HW_04 One Rule Threshold Classification for Multiple Attributes, Fall 2024
See Dropbox for Due Date
Thomas B. Kinsman

Homework is to be programmed in Python, R, Matlab, or Java.

When coding, assume that the grader has no knowledge of the language or API calls but can read comments. Use prolific comments before each section of code, or function call to explain what the code does, and why you are using it.

Create a directory named HW_NN_<LASTNAME>_<Firstname>_dir.
Do all your work in that directory. Zip up the entire directory and submit in one zip file.
The zip file should expand to one directory that contains: your code, your results, and so forth.

Your writeup should be called HW_NN_<LASTNAME>_<Firstname>.pdf, and other files should follow the same patterns so we can tell them apart. Substitute the homework number for NN.

The PDF file should stand alone: Answer the questions there, and include images of your results.

Feel free to look over each other's shoulders, 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.

Creating a One Rule – Feature Selection – using a threshold classifier:

The homework assignments build on each other. You can probably start with what you did for the last homework and modify it. Just be careful to watch for changes.

Read this:

- Please, read the entire assignment before you code anything.
- There are three attributes to consider:
  - Speed in miles per hour,
  - Number of lane changes in two minutes,
  - Brightness of the vehicle, from 0 to 10
- Of the attributes, select the one that gives the fewest number of mistakes.
- There will be questions about this homework later, on quizzes and exams.
- You will grow the skeleton of this homework later in other assignments.
- You will use the basis of this homework, to do the next homework as well.
  Make sure you make notes to yourself about what you did, so you can copy and reuse the code.

Given:
You are provided with a **new** data set, or sets.
The data includes the recorded speed, and the driver's intention:

The intentions for this homework are:

0 for not in any hurry at all, totally safe drivers
1 for being a normal driver, a bit anxious, but not being aggressive.
2 for drivers who are driving aggressively.
For this homework, this is your target variable.
Your goal is to predict when the intention is 2.

How to proceed:

1. Create a program named HW_NN_Lastname_FirstName.py, or .m, or .r, …
2. Read in all of the data:
   a. Speed in miles per hour.
   b. Number of lane changes in two minutes.
   c. Brightness of the vehicle, on a scale from 0 to 10.

3. Truncate the data to the nearest mile per hour.
   numpy.floor( ) will do the trick. That is, the bin size is 1 mile per hour for this assignment.

4. For each possible attribute:
   For each and every possible threshold speed, from low to high:
   a. Find the number of **non-aggressive** drivers who are > this threshold speed.
      This is the number of false alarms, or false positives, FA

   b. Find, the number of aggressive drivers who are <= this threshold speed.
      This is the number of false negatives, or FN.

   c. Find the number of **aggressive** drivers who are > this speed
      This is the number of **hits**, or **true positives**, TP.

   d. For this homework, the badness of each threshold is defined as:
      badness = FA + FN

   e. You want to minimize the badness.
      Find the threshold speed that minimizes this badness.
      Use this speed to create a classifier.

5. For this homework, have your program open another file named "HW_04_LastName_FirstName_Classifier.py" (Or whatever language you are using.). Have your program write out an entire classifier program.
   The classifier should use only one attribute for making its classification decision.

   The resulting program will need to use the correct attribute for classification.

   In other words, you will be given access to a test suite. You will run your program like this:

   ```
   python   HW_04_LastName_FirstName_Classifier   test_suite.csv.
   ```

   Your classifier should print out the number of cars in the test suite that are <= your selected threshold, and the number of cars that are > your selected threshold.

   This classifier code should pre-process the input data the same way your original code worked, or it will fail.
   In this assignment, your resulting program of code must parse and be runnable.

(continued)

Write-Up and Grading:

A. Create an output file called HW_NN_LastName_FirstName.docx, which gets printed to a PDF file later.
MAKE THE OBVIOUS SUBSTITUTIONS.

B. Put your name at the top, HW_NN_, and the course number.

C. COPY THE FOLLOWING QUESTIONS and ANSWER THEM:

D. Do not submit any data, nor the test_suite. I have copies of them.

E. In your PDF, show the resulting one-rule threshold classifier code.
What was your one-rule for this assignment?
Which attribute did it use?
It is just an if-statement.
Copy and paste your code from your output classifier file.
Please use black ink on a white background.
Do not use a screen capture, the fonts do not scale. (2)

F. Run your code to produce your classifier program.
Your code must produce the resulting classifier. (2)

G. Run your resulting classifier on the supplied test suite.
Your resulting classifier must run. (2)

H. Report how many Aggressive drivers did your classifier routine find in the test suite?
Have your classifier print out the number of cars <= the selected threshold.
Have your classifier print out the number of cars > the selected threshold.
Report these numbers in your write-up. (2)

I. **Conclusion:** Write up what you learned here using at least three paragraphs. (2)

What did you discover? Were the results what you expected? What was surprising?

Was there anything particularly challenging? Did anything go wrong?

Provide strong evidence of learning.

Write a conclusion that describes what you learned in this homework.
Points are taken off for writing with bullet points or checkmarks.

J. **Totally Optional Bonus Problem: (+1)**

If you are the kind of person who is stressed doing anything unnecessary, do not do this part.
This is just for those who enjoy a challenge, and want to exercise their abilities.

On one graph, plot the three different ROC curves for the three different possible threshold classifiers.
Use different line styles (solid line, dashed line, dotted line) for each attribute.
Use a legend or label so that we can see which attribute is associated with each curve.

Make sure you use square axes, so that it looks correct.
Given these ROC curves, does the resulting attribute selection make sense to you?

Penalty Rubric:
We expect that you can do all the things above. Getting the code in, in one directory, with the correct name on it, is what we call "table stakes". If you fail that, you lose points for that. You start with 10 points, and your grade goes down from there.

You do not submit supporting code.	(8)
Your code is handed in with a zip file that creates more than one directory when unzipped.	(2)
Your main program must create the classifier program, or you lose 5 points out of 10. (This is the main goal of these homeworks.)	(5)
Your main program must run.	(1)
Your main program is commented well.	(2)