Reflection

This time we performed a multi-classification prediction on the given dataset. I chose the Naive Bayes classifier for research.

The reason for choosing is:

1. Bayesian classifier is a well-behaved and quite simple method for studying multi-classification problems

2. The Bayesian classifier uses a simple idea to perform a method that is theoretically superior to the baseline decision tree, but often cannot reach the accuracy of tree methods such as decision trees, so I tried the Naïve Bayes this time and wanted to find why the Bayes classifier cannot achieve high accuracy.

m

In the process of determining the baseline of our group, we were not sure whether we should choose decision tree or naive Bayes. In the end, we decided to use the decision tree as the baseline. The reason is

1. Naïve Bayes theoretically have a higher accuracy than decision trees. It will be strange if we choose a method that is better in theory as the baseline.

2. At the same time, although both are relatively simple and accurate methods, most of my teammates’ methods are modified methods based on decision trees, and choosing Naive Bayes as the baseline did not provide enough reference significance for their models.

In the process of coding, I still encountered some problems. The first problem I encountered was the difficulty of selecting variables. I tried to use the ‘nvz’ function to select variables like my teammates, but in this way the filtered variables did not met the requirement of the naïve Bayes model. In the end, I still filtered the variables based on the variance. This is quite a naïve way. In the last few days I read some reports of the deeper learning of data, I found some excellent methods for dealing with sparse matrices. Unfortunately, this time I do n not have time to applied these methods to my model and I will try them during the holiday.

The second problem is that the code takes too much time to run, sometimes it takes almost 10 minutes to run on a small data set like this. Although I know the extreme long running time is due to the higher computational complexity of Bayes algorithm, I still tried another package which claimed a much faster running speed, but the result was not good. The algorithm is not a well-performed classifier with high accuracy. At the same time, it can only do the prediction of 0-1 classification. The result of multi-classification calculated by this algorithm is very poor.

The last problem is that we can not decided a concise descriptive visualization of our model. We have to use the tasteless accuracy and confusion matrix as final result of our project. I once searched some interesting or useful way to visualize our model, but unfortunately the packages for visualization are all based on Python and we can not apply them to our results.

Anyway I think this work have shown a deeper level of knowledge in data processing to me. I have learned a lot from my thinking and practice.