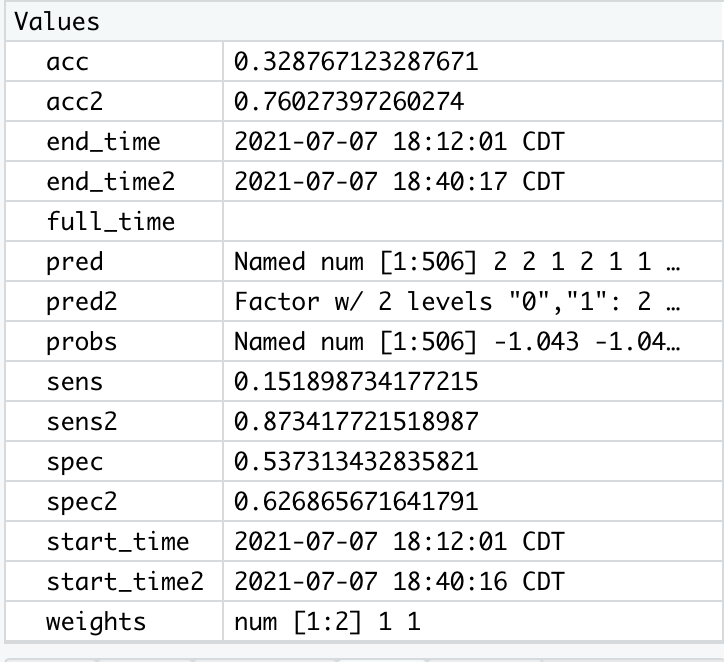
Mustafa Khan

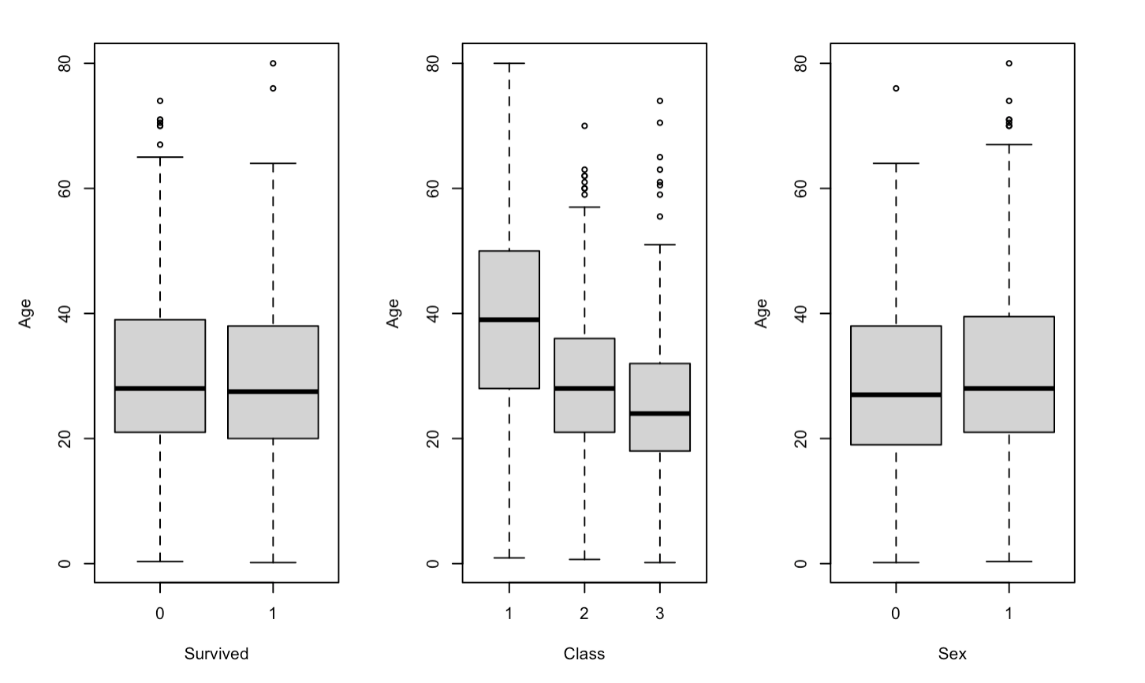
CS 4375

7/7/2021

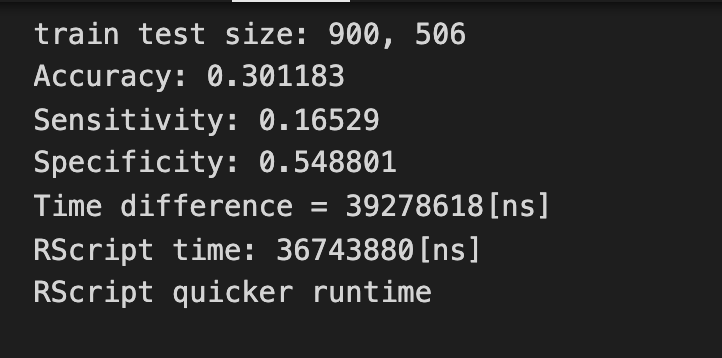
Report

For the first part we used logistical regression to make predictions on the data from the Titanic data frame. We firstly split the data to train and test data and from there I started to record the start time. Using pclass as a predictor for survived, I was able to form a model. Using the model, I was able to test it on the test data and get a table of predictions and survived as an integer. The table gave me the true positive, false positive, true negative, and false negative values correctly. Using these values, I calculated the accuracy, the sensitivity, and the specificity. 

The values with the “1” after them are the values I obtained using the logistical model. Note, the time difference was 0.02 seconds. The runtime for the naïve bayes was 0.0378 seconds. The logistical model was a tad bit faster. On the graphs I predicted Age from the train values of survived, class, and sex using box plots for each. The average age of females compared to males was relatively equal with a few outliers. The age distribution when it came to classes was different. Class 1 were the eldest, followed by class 2 and then class 3. The average age of the people who survived versus the ones that perished is relatively equal too, landing at around 30 years of age.



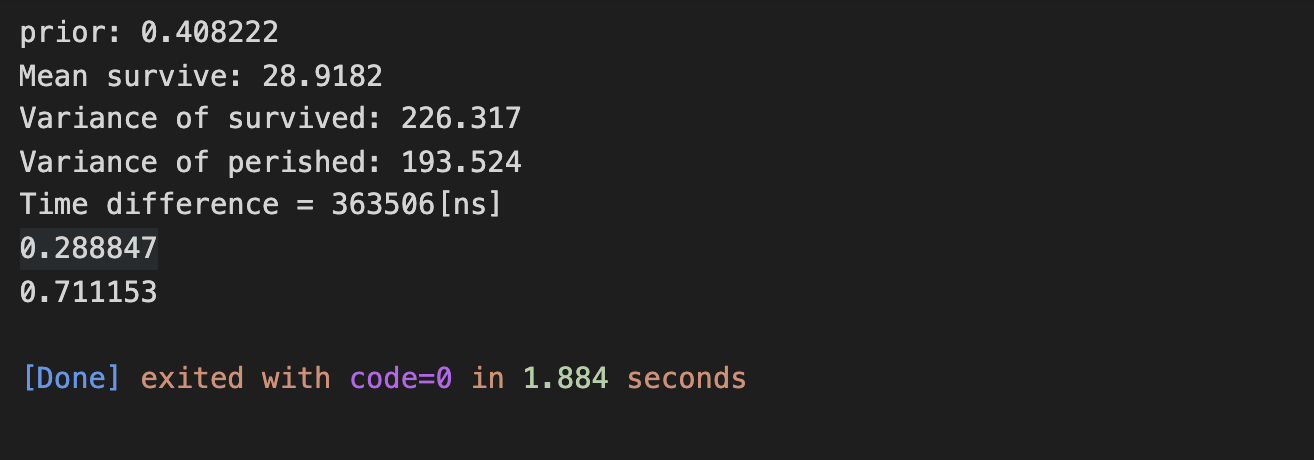
