Big Data Analytics

Nihal Surendra Parchand

np9603@rit.edu

Homework 3

1.

a. In the case of maximizing the public safety on roads, it is safer to go with the lower speed as people have more time to react and accidents can be avoided. So, for breaking ties between speeds with same lowest misclassification rate, we choose the slower speed as threshold.

b. In the case of maximizing the public trust on the police officers that they are not getting pulled off unfairly, we should break ties by choosing the higher speed as threshold if the speeds have the same misclassification rates.

c. The best threshold = 58.5 mph

```
--------- Cost function = number of false alarms + number of missed speeders ---------

Best number wrong/cost function  = 179
Best threshold  = 58.5 mph
```

For breaking ties, we are choosing the higher threshold as we want to maximize public trust on police officers. So, in order to choose the higher threshold for same cost function/misclassification rate we use <= for comparing it with current best cost function and then overwrite the threshold value with higher speed if two speeds have the same cost function.

```
if number_of_wrong <= best_number_wrong:
    best_number_wrong = number_of_wrong
    best_threshold = threshold
```

If safety is the priority then we should consider threshold < 58.5 mph as safe. For our case, I considered threshold <= 58.5 mph as safe because our goal is to maximize the public's trust in police officers. So, people driving at 58.5mph would not be caught as we are using the higher threshold as compared to threshold < 58.5 mph.

d. If the cost function is such that missed speeders are two times worse than the false alarm then we use the formula:
Cost function = number_of_missed_speeders + (2 * number_of_false_alarms)
The results are similar to what I predicted that the threshold value should increase and it increased to 60 mph. For threshold = 0 the number of false positives or the false positive rate is 1 and false negatives or false negative rate is 0. And as threshold increases, the FPR decreases and the FNR increases. This makes sense because if we increase the threshold speed, then more number of speeders will get away because the threshold set is too high. The screenshot shows that the number of missed speeders (123) is almost twice the number of false alarms (59) while minimizing the cost function. Also the

```
-------- Cost function = number_of_missed_speeders + (2 * number_of_false_alarms) --------

Best number wrong/cost function with double false alarms = 241
Best threshold with double false alarms = 60.0 mph
Best false alarm rate = 0.11591355599214145
Best true positive rate = 0.7583497053045186
Number of aggressive speeders who did not get caught = 123
Number of non-reckless drivers who got pulled over = 59
```

e. I think the cost function is the sum of missed speeders and false alarms and the regularization used here is the additional false alarms which we are adding to this temporary cost function. It is penalizing the number of missed speeders (false negatives) as the resulting cost function increases the threshold and allows more number of aggressive speeders to get away. Even if it is reducing the false alarm rate, we should not use this cost function because this reduces the safety on the roads as we are using the higher threshold.

f. For the given training data, the number of aggressive speedeer who did not get caught (false negatives) = 86

