機器學習概論 Introduction to Machine Learning
Assignment 2 – Diabetes Prediction
Department: 理學院學士班
Student ID: 108020017 Name: 黃珮綺
-

1. Attributes setting of the random forest model

In the advanced part, I use 80% of input data as training data and the rest as validation data. There will be a total of 15 trees in the forest. Each tree is built based on 5 features, which is the square root of total number of features, and 80% of training data is taken as instances for training the tree. Also, I filtered some tree with low f1 score (0.65 as threshold) in order to ensure the accuracy rate of prediction.

2. Difficulty I encountered

During programming, I often encountered an index error in Dataframe. The most frequent error occurs when the index of the input data doesn’t start from 0. After splitting the data, instance of index 0 is separated to one side, and the index of the other side start from a random number other than 0. When process data of the side without index 0, the error then occur since the program could only process data with indexes starting from 0.

Also, I found it took a lot of time on training a forest. On average, it takes 5~10 minute to train a tree, without considering the accuracy rate. It is due to the large amount of data and the time latency on reading data of type Dataframe.

3. Solve the difficulty and reflection

For the index error, I re-index the data every time when I split the data or any possible move that will effect the indexing. As for the time latency, I adjust the size of training data and reduce the number of trees to 15. Although more training data and more trees may help improve the accuracy rate, the improvement will be slight after some threshold. So I just choose a reasonable number of data and trees. At last, the in this assignment I didn’t implement any methods on accelerating the speed of reading data of type Dataframe, but this could be done by changing the data type or implement the training process on C code or otherwise.