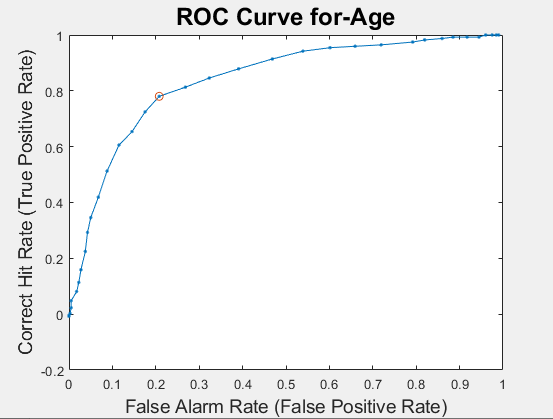
Here, in this experiment, I performed classification by implementing the concept of the confusion matrix as there are known groups under which the data should be separated and then eventually become a classifier for future verification data. This program is written in Matlab. It follows the same requirement for accepting input from the user, i.e via command line. There will be a prompt that will indicate for the csv data to enter and then the classifier and the trained program will perform its necessary functions.

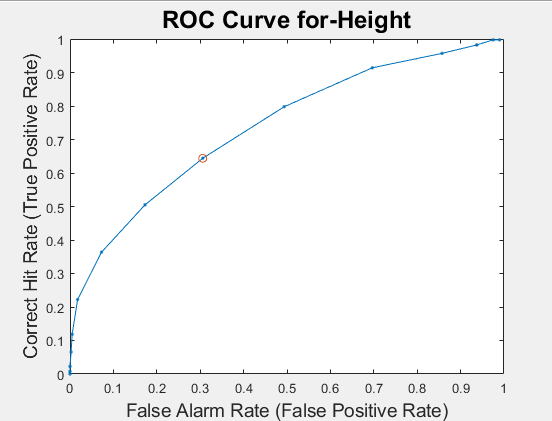
The data for Abominable\_Data\_FOR\_CLASSIFICATION\_HW\_v2201\_720 file for classified by isolating the age and height values from the csv file. Here, each file was quantized differently. The data for the height for quantized to the nearest 5 cm and the data for the age was quantized to the nearest 2 years using the floor function. The target class is Bhuttan and it will be represented as ‘+1’. Thus, ‘-1’ for Assam.

For the mentor program, after quantizing the data, I extrapolated the edges for each of the data lists. Using these respective edges, I was able to calculate their individual thresholds and thus find the best thresholds by comparing the cost functions.

This method was processed through the Age and Height data sets. This resulted in respective threshold indices, best cost functions (the ones with the least amount of errors) and the threshold values. To find the best classifier, I compared the cost functions of the Age and Height data sets and found that the Age data set gave the least error value of 170. Thus, the trained file is written to highlight the age differentiation.

Furthermore, ROCs were also created for each of the data sets as they are shown below:





In comparison with the Otsu’s method, the values to roughly give the same result in terms of age as being the threshold, however, Otsu’s method was pertaining to the clustering concept where there had been no knowledge about the data that was being passed down and how exactly it would be analyzed to get the threshold value. As mentioned above, for this experiment, the confusion matrix was used.

One difficulty I had here was trying to identify the mathematic formula for the False positive and the false positive rate. For false positive, I realized that my mistake was that I was over complicating the formula and not making the efficient. It took me time to realize how to efficiently extrapolate the data from the class data file that stores the names of Assam and Bhuttan. The rate was especially tricky as I wasn’t able to realize that I needed to incorporate the total number of Bhuttans and the total number of Assams within the formula.

Overall, I enjoyed this experiment due to the small complications I had. It made me think through the implementation of the concept.