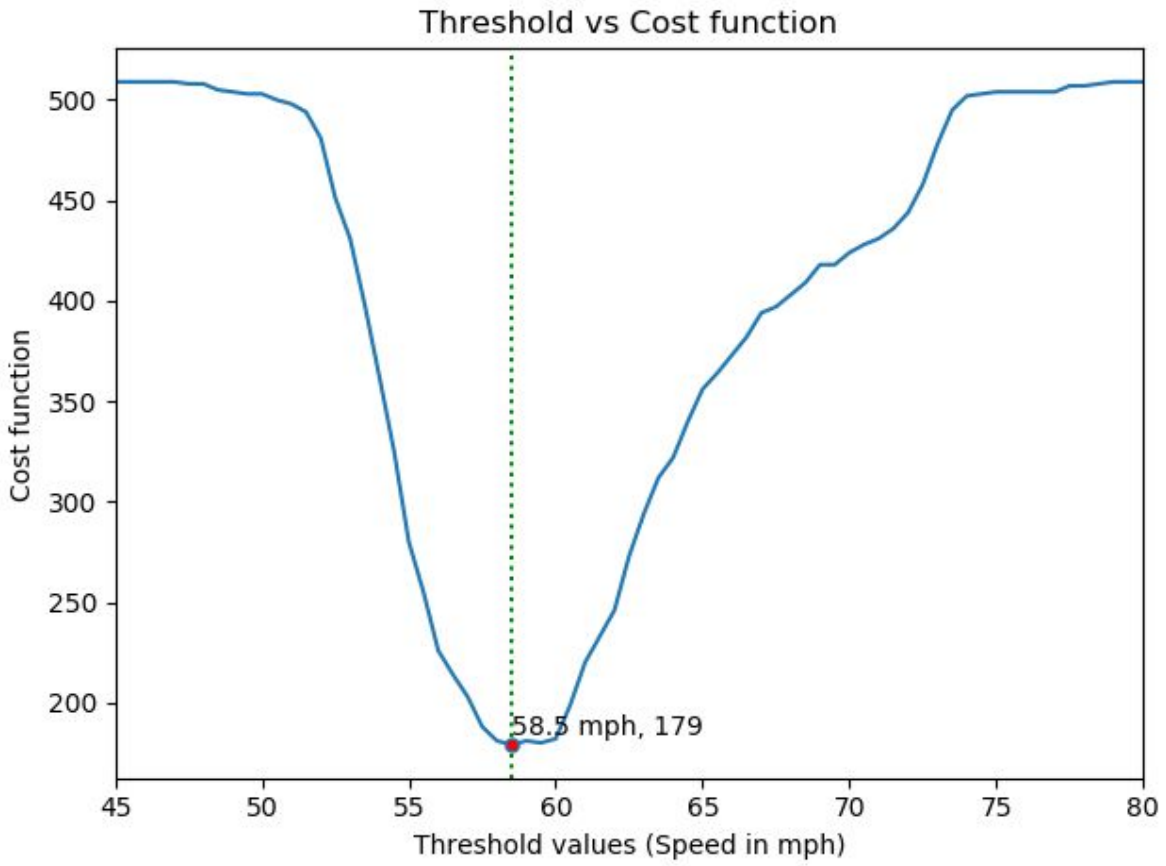
```
-------- Cost function = number_of_missed_speeders + number_of_false_alarms --------

Best number wrong/cost function  = 179
Best threshold  = 58.5 mph
Best false alarm rate = 0.18271119842829076
Best true positive rate = 0.831041257367387
Number of aggressive speeders who did not get caught = 86
Number of non-reckless drivers who got pulled over = 93
```

g. For the given training data, the number of non-reckless drivers who got pulled over (false positives/ false alarms) = 93

