Curses, Clustering, and Classification
See Dropbox for Due Date
Thomas B. Kinsman

Homework is to be programmed in Python, R, or Matlab. (There is a rumor that you need to use Python. That is hogwash. Matlab or R are much easier than Python, but only if you are familiar with them.)

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.

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

Create a directory named HWNN_<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 HWNN_<LASTNAME>_<Firstname>.pdf, and other files should follow the same patterns so we can tell them apart. Substitute the homework number for NN.

Feel free to look over each other's shoulders, at each other's work, but do you own work.
Do not hand in copies of each other's code.

1D Clustering and Classification using a threshold.

The homework assignments build on each other.
You will start with this homework for future homework assignments.

Read this:

- Read the entire assignment before you code anything.
- This is not too bad.
- There will be questions about this homework later, on quizzes and exams.
- **Remember:** 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.

How to proceed:

0. Create a program named HW_NN_Lastname_FirstName.py, or .m, or .r, …

1. Part A.
   Allocate 10,000 Gaussian random values, with zero mean, and a standard deviation of 1.0, allocate five vectors of these:
   The first vector we will call X, the second vector we can call Y, … Z, … S, … and T.
   a. For all of the X values, let dist = sqrt(x^2);
      Find the fraction of the data that is within 1 standard deviation of the origin.
      (i.e. What fraction of the data has a distance <= 1.0?)
   b. For all of the (X,Y) values let dist = sqrt(x^2 + y^2);
      Find the fraction of the data that is within 1 standard deviation of the origin.
      (i.e. What fraction of the data has a distance <= 1.0?)
   c. For all of the (X,Y,Z) values let dist = sqrt(x^2 + y^2 + z^2);
      Find the fraction of the data that is within 1 standard deviation of the origin.

   d. For all of the (X,Y,Z,S) values let dist = sqrt(x^2 + y^2 + z^2 + s^2);
      Find the fraction of the data that is within 1 standard deviation of the origin.

   e. For all of the (X,Y,Z,S,T) values let dist = sqrt(x^2 + y^2 + z^2 + s^2 + t^2);
      Find the fraction of the data that is within 1 standard deviation of the origin.

   f. Create a plot of the amount of data within one standard deviation of the origin, versus the number of elements in the vector. (The first point, for only one vector, should be around 68 percent.
      The numbers should drop off after that.

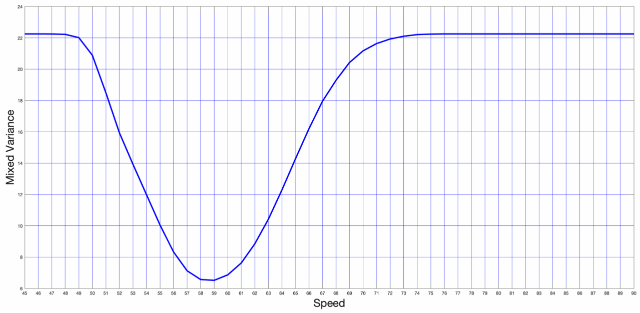
2. **Part B:**
   Read in all of the data for all the traffic stations.
   Quantize the data to the nearest one mile per hour. (Round in the data to the nearest mile per hour.)

   Ignoring the intention of the driver, plot the mixed variance of the two sets versus the threshold.
   This will help you solve Otsu's method.

   Here is how to compute Otsu's method.

   For each possible threshold speed, from low to high:
   a. Split the data into two groups:
      i. The left group is the data that is <= the threshold
      ii. The right group is the data that is > the threshold
   b. Compute the fraction of data in the left group.
   c. Compute the variance of the left group.
   d. Compute the fraction of data in the right group.
   e. Compute the variance of the right group.
   f. The mixed variance, for a given threshold, is b*c + d*e.
      The mixed variance is the fraction in the left group, time the variance of the left group,
      plus the fraction of the data in the right group, times the variance of the right group.

   Compute this for all thresholds and plot the values:



3. **Part C:**
   For each possible threshold speed, from low to high, using a linear search:
   a. Find the number of non-speeders who are > this threshold
      This is the number of false alarms.
   b. Find the number of speeders who are <= this threshold
      This is the number of misses, or false negatives.
   c. Find the number of mistakes total.
   d. On the same graph, plot, as a function of speed:
      i. The fraction of false alarms at each speed (in red)
      ii. The fraction of misses at each speed (in blue)
      iii. The fraction of mistakes at each speed (in magenta, or purple)

Write-Up and Grading:

A. Create an output file called HW_NN_LastName_FirstName.docx

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

C. COPY THE FOLLOWING QUESTIONS and ANSWER THEM:

D. What was the general trend as the number of elements in each vector went up,
what happened to the amount of data within one standard deviation of the origin?

In other words, as the number of attributes used to measure a data point increases, what happens to the density of the records? (2)

E. Show a bar graph of the entire histogram of speeds, by intention. Remember that the speed limit is 55 mph. How might we describe this data? Is it a mixture model? What kind of mixture model?
What do you notice about it? What is odd?
Speculate about why various lumps are the way they are? (2)

F. In your PDF, show your graph of Mixed Variance versus speed. (2)
This could be used to break the drivers into two groups.
What speed would you use to split the drivers into two groups?

G. In your PDF, show your graph of the numbers of false alarms, misses, and mistakes as a function of speed. You want to minimize the number of mistakes.
What one-rule would you use to decide if a driver was trying to speed? (2)

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

What did you discover?
Was anything unusual?
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.

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 run.	(1)
Your main program is commented well.	(2)