HW03: One Dimensional Classification – with ROC Curve
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.

Put everything in one directory named HW_03_<Lastname>_<FirstName>. Then zip the entire thing up and put everything in that one ZIP file into the associated MyCourses 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 0.02, (i.e. 1/50). This might or might not be different from the binning that was used in the last homework.

Set your test thresholds to 0.02, 0.04 0.06, 0.08, etc… and explicitly state whether you split at <= the threshold or < the threshold, >= 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.

2. You are provided with a file of classified data for training.

   There are some essay questions here. Write at least a paragraph for each, explaining your mindset and justifying your answers. Answers that are too short will be rejected.

   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 same attributes as in HW02, *and the class of the cantaloupe*. This data is simulated, but in real life, this data would come from carefully cutting open several hundred melons.

   a. (½) Imagine that we are trying to maximize customer satisfaction of melons. We want to be sure that the melons that customers take home (thinking that they are ripe) are really ripe. How would you break a tie if two different "ripeness" thresholds had the same minimum misclassification rate?

   b. (½) Imagine that you are trying to maximize melon sales. How would this change your answer to the last question.

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

d. (3) Using the techniques covered in class write a program to find a threshold that indicates that a melon is ripe. Design this so that it minimizes the total of (false alarms + false accusations). (Implement it yourself in code. Do not use a function).

In case of a tie, maximize the public's trust that a cantaloupe is ready to eat

Be sure to comment your code so that the graders can understand what your thoughts were.

e. (½) What threshold value did you compute? (To the nearest 0.02.)
State this clearly in terms of the relationship. Are the melons ripe if they are over this threshold or under this threshold? Be very clear so that you can be graded easily and correctly.

f. (¼) For the given training data, how many ripe melons does your decision miss?

g. (¼) For the given training data, how many unripe melons does your method say are ripe?

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

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.

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.

6. (1) Write a program to classify the melons given in the file MELONS_TO_CLASSIFY_2016.csv. Indicate if they are ripe with a 1 (ready to eat), 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 melon is not ripe and a 1 if it is either ripe or rotting.

The data to classify is in the file, MELONS_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?

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 circle the location of best threshold on the ROC curve. (In other words, circle the point or points with the lowest misclassification rate.) Label the axes correctly.