BIG DATA ANALYTICS (CSCI -720)

Homework-02 **Name: Rajkumar Lenin Pillai**

Question-1

Solution:

The implementation of otsu's method along with the plots is present in file q1.py

a.) There are no ethical issues in developing a machine which studies traffic volume for road planning in order to maximize traffic flow. This will actually help increasing the traffic flow and also reduce any possibility of an accident.

b.) Yes the ethical considerations change in this case, we can send speeding ticket to the reckless drivers but sending a paint ball that will stain the car permanently is wrong. Any damage to the car is not the proper way of handling this problem. The intentions of driver are responsible for reckless driving and not the car so sending a speeding ticket to the reckless drivers is enough as a punishment.

c.) The speed obtained from otsu's method to binarize the data is 61mph. The speed to best separate the two clusters 61 mph.

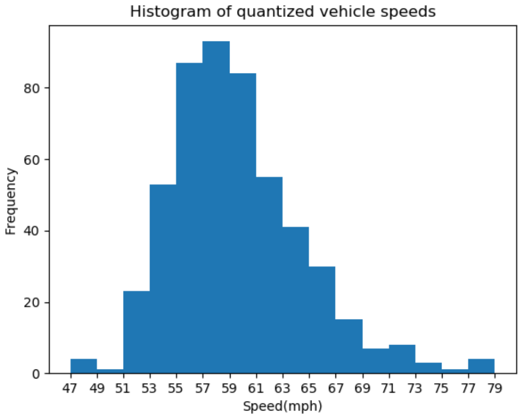
d.) The minimum mixed variance that resulted is 9.123116389957785

e.) If minimum mixed variance occurred twice ,the program chooses the minimum mixed variance which was computed first and the corresponding speed to sperate the two clusters because the python code to consider the minimum among computed mixed variance is as follows:

```
min_class_variance=min(class_variance_list)
```

For the dataset "DATA_v2185f_FOR_CLUSTERING_using_Otsu.csv "which is provided such situation didn't happen. But , the situation where minimum mixed variance occurred twice can happen in some other dataset.

f.) Plot of histogram of the quantized vehicle speeds.



Question-2

Solution:

a.) With norm factor = 50 and
Alpha = 1. The cost function yields the threshold as 64.72 (mph)
Alpha = 1/5. The cost function yields the threshold as 61.744 (mph)
Alpha = 1/10. The cost function yields the threshold as 61.372 (mph)
Alpha = 1/20. The cost function yields the threshold as 61.186 (mph)
Alpha = 1/25. The cost function yields the threshold as 61.1488 (mph)
Alpha = 1/50. The cost function yields the threshold as 61.0744 (mph)
Alpha = 1/100. The cost function yields the threshold as 61.0372 (mph)
Alpha = 1/1000. The cost function yields the threshold as 61.00372 (mph)

With norm factor = 20 and
Alpha = 1. The cost function yields the threshold as 70.3 (mph)
Alpha = 1/5. The cost function yields the threshold as 62.86 (mph)
Alpha = 1/10. The cost function yields the threshold as 61.93 (mph)
Alpha = 1/20. The cost function yields the threshold as 61.465 (mph)

With norm factor = 35 and
Alpha = 1. The cost function yields the threshold as 66.314 (mph)
Alpha = 1/5. The cost function yields the threshold as 62.06 (mph)
Alpha = 1/10. The cost function yields the threshold as 61.53 (mph)
Alpha = 1/20. The cost function yields the threshold as 61.265 (mph)

With norm factor = 10 and
Alpha = 1. The cost function yields the threshold as 79.6 (mph)
Alpha = 1/5. The cost function yields the threshold as 64.72 (mph)
Alpha = 1/10. The cost function yields the threshold as 62.86 (mph)
Alpha = 1/20. The cost function yields the threshold as 61.93 (mph)
Alpha = 1/25. The cost function yields the threshold as 61.186 (mph)

With norm factor = 5 and
Alpha = 1. The cost function yields the threshold as 98.2 (mph)
Alpha = 1/5. The cost function yields the threshold as 68.44 (mph)
Alpha = 1/10. The cost function yields the threshold as 64.72 (mph)
Alpha = 1/20. The cost function yields the threshold as 62.86 (mph)
Alpha = 1/25. The cost function yields the threshold as 62.488 (mph)
Alpha = 1/50. The cost function yields the threshold as 61.744 (mph)
Alpha = 1/100. The cost function yields the threshold as 61.372 (mph)
Alpha = 1/1000. The cost function yields the threshold as 61.0372 (mph)

From the above different values of norm factor and alpha which was tested we can observe that at alpha=1 with different norm factors above the cost function yields a different result in each case. In all cases decreasing the alpha below 1 /25 doesn't make much of a difference. The norm factor = 5 causes the "best" point to change and yields a different result for alpha = 1, 1/5, 1/10 ,1/20 , 1/25. Also we can observe that as the norm factor decreases the cost function yields a different result for different values of alpha.

