CS498 Homework 1

Xinchen Pan

January 30, 2017

3.1

Firstly I loaded the data set and used `createDataPartition` to get 80% of the data set for training and 20% for testing. I built the naive bayes classifier following

$$\begin{aligned} p(y|x) &= \frac{p(x|y)p(y)}{p(x)} \\ &= \frac{\prod_i p(x_i|y)p(y)}{p(x)} \\ &\propto \prod_i p(x_i|y)p(y) \end{aligned}$$

We then choose the largest value of $p(y|x)$. We assume each predictor follows a Gassusian normal distribution and is independent.

In my code, I created the bayes classifier with my training data first. Then I used testing data to test.

(a)

The training accuracy I got from the bayes classifier I made is 0.7691057, the test accuracy I got is 0.6228758.

(b)

We have a lot of 0 in the data set, so I changed all 0 to `NA` as questions said. There is basically no change for the training accuracy which is 0.761626. The test accuracy increased to 0.6509804.

(c)

The training accuracy I got by use `caret` package using 10-fold cross validation is about 0.7563882 and the test accuracy is about 0.751634.

(d)

For this problem I used `svmlight` for doing support vector machine algorithm. I used 80% for training and 20% for testing. The training accuracy is about 0.7788618. I got an accuracy of 0.7058824 for testing.

3.3

(a)

For this problem, I firstly created a new variable based on the response variable. If response variabe is greater than 0, then I set it to 1, else set it to 0. Then I used 85% of the data for training and 15% for testing. I ran the model ten times and reported the mean and the standard deviation of the accuracy. The accuracy for the training data is about 0.8092308. The test accuracy is about 0.8555556. The standard deviation is about 0.0459618.

(b)

We do not do any change for the response variable. I ran the model ten times and got a training accuracy of 0.5645499. The testing accuracy is 0.5755556 . The standard deviation is 0.060587.