Agglomeration You write yourself, and
Data Mining

Please read over the entire homework before starting. I'm vastly simplifying this to make our lives simpler. Work with a partner. You will need the both of you to figure out how to make Agglomeration run fast.

On my simple notebook computer, my code runs in 22 seconds, in Matlab. Matlab is an interpreted language, so it is slower than other languages. However, some students have complained that it takes them hours to have agglomeration run on their system. PLAN AHEAD. Think in optimal ways.

On the top of your homework should be the names of the two of you who worked together on this. Work together as paired partners, do your own teamwork. Let me know whom you worked with. Put both names on the homework, and make **one** submission to the dropbox. Hand in one copy of your team's code and write-up. You should be able to answer questions about your code if you see it again later. Or, you should be able to outline your solution if asked for it later on.

In the past, I have had students that were working in teams not know anything about the results. That is inexcusable. That means one person did all the work. The other one does not do well on exams.

Hand in your results, and the well-commented code, in the associated dropbox of one student of your team.

Think of this as a big "lab assignment", that you and your partner will work on.

As always, **Use prolific comments** before each section of code, or complicated function call to explain what the code does, and why you are using it. **Do not use single letter variable names.** Even FORTRAN allows long variable names now.

There is an extra, across the board, 25% penalty in this assignment for code that cannot be easily read. I don't want to pay graders to try to decipher hieroglyphics in your code. Use clear variable names and comments.

Britney's Boutique Bookstore:

Assume you work at BBB. **BBB** tracks each receipt by "Guest ID". In order to improve their statistics on the guests, we have consolidated 10 of each guest's most recent visits into a single record for 10 purchases. A data file will be provided for you at the usual place.

Note that: using the data on 10 trips instead of just one is a form of **noise reduction**. We remove small changes in the data that we don't care about by merging together several records. This improves the "signal to noise" ratio.

BBB already knows that many of their guests are family purchases. However, there must be other groups.

Your task is to identify the groups, and give them a "stereotype" code-name for the marketing department to use. If they exist, we need to know how their shopping trends differ from the other groups. What makes them special? What should we send them coupons for? Effectively, you are identifying each "prototype" shopper for the marketing department to pay attention to.

The details of your assignment follow: To simplify grading, the assignment must be very specific.

You are provided with the file **`HW_CLUSTERING_SHOPPING_CART_....csv`**. It contains data for the number of times various categories of items (attributes) were purchased by each guest, for 10 different visits.

Part A: Using Cross-Correlation for Feature Rejection and Selection:

In your write-up, copy the questions before answering them for more accurate grading.

1. **Using a package**, compute the cross-correlation coefficients of all attributes.
   Your resulting matrix should be n by n where n is the number of attributes.

   All values computed should be in the range [-1, 1].

   Record these values with two digits past the decimal point.

   [Please, if there is a God, do not let the students compute the cross-correlation of the data with the Record ID still in it. If you are reading this, be sure your partner does not make that mistake. I know you are smart enough not to do this, but your partner might not be. So, always check.]

2. Report: (1/10) point each)
   a. Which two attributes are most strongly cross-correlated with each other?
      Hint: It is the absolute value of the cross-correlation that matters, not how positive it is.
      [And don't forget to repeat these questions.]

   b. If someone buys lots of Manga books, else are they likely to buy (or not buy, depending on what is strongest.). In other words, what is the strongest Absolute value of CC with any other category? Are they also likely to buy Horror or Gifts? Or, are they NOT likely to buy Thrillers? Do the analysis and report the answer here.

   c. What other category is Fiction most strongly correlated with?

   d. What other category is Self Improvement most strongly correlated with?

   e. If someone buys cookbooks, what can you tell about them?

   f. If someone buys lots of classic novels, what can you tell about them?

   g. If someone buys NEWS, what can you tell about them?

   h. What do you know about people who buy Hairy Pottery?

   i. What are Thrillers most strongly associated with, or not associated with?

   j. What can we infer about people who buy Art & History books?

