Pratishta Prakash Rao

Big Data Analytics - Homework 3

1.

a) **Considering that we are trying to maximize public safety on the roads, how would you break a tie if two different speed thresholds have the same lowest misclassification rate? How would you set your threshold to the lower or the faster speed? Why?**

Solution:

In order to maximize public safety on the roads, it is important to have cars moving with the speeds within the speed limit in that area.Since the data set at hand doesn't have any other attributes except for aggressiveness, I'm assuming that the data was collected on one particular road, we would have access to the speed limit. Let's say that the speed limit in this particular road is 65:

Case 1: If the two thresholds with the same misclassification rates are 65 mph and 78 mph, the obvious answer would be to pick 65 mph in order to maximize safety which can happen with cars driving within the speed limit.

Case 2: If the two thresholds with same misclassification rates are below the speed limit say 48 mph and 55 mph, the answer here would be to choose 55 mph as slow drivers are also at the risk of causing accidents.

Case 3: If the two thresholds with same misclassification rates that are at 55 mph and 67 mph, the answer here would be to pick the 67 mph as this speed is closest to the speed limit.

b) **Imagine that you are trying to maximize how much trust the public has in the police officers, how would you break a tie if two different speed thresholds have the same lowest misclassification rate? Why?**

Solution:

One way of maximizing trust the public has in the police officers is when the police officers give out tickets to people who are over speeding. By giving out tickets to people who aren't over speeding and are following the rules can cause the people to lose trust in the officers. So I would pick a threshold that's closest to the speed limit.

c) **Define a cost function such that false alarms are just as bad as a miss.So, the cost function will be equal to (number false alarms + number of missed speeders).**
Using the techniques covered in class, write a program to find a threshold for a police officer to set their laser speed detector at so that it beeps in such a way that it minimizes this cost function.
In case of ties, maximize the public's trust that a police officer is not pulling people over for the fun of it. (Use the higher threshold.)
Here, I want you to round the speeds to the nearest 0.5 mph, sort them, and then try the speeds from slowest to the fastest.
What threshold value did you compute as the best threshold? (To the nearest 0.5 mph)
Solution:

The best threshold value is 62.5 mph.

d) **Suppose that we want a cost function such that false alarms are two times worse then a miss. So, the cost function will be equal to (2 x the number false alarms + 1 x number of missed speeders), or maybe it is (1 x the number false alarms + 2 x number of missed speeders). Guess how the threshold will move. What do you guess? Up or down? Faster or slower?**
Temporarily change your program to minimize the cost function, and check your guess. How did the threshold change? T e m p o r a r i l y modify the program to minimize this cost function.
What new threshold value did you compute for this new cost function? (To the nearest 0.5 mph) Does this change make sense? Why or why not?
Solution:

By multiplying the false alarms by 2, we are trying to minimize the false alarms, this can happen when the threshold is set higher, so my guess is that the threshold should be higher. By multiplying the misses we are trying to catch the speeders who are wrongly classified as non aggressive drivers. So reducing the threshold would mean higher chances of catching them, so my guess with this cost function would be that the threshold is lower.

The threshold obtained with cost function = 2 x the number false alarms + 1 x number of missed speeders is 63 mph

The threshold obtained with cost function = 1 x the number false alarms + 2 x number of missed speeders is 57.5 mph

Although the threshold with cost function that puts more weightage on reducing the false alarm is not very different from the original thresholds, I think my guesses are still pretty close.

e) **Decompose this temporary cost function in terms of an objective function and regularization. What is the regularization being used here? What does the regularization penalize?**
Are there any issues with the relative amount of regularization here?
Solution:

Cost function = 1 x number of missed speeders + 2 x the number false alarms
Where the regularization = 2 x the number of false alarms
The regularization is trying to penalize the number of false alarms
No, I don't think there are any issues with the relative amount of regularization.

f) **Change your program back to using the first cost function.**
For the given training data, how many aggressive drivers does this let through for the given data set?
And
g) **For the given training data, how many non-reckless drivers would be pulled over?**

Solution:
Cost function = # of false alarms + # of misses
number aggressive drivers let go(misses) : 104
number of non aggressive pulled over(false alarms): 11

Cost function = 2 times # of false alarms + # of misses
number aggressive drivers let go(misses) : 113
number of non aggressive pulled over(false alarms): 6

