Principles of Data Mining CSCI-720 – HW09 – Agglomerative Clustering
Due Monday March 14th, 2016 11:59 pm
Thomas Kinsman

Homework is to be programmed in R, Python, Java, or Matlab. (Not Excel.)

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

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

Worked with Manan Shah.
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. You should be able to answer questions about your code if you see it again later.

ASSIGNMENT:
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.

SSS already knows that most of their guests are family purchases. And, they have a second group of "party animals" that only buy groceries when they cannot find enough free food on campus.

In addition, SSS also suspects a third, yet unidentified, hidden group.

Your task is to identify this group, 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 two groups. What makes them special? What should be send them coupons for?

Effectively, you are identifying a third "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.

1. You are provided with the file `HW_09_SHOPPING_CART_v037.csv`. It contains data for the number of times various categories of items (attributes) were purchased by 33 guests, for 10 different visits.

   We used 10 different visits to help get enough data on each shopper to start to draw conclusions.

2. Implement agglomerative clustering to cluster the guests into 3 groups as follows:
   a. At the start of Agglomerative clustering, assign each record to its own cluster prototype. So, you start with 33 clusters and 33 prototypes of those clusters.
   b. Use the Euclidean distance between clusters as the distance metric.
   c. Use the 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.
   Record the size of the smallest of the two clusters that are merged together. There are questions about this later.
e. Cluster to completion and answer the following questions:

3. Guidelines and hints: These are only hints. Your approach might be different.
   a. 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.
   b. You need a separate data structure for each cluster's center of mass.
   c. It is convenient to use 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 three final clusters.
   d. You may want a separate data structure for the cluster's ID to make your life easy.
   e. 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 recompute all the distances between all the clusters on every pass.
   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.)
   g. It is convenient to have the lowest cluster labels persist through the progression, so that when you merge cluster 19 and cluster 29, the resulting cluster is now labeled 19. The final result is one large cluster labeled "cluster 1."
   h. At each stage, you need to keep track of several things. Update everything carefully.

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

1. Estimate how long this assignment will take. (½) (In prep for question #9 below.) **20 hrs.**

2. You will need to remove one of the attributes. Which one should you *always* remove? (½) **We will always need to remove the ID for the customers.**

3. You can keep all the other attributes, or remove more. Which attributes did you finally use? (½) **Used all the 12 attributes from Milk till Fruit.**

4. Submit code that shows the clustering process. (4 points, including code readability)

5. At each stage of clustering (from stage 1 to 32), what was the size of smaller cluster that was merged in? What does this indicate about the true number of clusters? (½)
   Iteration 1: 1 (16,25)
   Iteration 2: 1 (17,23)
   Iteration 3: 1 (17,23,10)
   Iteration 4: 1 (8,11)
   Iteration 5: 1 (0,9)
   Iteration 6: 1 (11,19)
   Iteration 7: 1 (0,16)
   Iteration 8: 1 (1,11)
   Iteration 9: 1 (4,8)
   Iteration 10: 1 (4,8,11)
   Iteration 11: 1 (6,7)
   Iteration 12: 1 (0,6,7)

6. When you have clustered to three clusters, report the guest id's in each of these three clusters. (1)
   Checked by putting data in R.
   Cluster 1 : (8,4,22,9,13,12,1)
   Cluster 2 : (14,15,30,26,24,25,2,32,17,8,31,19,18)
   Cluster 2 : (20,7,28,29,16,32,10,21)

7. What typifies the third cluster? What nick-name should we give these customers? (be polite) (1)

8. If we switched from "central link" to a "single link" merge step, what would you need to add to the algorithm when computing the distance between two clusters? (1)

9. How long did this assignment take? (½) **30 hours.**

10. Write a short answer question for the next midterm exam. If your question is used, you get the points on the exam. Part of the reason I ask this is to be sure you think about the questions that might be on the next exam. (½)

    Question based on which distance to minimize and which to maximize for agglomerative clustering.

11. Bonus: Generate a dendrogram of the clusters as they are being merged. (1) Show the code that demonstrates your understanding of this.

```
#Read file
setwd("E:\BDA Homework\HW09-Agglomerative Clustering")
HW_09_SHOPPING_CART_FOR_HUMAN_v037<-
read.csv("HW_09_SHOPPING_CART_FOR_HUMAN_v037.csv")
HW01_Agglomarative_data<-HW_09_SHOPPING_CART_FOR_HUMAN_v037
#Remove the Id column as we would not be using that.
temp<-data[,-1]
#Defaut method used is Eucledean distance
distance_matrix<-dist(temp)
#Hierarchical Clustering
cluster<-hclust(distance)
#Plot the data points
plot(cluster)
```