HW 09 Hierarchical Clustering
Latish Khubnani
Oct. 16th 2016

1. Estimate how long this assignment will take.
   It will take me 5 hours to complete this assignment.

2. You will need to remove one of the attributes in the CSV file. Which one should you certainly always remove?
   I will remove the ID or any attribute which is unique and is used to identify the data entry.

3. You can keep all the other attributes, or remove more. Which attributes did you finally use?
   I used all the attributes
4. Submit code that shows the clustering process. (4 points, including code readability)
   Please check the zip file.

5. At each stage of clustering (from stage 1 to 99), what was the size of smaller cluster that was merged in? What does this indicate about the true number of clusters?
   The smallest size that was merged in was 1.

6. When you have clustered to three clusters, report the guest id's in each of these three clusters.
   Cluster 1 : [7, 9, 10, 13, 17, 18, 20, 26, 27, 28, 30, 31, 32, 41, 47, 48, 50, 52, 56, 57, 59, 63, 67, 69, 72, 78, 79, 93, 94, 95, 97, 98, 100]

   Cluster 2 : [33, 4, 71, 73, 43, 12, 45, 44, 14, 76, 84, 21, 22, 54, 23, 60, 29]

   Cluster 3 : [1, 2, 3, 5, 6, 8, 11, 15, 16, 19, 24, 25, 34, 35, 36, 37, 38, 39, 40, 42, 46, 49, 51, 53, 55, 58, 61, 62, 64, 65, 66, 68, 70, 74, 75, 77, 80, 81, 82, 83, 85, 86, 87, 88, 89, 90, 91, 92, 96, 99]

7. What typifies the third cluster? What nick-name should we give these customers? (be polite)
   Observations:
   a. Summarizing the data from based on the ids in the clusters. People from cluster one have more Veggies, Cereal, Nuts, Rice, Yogurt and Fruits than milk, beer, meat and eggs.
   b. The second cluster of people have more beer, nuts, cereal and chips than veggies and fruits and this group is smallest in number.
   c. The third cluster has more cereal than all others and also buys everything else in numbers which are close to each other.

   From the observation above I can identify the the third cluster to be family cluster, the second one who consumes more alcohol and easy foods like nuts and cereal as party animals. The first group is using lots of veggies and nutrient rich foods and also they seem to buy less milk and meat, I would suggest these people are vegan/ vegetarians. We should provide them with coupons which encourage them to buy more cereal and provide discount on veggies as well.

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?
   I would need to change the function which calculates the average distance to the one which finds the shortest pairwise distance amongst the clusters (distance is calculated between points from each cluster)

9. How long did this assignment take?
   It took me 6.5 hours to complete the assignment

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.

Briefly explain the Silhouette Coefficient and give the formula for same
A: Silhouette Coefficient combine ideas of both cohesion and separation for individual points

S =1–a/b if a<b where

a = average distance of point to the points in the same cluster

b = min(average distance of point to points in any other cluster)