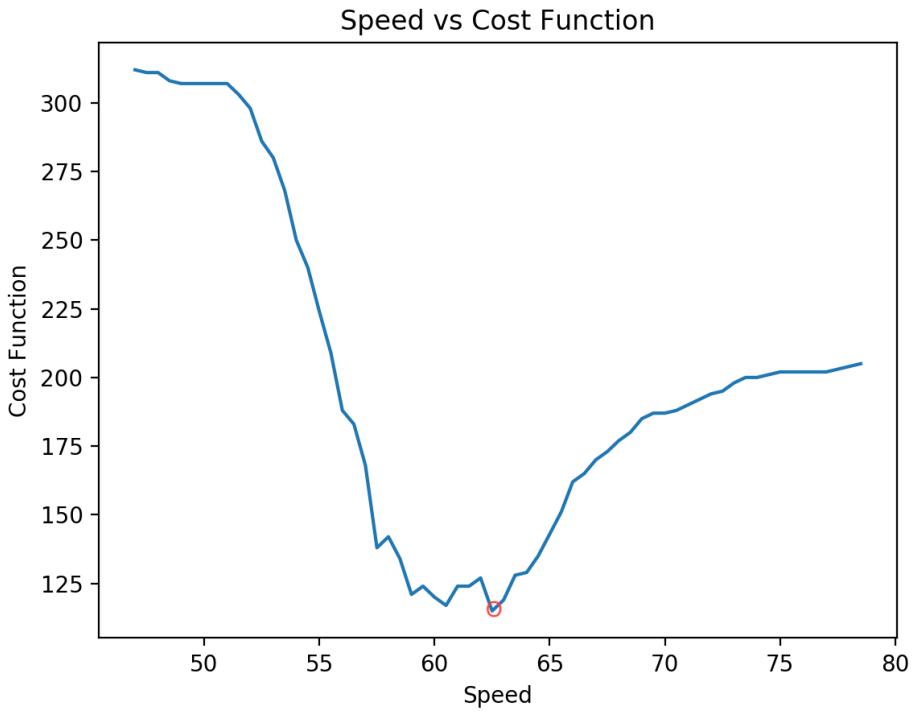
Cost function = # of false alarms + 2 times # of misses
number aggressive drivers let go(misses) : 22
number of non aggressive pulled over(false alarms): 116

h) **How does this value compare to the value you found using Otsu's method in the previous homework?**
Solution:
The threshold obtained from otsu's method is 60 mph which is lesser in comparison to the threshold obtained here.

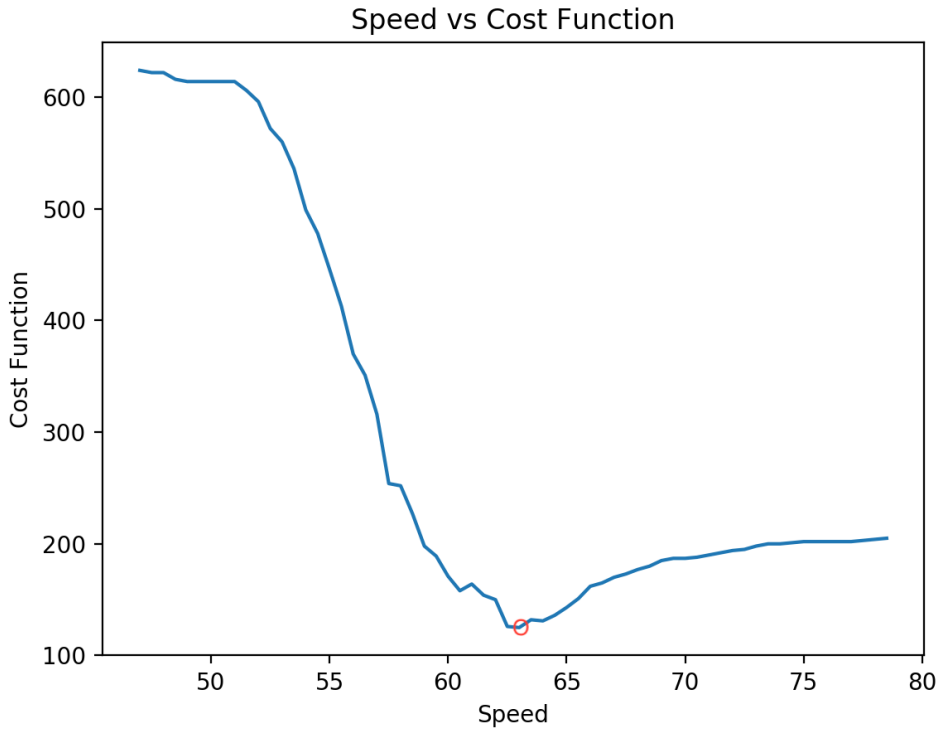
i) **Plot the cost function as a function of the threshold used. Label all axes.**

