistic regression and the best KNN ($K = 3$) by plotting the performance of each model in 50 experiments. I also created three separate box plots for pair-wise comparison of F1 scores with the p-values calculated from wilcox.test among the three models. In conclusion, the best model, LDA, was consistent on all drugs.