In the C code, I read all the survived and put it into a vector and with pclass as well. I made training and test data sets with these values. I made a matrix of pclassTrain with a vector of 1’s. After this I started the time off. I did matrix multiplication from scratch and stored individual values in a variable that I sent to sigmoid function. I calculated the error and updated the values of the weights. Using these values I was able to calculate accuracy, specificity, and sensitivity.



The run time was slower for the C++ program. R Script runtime:

**Time difference of 0.02917814 secs**

The time came in at about 0.0367. The accuracy, sensitivity, and specificity were all close to the R script values.

The naïve bayes model was made with training data that was split the same way it was in the logistics model. The prior survival probability was calculated along with the mean age and variance for both survived and perished. The likelihood of class1, 2, and 3 along with the likelihood of each sex surviving and perishing was also calculated. The age likelihood of survival and non survival given a certain parameter, which in this test case was age of 19, was calculated too. Because of that I was able to calulate the raw probability of survival which was 0.2888 in this case, and the raw possibility of not surviving, which in this case was 0.7111. Tested on R Studio, these test cases and values were correct. The accuracy was a little off comparing the C++ to R, accuracy was 0.32, sensitivity was 0.85, and specificity was 0.52, a few decimals off from the R script values.

In this particular case, the C++ run time was quicker, coming in at 0.003635 seconds, whereas R was:

**> print(paste("Run time was: ", end\_time2 - start\_time2))  
[1] "Run time was:  0.115835189819336**"

C++ was considerably faster.