Solution: Following are the plots obtained:

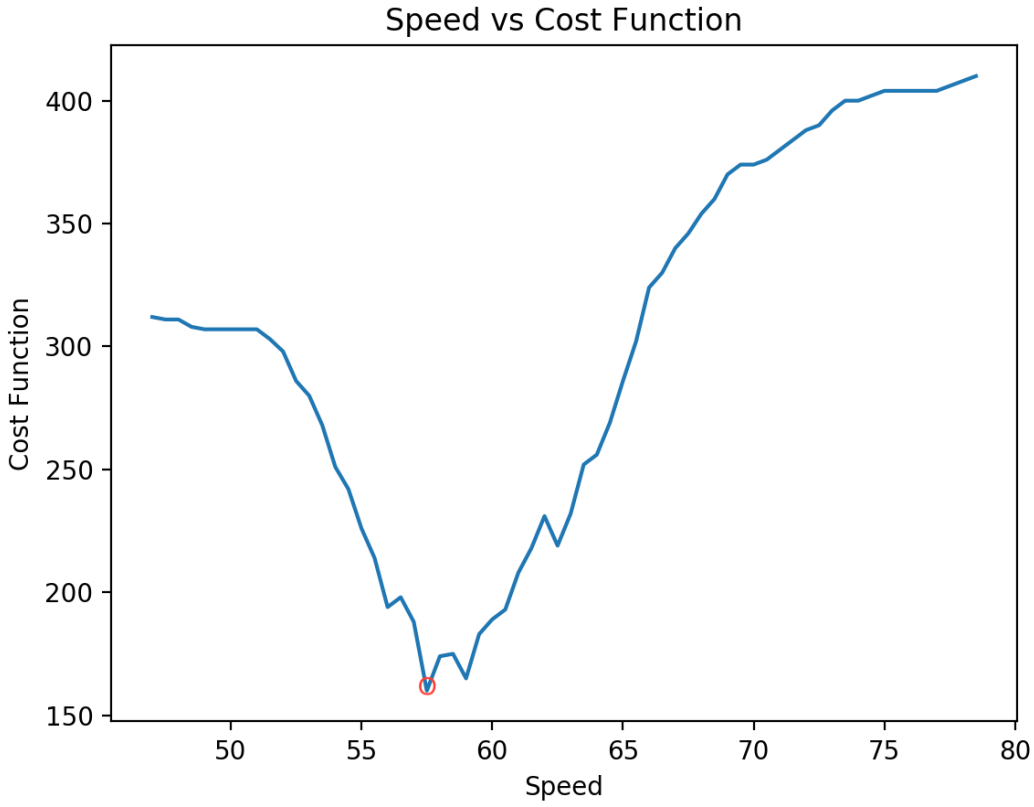
Cost Function = # of false alarms + # of misses



