Principles of Data Mining (420)
Spring 2022, Homework 06

Mohammed Mehboob
(mm2260@rit.edu)

March 23, 2022

Question-2:

(a) Which two attributes are most strongly cross-correlated with each other?

⇒ Veggies and Soda are the most strongly cross-correlated attributes in our data, with a cross-correlation of -0.83.

(b) What is the cross-correlation coefficient of Chips with cereal?

⇒ The cross-correlation coefficient of Chips with Cereal is 0.19

(c) Which attribute is fish most strongly cross-correlated with?

⇒ Fish is most strongly cross-correlated with Chips, with a coefficient of 0.23

(d) Which attribute is Veggies most strongly cross-correlated with?

⇒ Veggies are most strongly cross-correlated with Soda, with a coefficient of -0.83

(e) According to this data, do people usually buy milk and cereal?

⇒ The cross-correlation coefficient of Milk with Cereal is 0.012, which is fairly low. Hence, people who shop at SSS do not buy Milk and Cereal together.

(f) Which two attributes are not strongly cross-correlated with anything?

⇒ Taking a look at the heatmap, we can identify right away that Cereal, Eggs, Fruit, Beans and Salt have very low absolute cross-correlation coefficients across the board.

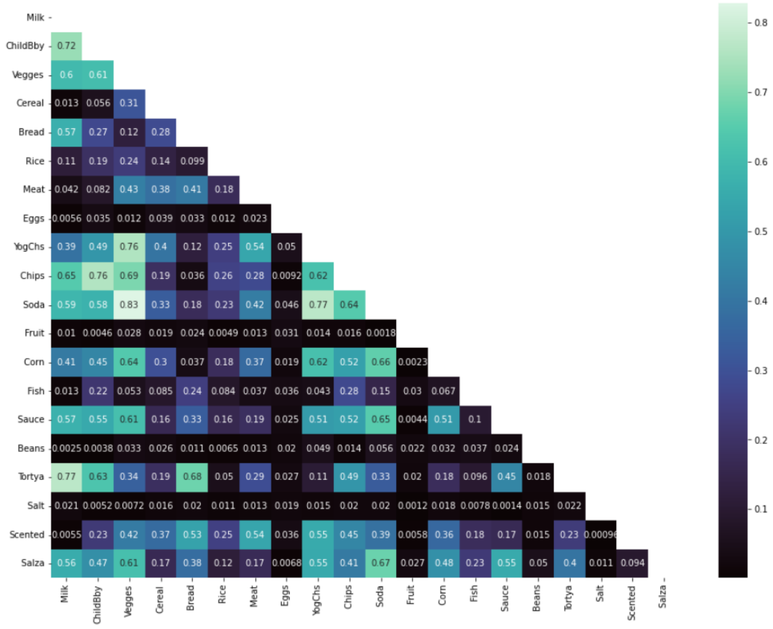


Figure 1: Cross-Correlation Coefficients Heatmap

As we can see in Figure-2, **Salt** and **Fruit** have the lowest cross-correlation coefficients out of all the attributes. Hence, these will be the two attributes which are not strongly cross-correlated with anything.

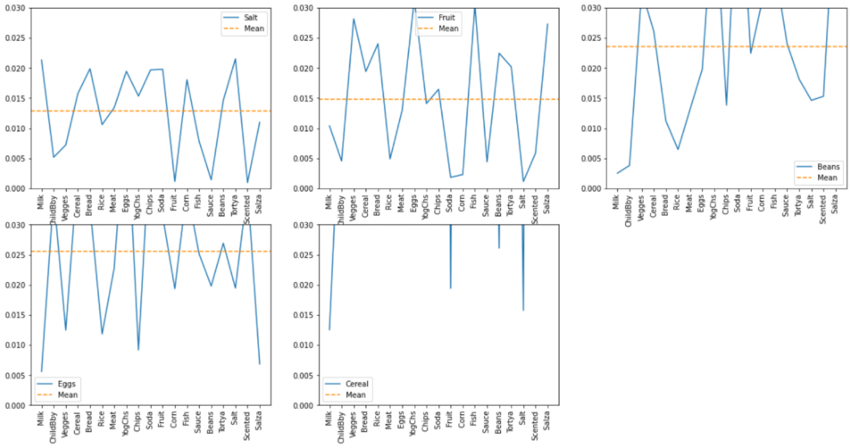


Figure 2: Comparing Cross-Correlation Coefficients to find Least Correlated Attributes

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

⇒ *We would remove variables which are highly correlated.* This is because we would mostly learn the same amount of information from them, given that both variables are independent, and it would allow for reduction in dimensionality which directly impacts performance.

Thus, if we were to delete two attributes from the SSS data, we could utilize the following data to make the decision:

Attribute-1	Attribute-2	Cross-Correlation Coefficient
Veggies	Soda	0.83
YogChs	Soda	0.77
Milk	Tortilla	0.77

Since Soda and Veggies are highly correlated, we could remove Veggies, since Soda would be a decent substitute for the same information.

