Tyler Paulen
10/26/2015
CSCI420-1
HW 6

1. Who did you work with? (½ points)

I worked alone.

2. You will probably need to toss one of the attributes. Which one? (Hint: you would always toss this attribute of each record, and would never want to use it for anything.) (½)

The ID would be tossed since it has no correlation with any of the other cart combinations. Cereal would also be tossed. It had the lowest standard deviation, so it was bought or not bought the most consistently.

3. Which attributes did you finally use? (½)

All attributes but cereal were used.

4. When you have clustered to three clusters, report the guest id's in each of the final three clusters. Compare with fellow classmates to assure you have a common agreement. (5 points, including code readability)

The clusters by ID are as follows:
1.0, 22.0, 4.0, 9.0, 13.0, 12.0, 8.0
2.0, 32.0, 5.0, 11.0, 19.0, 25.0, 18.0, 31.0, 17.0, 26.0, 14.0, 30.0, 24.0, 27.0, 15.0, 6.0
3.0, 29.0, 10.0, 21.0, 23.0, 16.0, 28.0, 33.0, 20.0, 7.0

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

The third cluster likes meat rice and nuts the most. The buyers like high protein meals, and would have a physically active lifestyle. Let's call this group the gym-goers

6. 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? (1)

The size of the smaller cluster was 7. The size of the cluster shows that the true number of true clusters is 3 because the cluster sizes were so large.

7. If we switched from "central link" to a "single link" merge step, what would you need to add to the algorithm that computes the distance between two clusters? (½)

The algorithm would need to loop through each cluster, and find the point that is closest to another point within a cluster, rather than just look at the cluster Center of Mass. This would increase the amount of time the algorithm would take.

Tyler Paulen
10/26/2015
CSCI420-1
HW 6

8. How long did this assignment take? (½)

> 4 hours.

9. Write a short answer question for the next midterm exam. As if your question is used, you get the points on the exam. Part of the reason

> Sometimes in a Naive Bayes problem, a probability can be zero. What can we do to solve this, and what is this problem called?