Cost Function = 2 times # of false alarms + # of misses



Cost Function = # of false alarms + 2 times # of misses

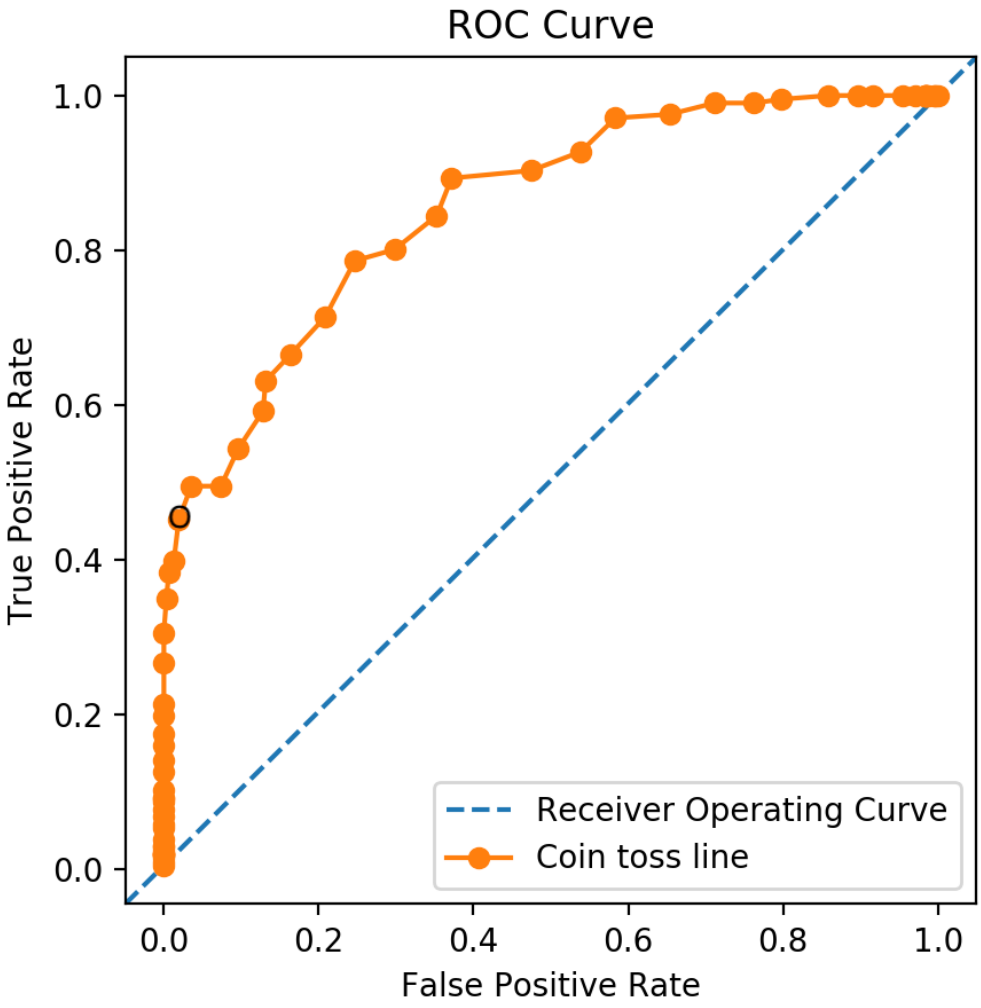


j) **Generate a receiver-operator (ROC) curve for this training data. Plot it, and put the location of the any thresholds on the ROC curve. Label the X and Y axes correctly. Try to make the axes square if possible, so that relative slopes can be compared. Circle any point (or points) on the ROC curve with the lowest cost function. Caution, there may be more then one of them.**

Solution:

