Principles of Data Mining CSCI-420 – HW04
Due Tuesday Night, February 10th, 2015 11:59 pm
T. Kinsman

Homework is to be programmed in Java, Python, Matlab, or R. It is mostly one function.

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 associated dropbox. Submit two files, one named: HW02_<LASTNAME>_<Firstname>_results.pdf, and the other named HW02_<LASTNAME>_<Firstname>_program.extension.

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.

1. (½ pts) Before you start, read the following description, and guess how long it will take you to finish the homework. Don't cheat. This isn't for me, this is for you to realize how long it *actually* takes to do homework. Guess how many hours it will take. Being off by a factor of 6 to 12 is not uncommon.

2. (5 pts) 1D Clustering using Otsu's method.

   You should get results that are similar to what you saw in the lectures, but not the same values, because you have different data.

   You are supplied with a set of speed observations for 128 cars, in the file UNCLASSIFIED_Speed_Observations_for_128_vehicles.txt.

   a. **Ethics:** We believe that these vehicles are composed of two underlying (latent) groups: those who are intentionally speeding (reckless drivers), and those who are trying to maximize safety, conserve fuel or perhaps waste time.

      You are developing a machine with studies traffic volume for road planning in order to maximize traffic flow. Do you have any ethical issues with doing this?

   b. You are being paid to develop a computer vision machine that will automatically send a speeding ticket to reckless drivers. Do your ethical considerations change?

   Write a program to:
   c. Quantize the vehicle speeds into bins that are [38 up to 40), [40 up to 42), … up to 80 mph.
   d. Implement Otsu's method to separate the vehicles into two clusters.
      That is, we are using Otsu's method to binarize the data. There are other methods.
      We are quantizing the data into two groups.

   e. What speed should we use to best separate the two clusters?
   f. What is the minimum mixed variance that resulted?
   g. Breaking Ties: How would your program handle a situation where the minimum mixed variance occurred twice? Does this situation happen?
   h. It might help to plot a histogram of the quantized vehicle speeds.

3. (3 pts) Exploratory Data Analysis:

You are also provided with some mystery data, in the file MysteryData.txt. It consists of two underlying groups. This data is pre-quantized to the nearest unit.

   a. What are the mode, median, and average values of this data?

   b. Remove the last value from the data, the 16, how do the mode, median, and average values change?

   c. Use your Otsu's clustering routine to split this data into two groups.
   What threshold best splits the data into two groups?
   What was the minimum mixed variance that resulted?

4. (3 times ½ pts) Reality versus perception.
   a. Report how long did it actually take you to do the homework (in hours)?

   b. Report the initial guess divided by the actual result to 2 significant figures.

   c. In a few sentences, what factors do you suppose there are that make it difficult to predict the time it takes to write software? [Can you find any references that support this?]

5. (1 pt) BONUS

   Plot a graph of the mixed variance for the car data in question 2,
   versus the value used to segment the data into two clusters.
   Clearly label the axes.