Question-3

Solution:

a.) Mean: 15.775

Median: 14.5

Mode: 7 ,8 ,16

b.) After removing the last value (16) from dataset :

Mean: 15.75

Median: 12.0

Mode: 7 , 8

The mode changed which is 7 and 8 as 16 is now deleted, the mean value of data changed by a small fraction where initial mean was 15.775 and after removing 16 it becomes 15.75 but the median value changed from 14.75 to 12 . The value 16 was present in the data 4 times. The value 16 when removed midpoint of the frequency distribution of values changes to 12 as a median does not overly estimate the central value because of the outliers.

c.) With the original dataset "Mystery_Data".csv the otsu's routine computes the following values:

The threshold that splits data into two groups = 22

The minimum mixed variance that resulted = 22.084375

After removing the last value (16) from dataset :

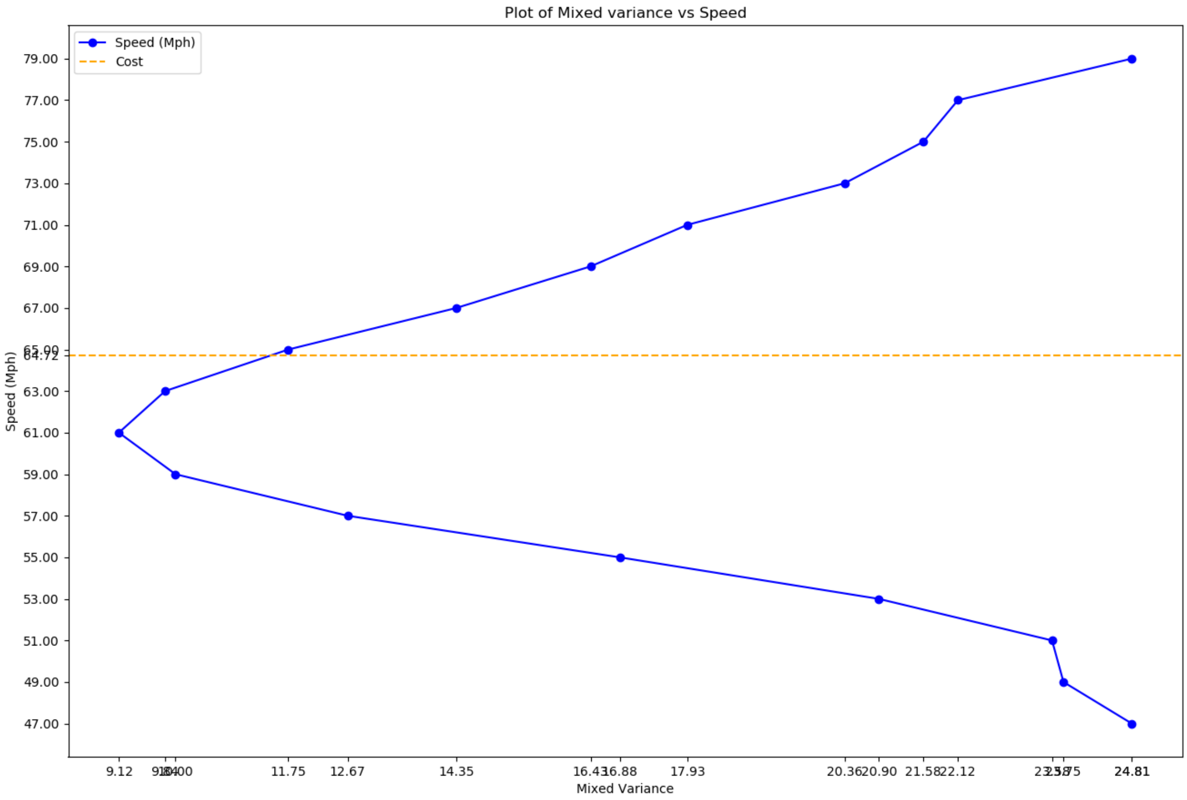
The threshold that splits data into two groups = 20

The minimum mixed variance that resulted = 22.40329218106996

Question-4

Solution:

Plot of the mixed variance for the car data in question 2, versus the speed.



The code to generate the above plot is present in file q3.py. The cost function yields a threshold 64.72 which is represented by orange dotted line since it contains the objective and also the regularization term.

Question-5

Solution:

a.) It took 4 hours to do this homework.

Initial guess to do the homework was 3 hours.

So , 3 /4 = 0.75 hours

b.) Factors that make it difficult to predict the time it takes to write software are lack of understanding of the complex problem , lack of knowledge to the write code to solve the problem. Also deciding which approach to use to solve the problem because some approaches may solve a part of the problem but not the actual problem and which approach will be correct can be decided most often by implementing the approach. Also testing the software for different test cases so tha if the testing fails the software needs to be fixed which takes more time.