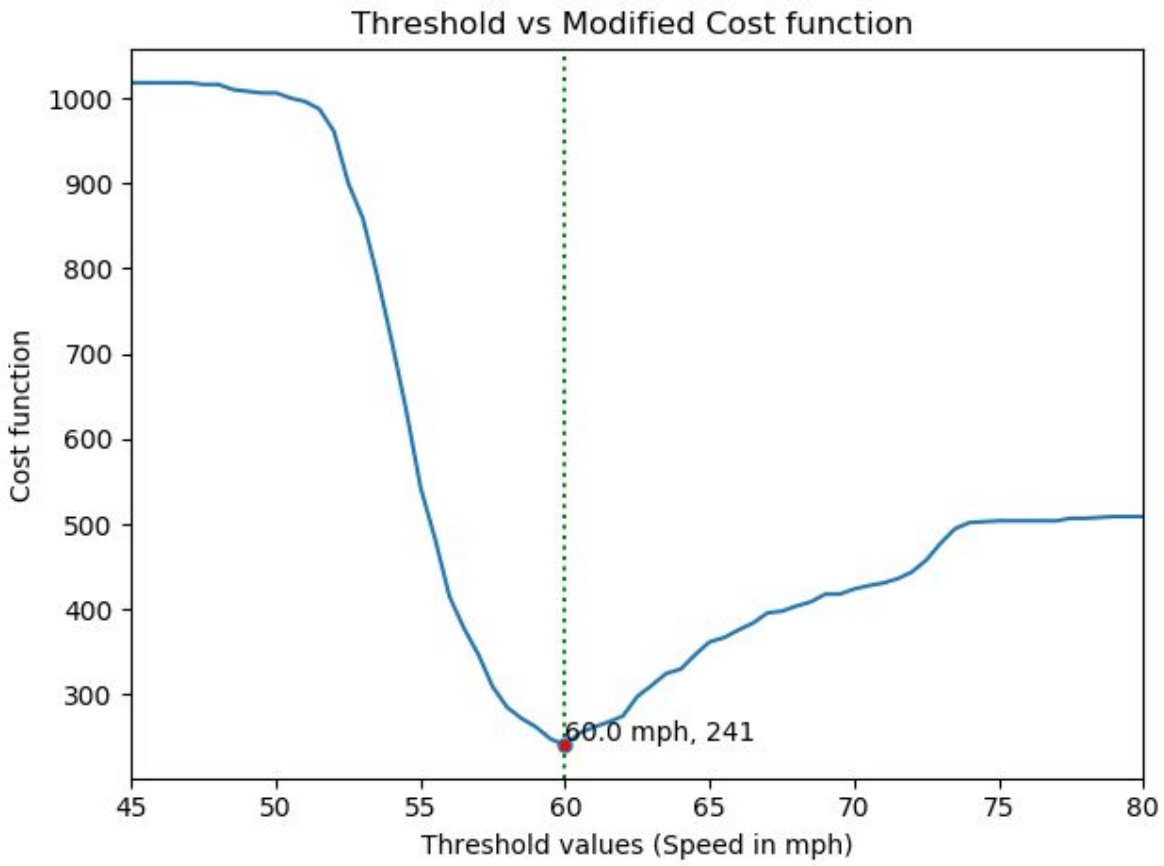
```
-------- Cost function = number_of_missed_speeders + number_of_false_alarms --------

Best number wrong/cost function  = 179
Best threshold  = 58.5 mph
Best false alarm rate = 0.18271119842829076
Best true positive rate = 0.831041257367387
Number of aggressive speeders who did not get caught = 86
Number of non-reckless drivers who got pulled over = 93
```

h. In Otsu's method, we got the threshold value as 60 mph with and without regularization for the given clustering data.
For classification, we got the threshold value as 58.5 mph without normalization and 60 mph when we change the cost function.