And since the next on the table is YogChs, which is also highly correlated with Soda, we could remove YogChs. Alternatively, we could also consider removing Milk since it is highly correlated to Tortilla.

To simplify matters, we can remove **Veggis and YogChs** .

(h) If buying fish is strongly cross-correlated with another item, and buying that item is strongly highly cross-correlated with a third item, is buying fish strongly cross-correlated with the third item?

⇒ Let us assume that strong cross-correlation between variables is a transitive property.

So if Fish is strongly cross-correlated with X, and X is strongly cross-correlated with Y, then Fish will be strongly cross-correlated with Y.

However, we know that:

Fish is strongly cross-correlated with Bread (cross-correlation coefficient = 0.24). Bread is strongly cross-correlated with Milk (cross-correlation coefficient = 0.57). However, the cross-correlation between Fish and Milk is 0.013, which is fairly low.

Thus, strong cross-correlation between variables is not a transitive property. It may be true at times, but it is not always the case.

Question-7: Implement Agglomerative Clustering by hand.

(d) Report the size of the last 20 smallest clusters merged.

Cluster-A ID	Cluster-B ID	$min(\lvert$Cluster-A$\rvert$, $\lvert$Cluster-B$\rvert)$	Centroid Linkage Distance (Euclidean)
268	2389	1	8.68
849	2363	1	8.71
1064	2379	1	8.81
57	2390	1	8.98
1123	2393	1	9.01
805	2394	1	9.17
969	2395	1	9.82
2392	2396	250	11.44
2391	2397	275	11.44
2386	2498	375	14.51

(e) Based on the previous answer, how many clusters do you think are in your data?

⇒ Based on the increase in minimum cluster size between the two being merged for the last three merges, and the merges coming before being between a singleton cluster and a larger cluster, we can say that there are three clusters in our data.

This is also reflected in the dendrogram:

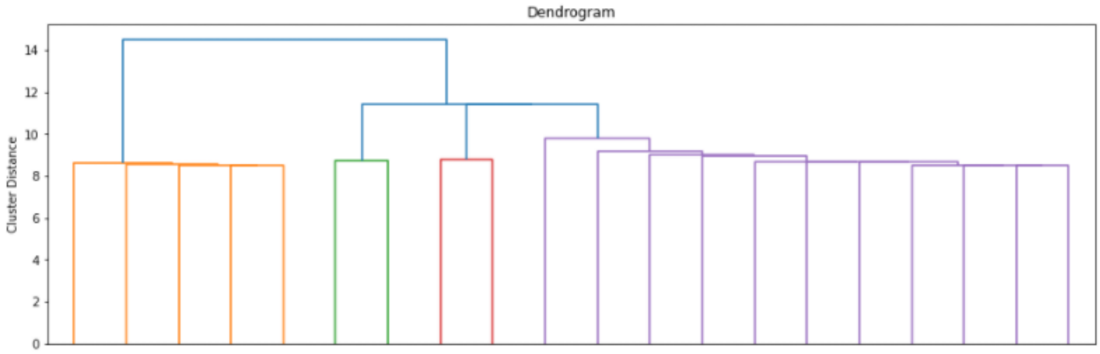


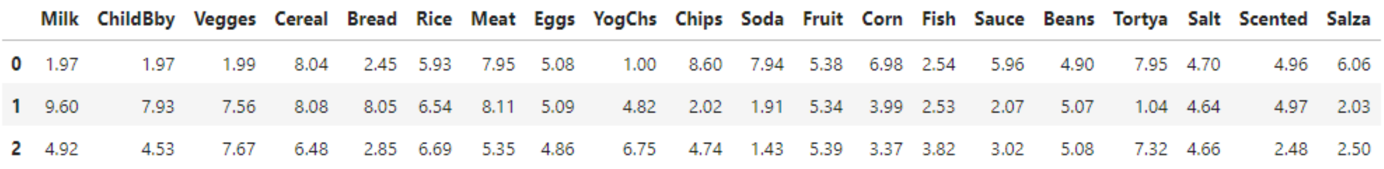
Figure 3: Dendrogram for Agglomerative Hierarchical Clustering

Question-8: Report the size of each cluster, from lowest to highest.

⇒ Taking into consideration the three sthat we say are the correct way of clustering our data for SSS, they will have the following sizes in terms of the original observation that they consist of from our test data:

Cluster Number	Cluster Size
1	275
2	375
3	550

Question-9: Report the average prototype of each of the clusters.