3. If you were to delete three attributes, which would you guess were irrelevant?
Why? (1)

Part B: Agglomeration:

4. Implement agglomerative clustering by yourself. **Do not use a package.**
   Cluster the guests into groups as follows:
   a. At the start of agglomerative clustering, assign each record to its own cluster prototype.
      Suppose we have 1000+ records.
      So, you start with 1000-plus clusters and 1000-plus prototypes of those clusters.
   b. Use the **Manhattan** distance between cluster centers as the distance metric.
   c. Use the center of mass as the prototype center, the center of mass of a set of records, to represent its center location in data space. And use the distance between these centers as the linkage method.
   d. Note: At each step of clustering, two clusters are merged together.
      Track the size of the smallest of the two clusters that are merged together.

      There are questions about this later. Write down the size of the smallest cluster in the last 20 merges.
      For example, if we merge a cluster of size 30 with a cluster of size 10, you remember that a 10 was merged in. Cluster to completion.
      Record and report the size of the last 10 smallest clusters merged.

   e. Based on agglomeration, how many clusters do you think are in the data? Why did you reach this conclusion? Support your guess. Can you support this guess with a dendrogram?

Discussion Questions – Copy and paste all questions so that you can understand the context of your answers later on, and so that the grader knows which question you are answering.

5. Report the size of each suspected cluster, from smallest to largest size.
6. Report the average prototype of each of these the clusters.
7. What typifies each of the clusters? What typical names should we give each of these prototypes? Is there a family group? Is there a gift-giving group? What typifies each group?

Guidelines and hints for agglomeration:

a. Do not start with all data points. Start with a subset, a test-suite. Design your test suite so that you know the answer before you start.

b. You need to keep track of all the records (guest id) that belong to each cluster. This is necessary because after each merge, you need to compute the new average (center of mass) of the entire cluster. This drifts after each merge.

a. You need a separate data structure for each cluster's center of mass – the cluster prototype.

b. It is convenient to have a data structure that records which cluster each record (guest id) is assigned to. This is the answer you are looking for ultimately when you get down to a few final clusters.

c. You may want a separate data structure for the cluster's ID to make your life easy.

d. For big data, to be computationally efficient, you only need to re-compute the distances involved with the two clusters that are merged, to all other clusters. For this assignment, you might need to forget about being computationally efficient – it is painful to debug. Just re-compute all the distances between all the clusters on every pass.

e. However, if you compute the distance from record ii to record jj, you do not need to compute the distance from record jj to record ii. Be smart. Don't take forever.

f. You need some way to select the shortest inter-cluster distance, without accidentally selecting the distance from a cluster to itself. [Otherwise, you will loop forever. It is always wise to avoid infinite loops. Just my suggestion, you do what you like...]

g. It is convenient to have the lowest cluster labels persist through the progression, so that when you merge cluster 19 and cluster 95, the resulting cluster is now labeled 19.
The final result is one cluster of all the data labeled "cluster 1."

h. At each merge stage, you need to keep track of several things.
Update everything carefully.

8. Dendrogram :

Create a Dendrogram, showing the last 20 clusters. Caution: other packages use different names for the linkage methods. Watch out. Some confuse central and average. Check your results with known test suites.

There are many tutorials on the web for plotting. You can use a tool or package for plotting a dendrogram. You can do it in R in about 10 lines. There are facilities in python to do this. You only need to show the top 20 clusters for the dendrogram to make sense. Does your dendrogram agree with your previous conclusions about cluster sizes?

By the way, I recommend you do the dendrogram first. It will give you valuable hints about the inherent structure in the data. Again, this is just suggestion, you do what you like. ☺

Part C: Write a summary and conclusion overall.

9. Write a conclusion about what you learned overall.

If each of you learned different things, tell me what each of you learned.

In college a one paragraph conclusion does not cut the mark. I expect at least half a page of significant conclusions. If I grade your work, I might only read the conclusion.