COMP 421 – HOMEWORK 07

REPORT

Özgür Taylan Özcan

Read data:

Initially, I read data from the given csv files and placed them into X_train, X_test and y_train data structures.

Preprocess & PCA

I applied preprocessing on the data. I used caret library for this purpose. This preprocessing includes centering, scaling, PCA and eliminating features with zero variance. I set the threshold for PCA as 0.95. This means that the selected features explain 95% of the variance. With the help of PCA, the feature size is decreased from 142 to 112. This improved the efficiency of my algorithm.

Implementation of Training and Prediction

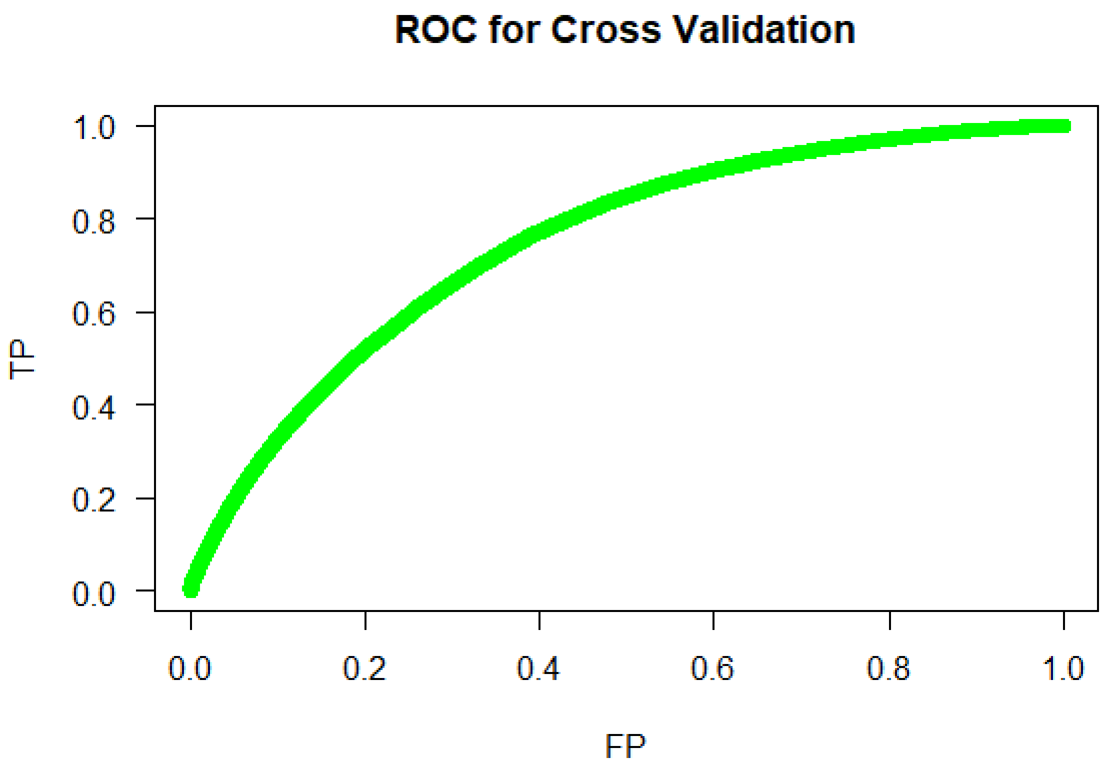
I sticked to the “No free lunch theorem” and tried a variety of different classification algorithms and their mixtures. I tested performances of each algorithm (using cross validation). At the end, I decided to implement a combination of multiple learners with different weights. I selected 3 base learners, namely logistic regression (stats library), decision tree (tree library) and LDA (MASS library). I picked them because they were time-efficient, and they made good AUROC values compared to others. I assigned weights (0.25, 0.35 and 0.40) to each of them, after testing some different weight combinations.
Each of these base learners returns probability predictions. Then, prediction vectors of all learners are summed up after being multiplied by their weights. The resulting vector contains final prediction values for the given data.

Implementation of Cross Validation

I implemented two different cross validation methods. These are K-fold Cross Validation and 5x2 Cross Validation. They are working as expected. Both of these algorithms use stratified folds (splits). K-fold algorithm takes too much time to apply on the training data, since it is very large. So, I preferred 5x2 method for testing my algorithms.

Apply Cross Validation on Training Data

After implementing the functions, I called 5x2 Cross Validation algorithm on the preprocessed training data. Then I printed the resulting AUC values and plotted the ROC curve (I used cvAUC library). The following figure shows my ROC plot:



Train & Predict

At the end, I trained my model using the preprocessed training data. Then I used this fit to predict probability scores for the given test data. Finally, as stated in the homework description, I wrote these predictions into the csv file.