i. Plotting the cost function as a function of threshold



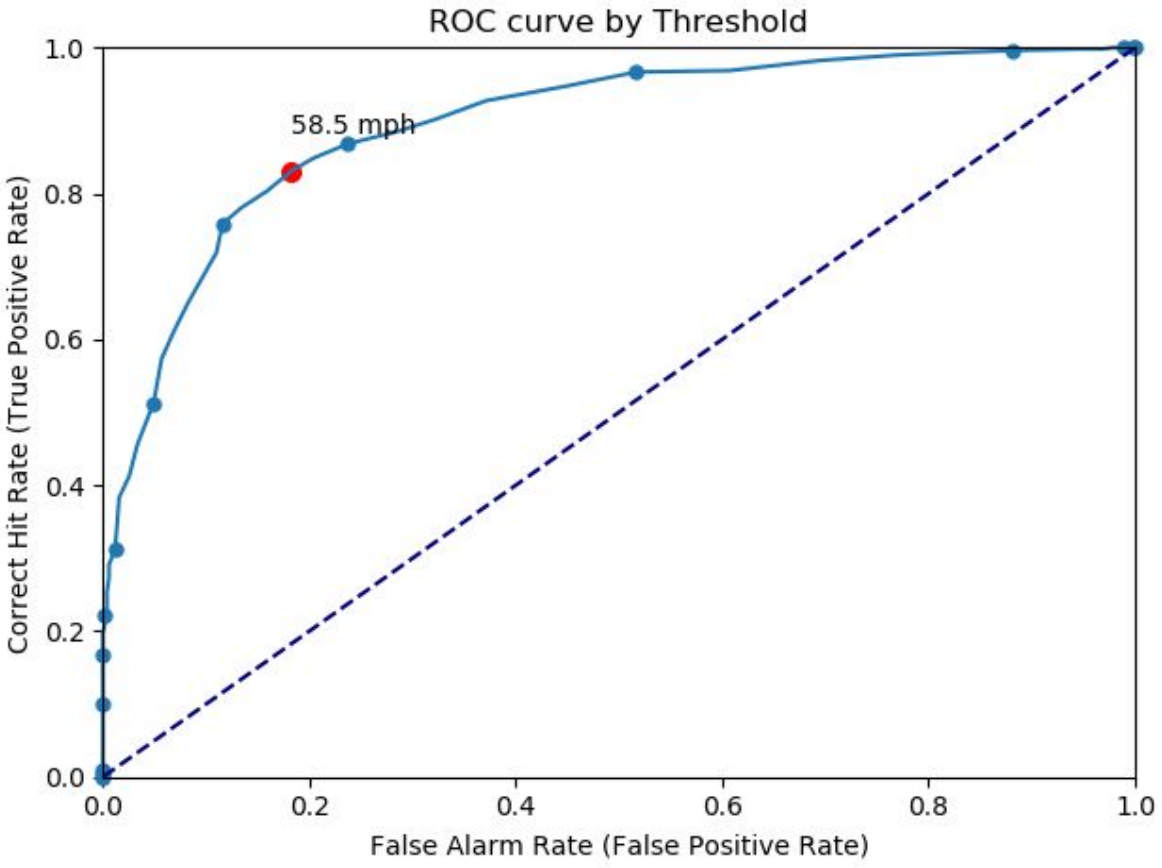
For the normal cost function, we got the best threshold as 58.5 mph and the best cost function is 179.



For the modified cost function, we got the best threshold as 60 mph and the best cost function is 241.

j. Receiver-operator curve (ROC curve)

