HW03: One Dimensional Classification
See DropBox for due date.
Thomas Kinsman

Homework is to be programmed only in one of the following languages. No other languages will be accepted. Please limit yourself exclusively to: Java, Python, Matlab, or R. The last three had good native graphics and plotting support.

Assume that the grader has no knowledge of the language or API calls, but can read comments.
Use prolific block comments before each section of code, or complicated function call to explain what the code does, and why you are using it. Put your name and date in the comments at the heading of the program.

Hand in your results, and the commented code, in the associated dropbox.

Feel free to look over each other's shoulders, and at each other's work, but do you own work.
Let me know whom you worked with. Do not hand in copies of each other's code.

CAUTION: The bin size here is to the nearest half of a mile per hour.
This is different from the 2 mph binning that was used in the last homework.

Set your test thresholds to 30.5 mph, 31.0 mph, 31.5 mph, etc…

1. (½ pts) Read through the entire homework, and estimate how long it will take to do this homework before you start the homework. Again, this is for your education. Don't cheat. Write it down before you start coding.

2. You are provided with a file CLASSIFIED_TRAINING_SET_FOR_RECKLESS_DRIVERS.csv.

   This is a text file you can open and read with any text editor. CSV stands for "comma separated values." It contains two columns: the speed vehicles were observed travelling at, and if the driver was trying to be reckless. The recklessness was based on an officer painstakingly interviewing the drivers. Some of these were pulled over for aggressive driving, such as following too closely or neglecting to signal lane changes. Some were pulled over because their inspection certificate had expired.

   a. (½) Imagine that we are trying to maximize public safety, how would you break a tie if two different speed thresholds had the same lowest misclassification rate?

   b. (½) Imagine that you are trying to maximize traffic flow, how would you break a tie if two different speed thresholds had the same lowest misclassification rate?

   c. (½) What design decision will you use for your threshold? Will the value below the threshold be for speeds <, or <= the threshold?

   (Continued next page.)

d. (3½) 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 the total of (false alarms and false accusations).

In case of a tie, maximize the public's trust that a police officer is not pulling people over for the fun of it. (Minimize false alarms instead of maximizing true positives.)

Here, I want you to round the speeds to the nearest ½ of a mph. (round(speed)).

Sort them, and then try ½ mph increments from the slowest to the fastest.
Compute the misclassification rate for each threshold.

You may lose points for poor comments.

e. (1) What threshold value did you compute? (To the nearest 1/2 of a mph)
State this clearly in terms of the relationship. Do you pull over cars going >= 55.5 mph, or cars going > 55.5 mph. Be very clear so that you can be graded easily and correctly.

f. (½) For the given training data, how many reckless drivers does your decision let through?

g. (½) For the given training data, how many non-reckless drivers would be pulled over?

h. (½) How does this value compare to the value you found using Otsu's method?

3. (1) Plot the misclassification rate as a function of the threshold used.
Do this using a program API. Do not use Excel.

Circle the points with the lowest misclassification rate.
There may be more than one of them.

4. (½) Report how long it really took to do the homework divided by how long you estimated it would take.

5. (½) Who else did you talk with about your homework?
Again, be sure to hand in your own work, with your own comments in the code to be sure you understand it.

Hint: Always attempt the bonuses questions. Homework assignments get harder later. Get the points now.

6. (1 pts) BONUS:

Generate a receiver-operator (ROC) curve for this training data.
Do this using a program API. Do not use Excel.

Plot it, and put the location of the any tie-thresholds on the ROC curve.
Label the axes correctly. Circle the points with the lowest misclassification rate.