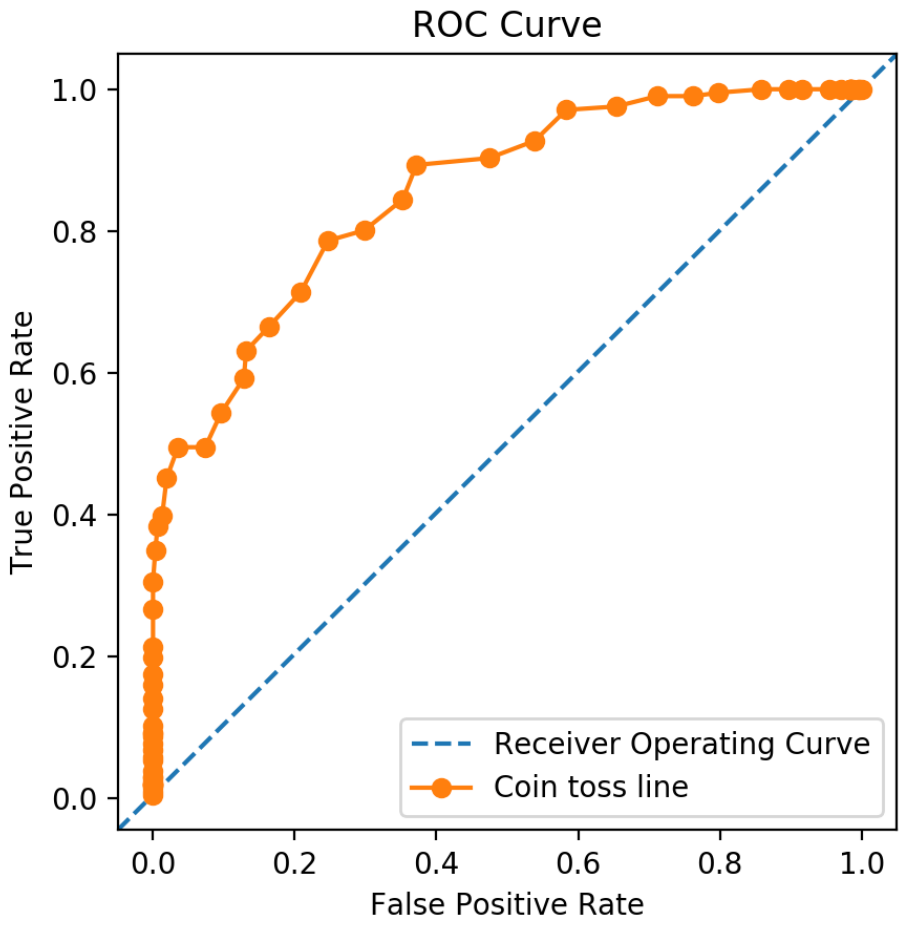
Pratishta Prakash Rao

ROC for Cost Function = # of false alarms + # of misses



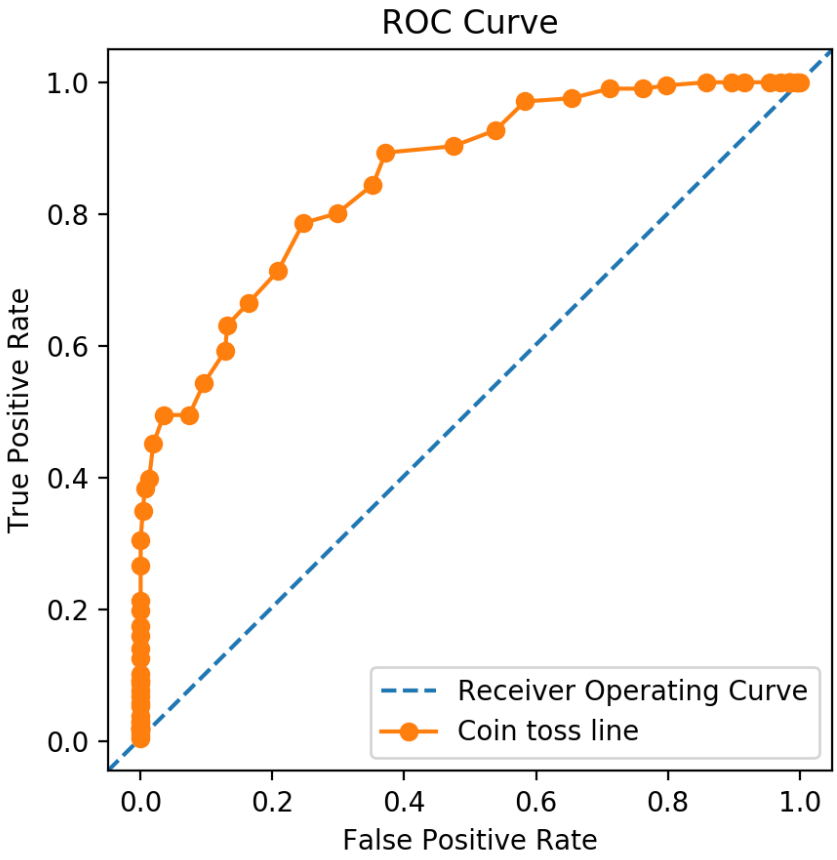
Pratishta Prakash Rao

ROC curve for Cost Function = 2 times # of false alarms + # of misses



Pratishta Prakash Rao

ROC curve for Cost Function = Cost Function = # of false alarms + 2 times # of misses



k)) **Conclusion:** Write up what you learned here using at least two paragraphs. How might you use a one-dimensional classifier with multi-dimensional data? Was there anything particularly challenging? Did anything go wrong? Provide evidence of learning.

Solution:

Things I learnt from this homework were significant. The first thing that I learnt was how n- dimensional data can be classified using one dimensional classification. With the given two - dimensional data, I calculated the false positive, true positive, false negative, true negative, found the lowest value of the cost function and obtained the threshold.

I also learnt about the importance of regularization and how it can play a role in calculating your cost function and how emphasis can be laid either on minimizing the mistakes or maximizing the accurate classification. I also realized the importance of ROC curves on how they can be used to determine the threshold for a classifier and how it can be used to choose between two or more classifiers.

For me the challenging part was understanding the role of regularization and how it can be used to manipulate the cost function. But after thinking about it for a while and practically doing it and looking at the result gave me a better understanding of the concept.

Initially, I didn't not find all possible thresholds and was just using the quantized and sorted thresholds. I did realize the mistake and corrected it. Also, while finding the elements of the confusion matrix, I was a little confused as a result of which my ROC curve was not right, but after I spent a little time together, I realized the error and corrected it.