Report AI Term project II

張書晧, 9917276

To be honest, I have no idea if the cut I made is correct or not. There are two methods that I use when I am making a cut for the intervals for the continuous data set. As for missing data (“?”), I simply ignore it and continue reading the next attribute or data.

The first method is to determine the intervals is to use Excel select and read through one particular attribute, which is sorted in descending order, and then create the intervals if I see a lot of “yes” output with respect to that particular attribute with the same input. For example, the interval 93.5<= x < 94.45 for the cons.price.idx attribute for the Bank dataset, I see a lot of “yes” output. Therefore, I will decide the interval to be around this range (93.5<= x < 94.45). The advantage of this method is to help me obtain more accurate prediction when my program is guessing the output. The disadvantage of this method is that I might have too many intervals and result of over-fitting.

The second method is not necessary for continuous data. I will make a huge cut if I see a lot of attribute possessing that input. For example, the native-country attribute in the adult dataset has a lot of input with United-States. As a result, I divide 41 different countries of the native-country attribute in to two main categories: United-States and non-United-States. This not only saves a lot of time, but also quickly separates the differences between people from different country.

I have recorded the intervals I that I use in the comment section right after I have read the Training data into the program. For that particular section, I have enhanced the description given and added the interval range for each attribute. For the Bank dataset, this comment is located from line 201 ~ 237. As for the adult dataset, I record the attribute intervals from line 1675 ~ 1698.

I don’t really like this project because building the database takes way too long where the actually algorithm of Naïve Bayes is not hard to implement. I spent over many hours copying and pasting the attributes and changing the index for the arrays which records different number of data input from the Bank’s and Adult’s training data and testing data.