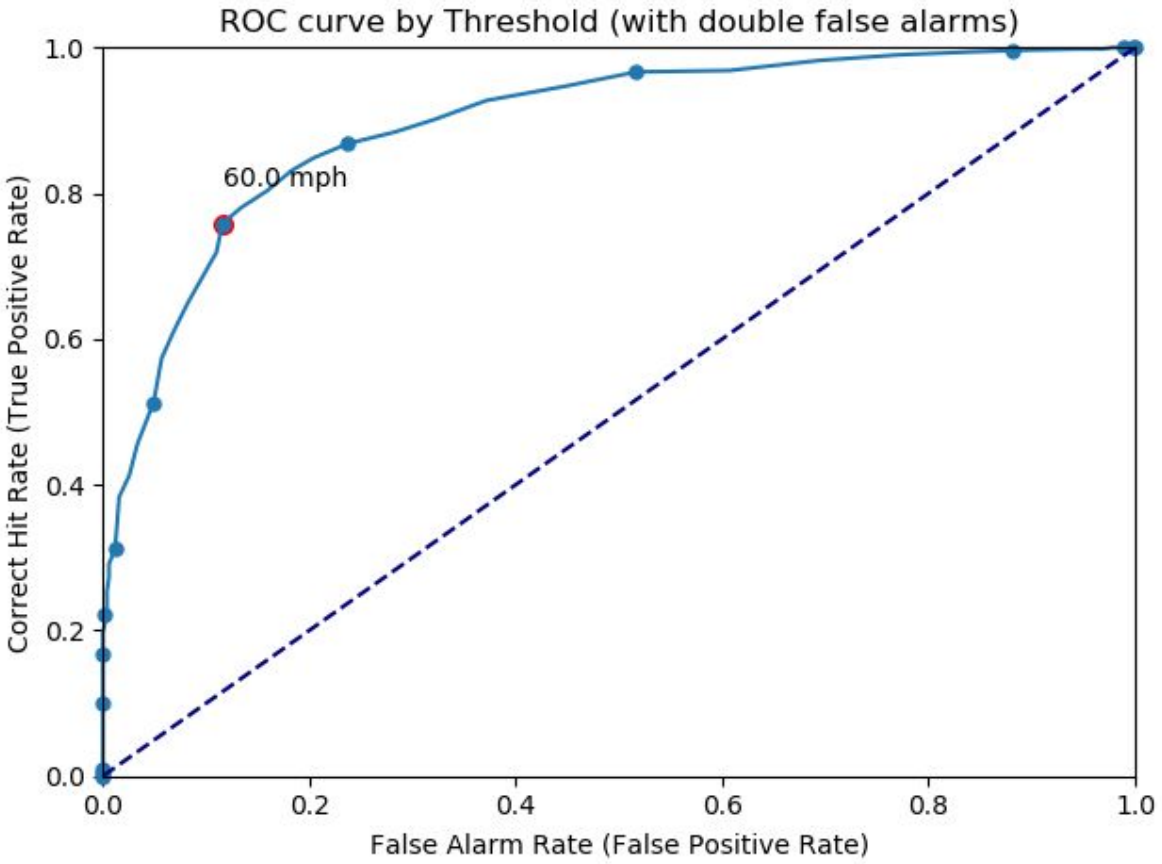
ROC curve for normal cost function



ROC curve for normal cost function

```
-------- Cost function = number_of_missed_speeders + number_of_false_alarms --------

Best number wrong/cost function  = 179
Best threshold  = 58.5 mph
Best false alarm rate = 0.18271119842829076
Best true positive rate = 0.831041257367387
Number of aggressive speeders who did not get caught = 86
Number of non-reckless drivers who got pulled over = 93
```

ROC curve for modified cost function




```
-------- Cost function = number_of_missed_speeders + (2 * number_of_false_alarms) --------

Best number wrong/cost function with double false alarms = 241
Best threshold with double false alarms = 60.0 mph
Best false alarm rate = 0.11591355599214145
Best true positive rate = 0.7583497053045186
Number of aggressive speeders who did not get caught = 123
Number of non-reckless drivers who got pulled over = 59
```

k. Conclusion: I learned about the 1-Dimensional classification algorithm. Firstly, the classification is different from clustering as we have a target variable (aggressive/ speeding or not) and we have to predict if the value is true for the target variable or not. I also learned about different terminologies like False positive, False negative, True positive, True negative. All these form a confusion matrix which helps in identifying which attribute is best for classification.

Four Possible Situations

		Actual Situation	
		TRUE	FALSE
Suspected Situation	TRUE	TP True Positive **HIT**	FP False Positive **FALSE ALARM**
	FALSE	FN False Negative **MISS**	TN True Negative **CORRECT REJECTION**

We can use the One-Rule to identify which attribute will provide the best classification results. As discussed in class, we build the frequency tables/confusion matrix for each attribute and choose the attribute which has the least misclassification rate. The goal is to minimize the incorrect classification/prediction of values. Hence, for a multi-dimensional data, we can use the 1-Dimensional classifier/ One-Rule to find out the best attribute for classification.

One challenging part where I got stuck for a while was to correctly interpret the data as in whether the false negatives should be incremented if speed <=threshold and aggressive == 1 or false positives. False negatives are the cases that we wanted to find but we missed, but if speed was already less than the threshold then shouldn't the driver be considered as false positive. But I figured out the actual interpretation later and I understood that the aggressive part was making the case as false negative. One last point that I wanted to add was that for this case we can allow false negatives to exist for the best threshold and cost function but if it's a medical or something serious issue then the cost function should be such that it minimizes the false negatives first. In short, the miss rate/false negative rate should be minimized to build a good classifier.