Exercise 3 Report

Name: Wenchang Liu, ID: 10406141

Part 1: This part is to build a naive bayes classifier for the discrete spam email data, the whole process can be divided into the following steps:

1. Load the datasets according the input filename.

2. Loop the training samples, counting the appearances of each value in each feature. resulting in different label values, and put the numbers we counted into a 3 dimensional array. We also need to count the total appearances of each category.

3. Iterate the matrix we built in the previous step and calculate the probabilities for each counts, and we also need to calculate the probability of each category.

4. In the testing phase, for any given testing sample, we iterate each attribute of this sample, and look up in our calculated conditional probabilistic matrix to find the probability for this attribute value, and then we multiplicative those probabilities for each class. Then we compare the results among classes to find the largest one as our prediction.

5. Observe the classification results by calculating confusion matrix and accuracy.

Part 2: This part is to implement the classifier on continuous data, the process should contain the these steps:

(1) Classifier for avc_c2.mat

1. Check the dataset whether it holds continuous data or not.

2. Count the total appearances and calculate the probability of each category as we did in part 1.

3. Split the data set into n parts (n is the number of categories), and then we calculate each attribute's average value and standard deviation of each class respectively.

4. In the testing phase, we use normal distribution and the average value and the standard deviation we trained to get to calculate conditional probability instead of looking up the conditional probability table, and other steps are the same as part 1.

(2) 10 fold cross validation on "spambase.data"

1. Check the filename whether it is "spambase.data"

2. Randomize the spambase dataset samples and then extract last column(label data) from the spambase.

3. Use a for loop from 1 to 10, and for each loop, we select 460 samples for testing and the rest of the samples for training. We can still use the same classifier for continuous data in part 2 (1), and we store these 10 accuracies into an array.

4. Then we can calculate the average accuracy and standard deviation of the 10 fold cross validation, and we can also draw a plot with these results.

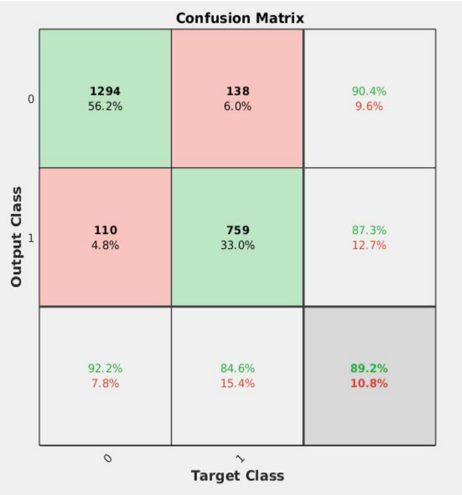
Q&A-1. In discrete naive Bayes, we need to learn conditional probabilities of each value of each attribute resulting in different classes; In continuous naive Bayes, we need to learn each attribute's average value and standard deviation of each class and then we could use these to calculate conditional probabilities according to normal distribution at testing phase.

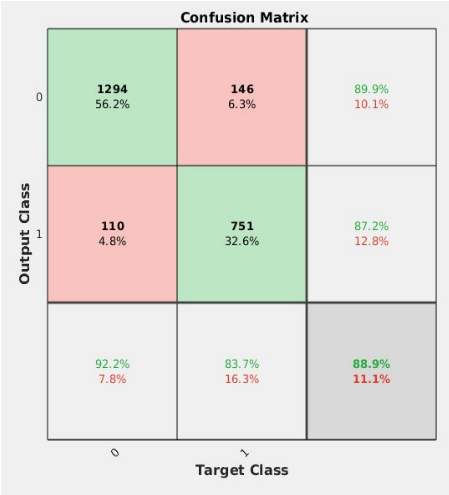
Q&A-2. For part 1, I use a 3 dimensional array, sizing of attributes value number X categories number X features number, and a array(size of categories number X 1) to store my "Parameter List", For part 2, I also use a 3 dimensional array, sizing of 2 X categories number X features number, and the same array to store my "Parameter List".

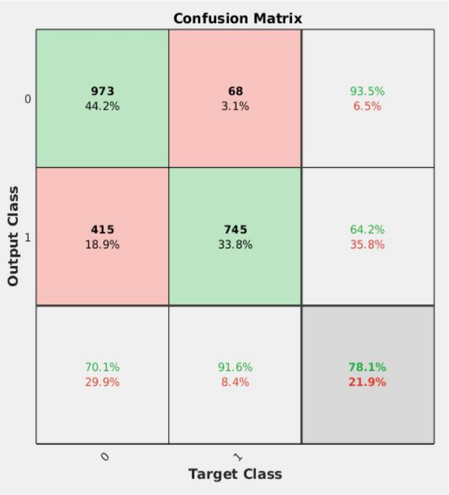
Q&A-3. Test results:

Part 1 Accuracies and confusion matrices:

Accuracy on av2_c2.mat: 0.888744
Accuracy on av3_c2.mat: 0.892221
Accuracy on av7_c3.mat: 0.866087





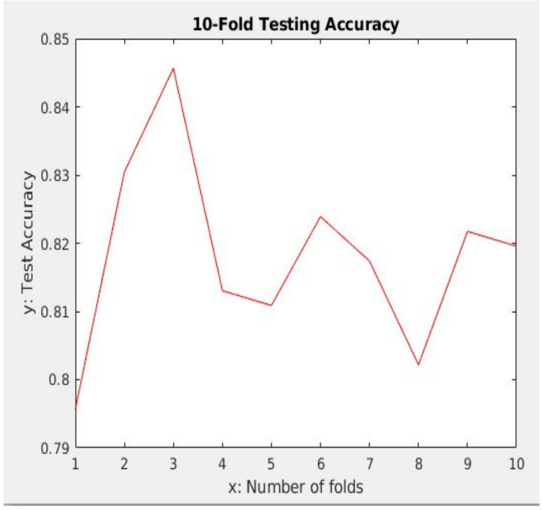


Confusion matrix for av7_c3.mat:

1201	0	53
0	635	37
84	134	156

Part 2 Accuracy and confusion matrix:
Overall Accuracy on Dataset avc_c2.mat: 0.780554
Q&A-4. Test results for spambase:
10-Fold standard deviation of accuracy = 0.014128
Overall Accuracy on Dataset spambase.data: 0.818043

Cross-validation motivation: It is a good way for us to evaluate our classifier if it is overfitting or underfitting, especially when we are not having a large amount of data to train and test. Using cross-validation, we can make full use of our data at hand, we not only care about the average accuracy of these 10 fold validation, but also pay attention to the standard deviation of it. If the results fluctuate a lot, then maybe our training process is overfitting.



Q&A-5.

(1) For "avc_c2.mat", the 41st attribute of class 1 has standard deviation 0, which will cause the result of normal distribution equals Nan, and all the classification will get the same result -- class 0. So what I have done here is to set a very small standard deviation instead of 0, and adjust this std to make the testing accuracy reach a sensible value.

(2) For part 2, when I was implementing 10 fold cross validation on "spambase.data", I found it important to randomize all the data, and then I can use a for loop to select training/testing sets. Because we want to minimize the influence caused by training/testing sets selection, so the most impartial way is to select them randomly.