Agglomeration You write yourself, and K-Means using a package, Data Mining

Work with a partner. Please read over the entire homework before starting. I'm vastly simplifying this to make our lives simpler.

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.

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.

Sam's Spiffy Supermarket:

Assume you work at SSS – (Sam's Spiffy Supermarket). SSS 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: this is a form of noise removal. We remove small changes in the data that we don't care by merging together several records. This improves the "signal to noise" ratio.

SSS already knows that many of their guests are family purchases. In addition, SSS also suspects several other groups: maybe Hispanic food buyers, vegetarian food buyers, a group who eats a lot of fish, and gluten-free shoppers. And, they think there is another group of "party animals" that only buys groceries when they cannot find free food on campus.

Your task is to identify the groups, and give them a shortcut 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_PCA_SHOPPING_CART_v….csv`. It contains data for the number of times various categories of items (attributes) were purchased by guests, for 10 different visits.

Part 1: Using Cross-Correlation for Feature Rejection and Selection:

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

1. Compute the cross-correlation coefficient of all attributes. Use a package to do this.
   Your matrix should be n by n where n is the number of attributes.

   All values computed should be in the range [-1, 1].
   [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 this mistake. I know you are smart enough not to do this, but your partner might not be. So, always check.]

2. Report:
   a. Which two attributes are most strongly cross-correlated with each other? (¼)
   [Don't forget to repeat the questions.]

   b. Which attribute is fish most strongly cross-correlated with? (¼)

   c. Which attribute is meat most strongly cross-correlated with? (¼)

   d. Which attribute is beans most strongly cross-correlated with? (¼)

   e. Which one attribute is least correlated with all other attributes? (¼)

   f. Which second attribute is least correlated with all other attributes? (¼)

   g. If you were to delete two attributes, which would you guess were irrelevant? (¼)

   h. If buying fish is strongly cross-correlated with buying cereal, and buying cereal is strongly cross-correlated with buying baby products, is buying fish strongly cross-correlated with buying baby products? Can you explain this? (¼)

Part 2: Agglomeration:

3. 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.
      So, you start with 800-plus clusters and 800-plus prototypes of those clusters.
   b. Use the Euclidean 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. (5) 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 18 smallest clusters merged.
   e. (3) 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?

Guidelines and hints for agglomeration:

a. Use as much code as you can from previous homework assignments.
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 six 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, forget about being computationally efficient – it is painful to debug. Just re-compute all the distances between all the clusters on every pass.
e. 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.)
f. 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."
g. At each merge stage, you need to keep track of several things. Update everything carefully.
h. There are many tutorials on the web for 3D plotting, including:matplotlib.org/mpl_toolkits/mplot3d/tutorial.html

Discussion Questions – Copy and paste these so that you can understand the context of your answers later on:

1. (1) When you have clustered to six clusters, report the size of each cluster, from lowest to highest.
2. (1) When you have clustered to six clusters, report the average prototype of these six clusters.
3. (1) What typifies each of the six clusters? What name should we give each of these prototypes?
4. (5) Write a conclusion about what you learned overall.
   If each of you learned different things, tell me what each of you learned.

Part 3: K-Means:
Using a package, such as sklearn, find six clusters of the data. Compare the cluster centers and numbers found using k-Means to those found using Agglomeration.

Part 4: Write a conclusion overall.