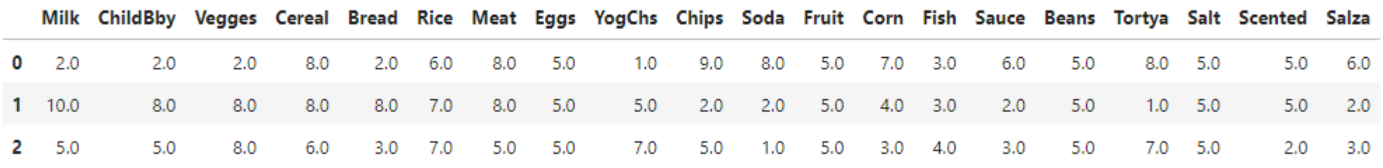


	Milk	ChildBby	Vegges	Cereal	Bread	Rice	Meat	Eggs	YogChs	Chips	Soda	Fruit	Corn	Fish	Sauce	Beans	Tortya	Salt	Scented	Salza
0	1.97	1.97	1.99	8.04	2.45	5.93	7.95	5.08	1.00	8.60	7.94	5.38	6.98	2.54	5.96	4.90	7.95	4.70	4.96	6.06
1	9.60	7.93	7.56	8.08	8.05	6.54	8.11	5.09	4.82	2.02	1.91	5.34	3.99	2.53	2.07	5.07	1.04	4.64	4.97	2.03
2	4.92	4.53	7.67	6.48	2.85	6.69	5.35	4.86	6.75	4.74	1.43	5.39	3.37	3.82	3.02	5.08	7.32	4.66	2.48	2.50

Figure 4: Average Prototypes

Question-10: What typifies each of the clusters? What typical names should we give each of the prototypes? Is there a gluten-free group? Is there a family group? Is there a group of party animals? Are there vegans? Are there healthy eaters? What typifies each group you found?

Let's round the prototype averages to the closest integer, so that we can make it a little easier to assess each of them individually:



	Milk	ChildBby	Vegges	Cereal	Bread	Rice	Meat	Eggs	YogChs	Chips	Soda	Fruit	Corn	Fish	Sauce	Beans	Tortya	Salt	Scented	Salza
0	2.0	2.0	2.0	8.0	2.0	6.0	8.0	5.0	1.0	9.0	8.0	5.0	7.0	3.0	6.0	5.0	8.0	5.0	5.0	6.0
1	10.0	8.0	8.0	8.0	8.0	7.0	8.0	5.0	5.0	2.0	2.0	5.0	4.0	3.0	2.0	5.0	1.0	5.0	5.0	2.0
2	5.0	5.0	8.0	6.0	3.0	7.0	5.0	5.0	7.0	5.0	1.0	5.0	3.0	4.0	3.0	5.0	7.0	5.0	2.0	3.0

Figure 5: Average Prototypes (Rounded to nearest integer)

- We can see that the first cluster (table Index-0) tends to buy things like Cereal, Rice and Meat, Chips, Soda, Tortillas and such more than other things. We could potentially typify this cluster as somewhat of a college student or party animals perhaps.

- The second cluster (table index-1) tends to buy items like Milk, ChildBby, Veggies, Cereal, Bread, Rice, Meat, and comparatively, tends to not buy items like Soda, Chips, Sauce or Tortillas. This could be typified as a shopper who tends to buy groceries for regular/daily use, and perhaps someone who tries to stay healthy rather than not.

- The third cluster (table index-2) tends to buy everything in proportion, so this might be something along the lines of a family group typification. Where perhaps someone in the family does not eat bread as much, or corn; And where soda is avoided.

Overall Conclusion:

I appreciated how the assignment walked me through the process of discovering how feature selection works. Understanding concepts like dropping highly cross-correlated attributes, since they only add calculation overhead and can very well be removed and the variable they were highly cross-correlated with can act as a substitute.

It was also cool to see, and to try to prove that strong cross-correlation is not transitive. In trying to understand this, I came across a post by Terrence Tao on the topic, which was farily cool (https://terrytao.wordpress.com/2014/06/05/when-is-correlation-transitive/).

Since I was using centroid linkage for the agglomerative hierarchical clustering, it yielded a very interesting result when I was performing computations on the 'D' csv data instead of the 'H' csv data, where the distance was not monotonic. What happened was that the newly formed cluster yielded a shifted center of mass / centroid, and that centroid yielded a lower distance. So the resulting dendrogram had a strange shape. This prompted me to add the second, package-generated, dendrogram to my code. Since, I was nearly convinced that I was doing something wrong. Seeing the package-generated dendrogram showed the same result was of some relief.

Speaking of dendrograms, I had the worst time – I assigned a new cluster ID to every newly merged cluster. Nonetheless, my cluster IDs started indexing from 1. So when I generated my linkage-matrix, scipy would prompt me with an error saying that I was using 'non-singleton' clusters. This was very frustrating, since I couldn't for the life of me, figure out what was going wrong. Eventually, after a lot of pain and suffering, I figured out that all I needed to do to fix it was to begin my cluster IDs from 0.

Overall, Typification felt fairly difficult to me. It was not super intuitive, since it could be anything out of a few different possiblities, except for the very obvious cases.