Report for Assignment

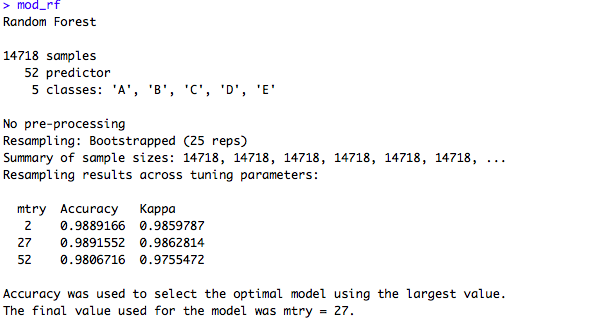
Liang Qinyu

The first step is data cleansing. After download the training and testing data, I replaced all the lacked data with NA and delete all the column which almost only contained NA. I assumed the column doesn’t contain any useful number if the mean of the column is 0. Besides, since I focused on analyzing the data from wearable accelerometers, the users name and time data are also removed for their irrelevance.

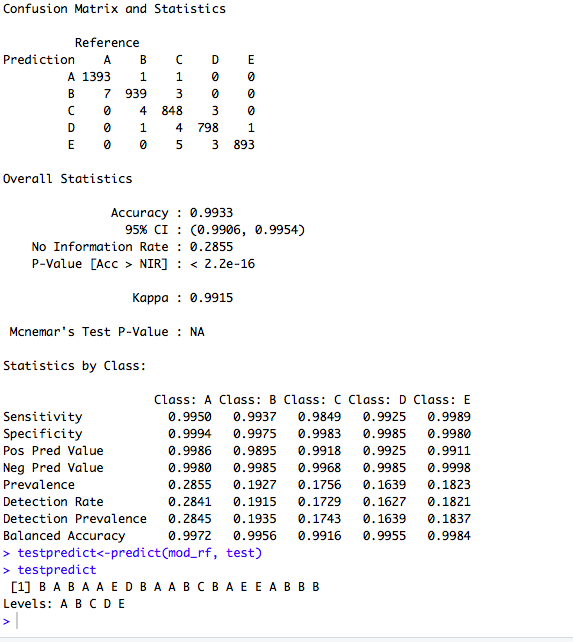
The second step is preprocessing. I divided training data into 2 parts, 0.75 of training data is for training and the rest is for cross-validation(in-sample).

The third step is modeling. Random forest algorithm is used for the model building, because random forest is a high-accuracy classifier on multi-variable data, and it could process abundant training data. What’s more, when decides the classification, random forest algorithm gives the importance of variable. When there’s data missing, it still could maintain the accuracy.

As the data given by Coursera needs is abundant and the question requires us recognizing a pattern from it and use the pattern to classify the testing data. I choose the RF as model.



Meanwhile, I used K-fold cross validation to run the cross validation so that I could repeat using the random sample to train and validation. This method avoids wasting data and overfit



The fourth step, prediction. From the confusion matrixes, the accuracy of in training prediction is 1 and the accuracy of out-of-sample prediction is 99.3%. Thus, there’s 0.7% out-of-sample error. The model works well.

The out-of-sample error shows how the model would perform in real practice. It happens because of overfitting, which means the algorithm matches the data too well. Overfitting happens if the model is too complicated and sometimes capture the noise instead of signal.

The fifth step predict in testing data. The answer is showed on the graph.