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 have 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 everything in one ZIP file in the associated dropbox. This should include: a) your write-up, b) your code, and c) your classifications. Put them all into one ZIP file. Assure that it can be unzipped correctly without outside libraries.

I encourage you to look over each other's shoulders, and to cross-check 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… and explicitly state whether you split at <= the threshold or < the threshold.

1. (¼ 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. **8 hrs.**

2. You are provided with a file CLASSIFIED_TRAINING_SET_FOR_RECKLESS_DRIVERS_2016.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?
   In case of a tie, to maximize public safety I will select the least value for threshold. This means that the threshold would be reduced and number of people getting caught (speeding by very little amount) might increase.

   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?
   In case of a tie, to maximize public safety I will select the maximum value for threshold.

This means that the threshold would be increased and number of people getting caught might decrease. This type of threshold should be used on highways.

c. (½) What design decision will you use for your threshold? Will the value below the threshold be for speeds <, <=, >, or >= the threshold?
Value below the threshold (value at which the cop will not be ticketing you) should be below the threshold. i.e. <u>Value below threshold< Threshold.</u>

d. (3) 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). (Implement it yourself in code. Do not use an R function).

(Continued next page.) 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 ½ speeds to the nearest of a mph.(in other words `round( speed )`).

Sort them, and then try ½mph increments from the slowest to the fastest.
Compute the misclassification rate for each threshold.
You will lose points for poor comments.

e. (½) 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.
Using this algorithm <u>cars going at speed >58.5 should be caught</u>.

f. (¼) For the given training data, how many reckless drivers does your decision miss? **4 misses**

g. (¼) For the given training data, how many non-reckless drivers would be pulled over? **4**

3. (1) Plot the fraction of cars misclassified as a function of the threshold used.
<u>Do this using a program API. Do not use Excel</u>.

Put a circle around the points with the lowest misclassification rate.
There may be more than one of them.

4. (¼) Report how long it really took to do the homework divided by how long you estimated it would take.
   Time actually required: 12 hrs. Ratio: 8/12=.67

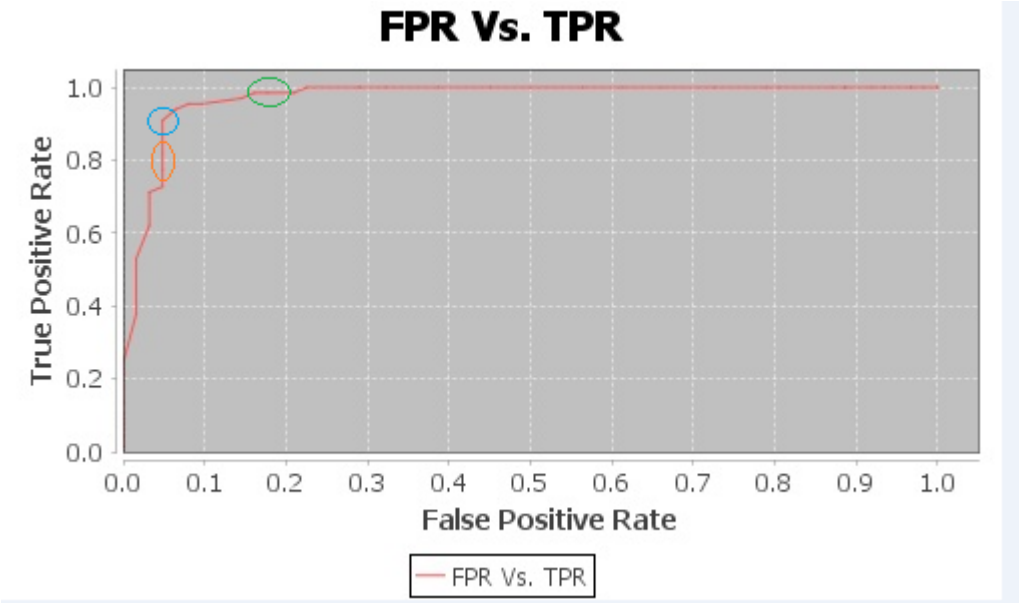
5. (½) Who else did you talk with about your homework? You *must* cross-check with at least one other person. Again, be sure to hand in your own work, with your own comments in the code to be sure you understand it. **Discussed the answers for threshold and misclassification rate with Manan Shah.**

6. (1) Write a program to classify the car speeds given in the file SPEEDS_TO_CLASSIFY_2016.csv. Indicate if they are speeding, as a 1 if they are speeding, and a 0 if they are not. What we are looking for here is a vertical list with a 1 or a 0 on each line. The first line is a 0 if the first speed is not speeding and a 1 if you think the first speed is speeding.

   The data to classify is in the file, SPEEDS_TO_CLASSIFY_2016.csv. Put your classification results in a file called "HW_03_<Lastname>_<FirstName>_CLASSIFICATIONS.csv".

7. (½) You just generated a classifier using some given data. What would you need to actually evaluate how good your classifier was?
   The accuracy of the classifier built by me can be checked by verifying the recklessness values determined by the algorithm implemented and the actual values given as training dataset (CLASSIFIED_TRAINING_SET_FOR_RECKLESS_DRIVERS_2016.csv).

8. (1 pts)
   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 point(s) with the lowest misclassification rate.



Blue circle: Lowest misclassification rate.
Orange circle: Points having ties.
Green circle: Points having ties.