Read Me Document

The data sets we have used in our Respiratory viral DREAM challenge 11: Ensemble of predictions project are as below:

Data:

Training data:

ViralChallenge_training_CLINICAL.tsv

ViralChallenge_training_EXPRESSION_RMA.tsv

Leaderboard test data:

ViralChallenge_test_SYMPTOMATIC_SC2.csv

ViralChallenge_test_Phase1_CLINICAL.tsv

ViralChallenge_test_Phase1_EXPRESSION_RMA.tsv

ViralChallenge_test_Phase2_CLINICAL.tsv

ViralChallenge_test_Phase2_EXPRESSION_RMA.tsv

Independent test data:

RespiratoryViralChallenge_IndependentTest_RSV_Time0_CLINICAL.tsv

RespiratoryViralChallenge_IndependentTest_RSV_Time0_EXPRESSION_RMA.tsv

RespiratoryViralChallenge_IndependentTest_RSV_Time24_CLINICAL.tsv

RespiratoryViralChallenge_IndependentTest_RSV_Time24_EXPRESSION_RMA.tsv

R code for 5-fold cross validation: CV.R
Assessment code: scoring.py

Acquiring the Predicted Probabilities

We run the code for T0 that we have (3 models) to get predicted probabilities for each sample id, for each model through Xiao's CV

Model 1: ViResPred
predictions from https://www.synapse.org/#!Synapse:syn7208407

install.packages("ada", repos = "http://cran.us.r-project.org")
install libraries: e1071, ada, randomForest and class from Rstudio Packages

Shuffle: First we shuffle our set for training
We used a fixed seed number “set.seed(588)” in our code.
We randomly shuffle the data
Then we break the dataset into 5 folds (k-fold, k=5)

Feature selection:
Input:
file.name : data frame

Return:
a vector of p-values
We use Feature selection to find the lowest p-value

Model:
it gets train and test data as input to train the model and return the prediction result
We have used the Lasso Regression, KNN and SVM methods to build models for generating our Ensemble model.

We then performed 5 fold Cross Validation
#==========kfoldCV==========
Input
data: dataframe contains predictors and target variable
round: number of rounds we want to repeat cross-validation

Model:the function name of the model:
logistic, svm.run, ada.run, forest, knn.run e.g. logistic(train, test)
K: number of folds for crossValidation
X: number of sample from each of y=0 and y=1

Return
average true_auroc & average true_auprc

Bagging- We perform the Bagging Ensemble method to prepare our models
    ## UnderSampling of both class

feature selection
    ## Define the p-value threshold

Train classification model on data using selected features, then print and save predicted and actual values to compare our model and score it.

Main Function
read the result file and data file, and combine them to one table
#create the prediction probability table for the models