

Home Work 2

CSCI - 720 Big Data Analytics

Collaborators: None

1.

You will be provided with a file of driver speeds, several other attributes, and if they are trying to drive aggressively. (See the column headers.) The aggressiveness was based on an officer painstakingly interviewing the drivers.

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?

Ans. To maximize public safety we must attempt to minimize the number of false negatives(missed speeders). The cost function should be in such a way that we move the threshold point towards the **lower end**. The intention of the driver has no bearing because they have the same lowest misclassification rate.

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?

Ans. To maximize public safety we must attempt to minimize the number of false positive(false alarms). The cost function should be in such a way that we move the threshold point towards the **lower end**. The intention of the driver has no bearing because they have the same lowest misclassification rate.

c. Define a cost function such that false alarms are three times worse than a miss. So, the cost function will be equal to $3 \times \text{the number false alarms} + \text{number of missed speeders}$,

or maybe it is $3 \times$ the number false alarms + number of missed speeders. Guess and check your results.

Ans. Submitted as a part of the code

d. What threshold value did you compute?

Ans. 65 mph

e. For the given training data, how many aggressive drivers does this let through for the given data set?

Ans. 129 drivers are let through

f. For the given training data, how many non-reckless drivers would be pulled over?

Ans. 15 drivers are pulled over

g. How does this value compare to the value you found by Otsu's method?

Ans. The threshold for Otsu's method is 61 mph

h. Plot the cost function as a function of the threshold used.

Ans. Submitted as a part of the code

i. Conclusion

Write up what you learned here.

Was this a complete waste of time?

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.

Ans.

We learned how to implement one dimensional data classification. I

learnt how 1-D clustering compares with 1-D classification. No, it wasn't a waste of time. You can use a one-dimensional classifier with multidimensional data by choosing a single target variable. I did not find anything which was particularly challenging. There was nothing specifically that went wrong.