HW06
Authors: Prionti Nasir, Justin Palmer

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

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

1. Using a package, compute the cross-correlation coefficients of all attributes. Your resulting matrix should be n by n where n is the number of attributes. All values computed should be in the range [-1, 1]. Record these values with two digits past the decimal point. [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 that mistake. I know you are smart enough not to do this, but your partner might not be. So, always check.]

	Milk	ChildBby	Vegges	Cereal	Bread	Rice	Meat	Eggs	YogChs	Chips	Soda	Fruit	Corn	Fish	Sauce	Beans	Tortya	Salt	Scented	Salza
Milk	1.00	0.72	0.60	0.01	0.57	0.11	0.04	0.01	0.39	-0.65	-0.59	-0.01	-0.41	-0.01	-0.57	-0.00	-0.77	-0.02	0.01	-0.56
ChildBby	0.72	1.00	0.61	-0.06	0.27	0.19	-0.08	-0.04	0.49	-0.76	-0.58	-0.00	-0.45	-0.22	-0.55	0.00	-0.63	-0.01	-0.23	-0.47
Vegges	0.60	0.61	1.00	-0.31	0.12	0.24	-0.43	-0.01	0.76	-0.69	-0.83	0.03	-0.64	0.05	-0.61	0.03	-0.34	-0.01	-0.42	-0.61
Cereal	0.01	-0.06	-0.31	1.00	0.28	-0.14	0.38	0.04	-0.40	0.19	0.33	0.02	0.30	-0.09	0.16	-0.03	-0.19	0.02	0.37	0.17
Bread	0.57	0.27	0.12	0.28	1.00	-0.10	0.41	0.03	-0.12	-0.04	-0.18	-0.02	-0.04	0.24	-0.33	0.01	-0.68	-0.02	0.53	-0.38
Rice	0.11	0.19	0.24	-0.14	-0.10	1.00	-0.18	-0.01	0.25	-0.26	-0.23	-0.00	-0.18	-0.08	-0.16	0.01	-0.05	0.01	-0.25	-0.12
Meat	0.04	-0.08	-0.43	0.38	0.41	-0.18	1.00	0.02	-0.54	0.28	0.42	-0.01	0.37	-0.04	0.19	-0.01	-0.29	-0.01	0.54	0.17
Eggs	0.01	-0.04	-0.01	0.04	0.03	-0.01	0.02	1.00	-0.05	0.01	0.05	0.03	0.02	-0.04	0.03	-0.02	-0.03	-0.02	0.04	0.01
YogChs	0.39	0.49	0.76	-0.40	-0.12	0.25	-0.54	-0.05	1.00	-0.62	-0.77	0.01	-0.62	0.04	-0.51	0.05	-0.11	0.02	-0.55	-0.55
Chips	-0.65	-0.76	-0.69	0.19	-0.04	-0.26	0.28	0.01	-0.62	1.00	0.64	-0.02	0.52	0.28	0.52	-0.01	0.49	0.02	0.45	0.41
Soda	-0.59	-0.58	-0.83	0.33	-0.18	-0.23	0.42	0.05	-0.77	0.64	1.00	0.00	0.66	-0.15	0.65	-0.06	0.33	0.02	0.39	0.67
Fruit	-0.01	-0.00	0.03	0.02	-0.02	-0.00	-0.01	0.03	0.01	-0.02	0.00	1.00	0.00	-0.03	0.00	0.02	0.02	0.00	-0.01	0.03
Corn	-0.41	-0.45	-0.64	0.30	-0.04	-0.18	0.37	0.02	-0.62	0.52	0.66	0.00	1.00	-0.07	0.51	-0.03	0.18	0.02	0.36	0.48
Fish	-0.01	-0.22	0.05	-0.09	0.24	-0.08	-0.04	-0.04	0.04	0.28	-0.15	-0.03	-0.07	1.00	-0.10	0.04	0.10	-0.01	0.18	-0.23
Sauce	-0.57	-0.55	-0.61	0.16	-0.33	-0.16	0.19	0.03	-0.51	0.52	0.65	0.00	0.51	-0.10	1.00	-0.02	0.45	0.00	0.17	0.55
Beans	-0.00	0.00	0.03	-0.03	0.01	0.01	-0.01	-0.02	0.05	-0.01	-0.06	0.02	-0.03	0.04	-0.02	1.00	-0.02	-0.01	-0.02	-0.05
Tortya	-0.77	-0.63	-0.34	-0.19	-0.68	-0.05	-0.29	-0.03	-0.11	0.49	0.33	0.02	0.18	0.10	0.45	-0.02	1.00	0.02	-0.23	0.40
Salt	-0.02	-0.01	-0.01	0.02	-0.02	0.01	-0.01	-0.02	0.02	0.02	0.02	0.00	0.02	-0.01	0.00	-0.01	0.02	1.00	-0.00	-0.01
Scented	0.01	-0.23	-0.42	0.37	0.53	-0.25	0.54	0.04	-0.55	0.45	0.39	-0.01	0.36	0.18	0.17	-0.02	-0.23	-0.00	1.00	0.09
Salza	-0.56	-0.47	-0.61	0.17	-0.38	-0.12	0.17	0.01	-0.55	0.41	0.67	0.03	0.48	-0.23	0.55	-0.05	0.40	-0.01	0.09	1.00

2. Report:

   a. Which two attributes are most strongly cross-correlated with each other? Hint: It is the absolute value of the cross-correlation that matters, not how positive it is. [And don't forget to repeat these questions.]

   The two attributes that are most strongly cross-correlated with each other are Veggies and Soda. They have the highest absolute value of all attributes.

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

   Chips and cereal have a cross-correlation coefficient of 0.19

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

   The attribute fish is most strongly cross-correlated with is chips, since that attribute has the greatest absolute value.

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

The attribute that veggies is most strongly cross-correlated with is soda, since soda has the highest absolute value for veggies.

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

According to this data people do not usually buy both milk and cereal, since they are not strongly cross correlated with each other.

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

The two attributes that are not strongly cross-correlated with anything are salt and fruit. The maximum values of their cross-correlation coefficients are ≤ 0.3

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

Salt and fruit seem to be the most irrelevant given they are least strongly correlated with other items.

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? Can you explain this?

No, fish is not necessarily strongly cross-correlated with the third item. Since cross-correlation is not transitive, we cannot say that fish is strongly cross-correlated with the third item for certain. For instance, bread is strongly correlated with milk and milk is strongly correlated with baby food, but bread is not strongly correlated with baby food.

Part B: Agglomeration

3. Implement agglomerative clustering by yourself. Do not use a package. Cluster the guests into groups as follows:

d. At the start of agglomerative clustering, assign each record to its own cluster prototype. Suppose we have 800+ records. So, you start with 800-plus clusters and 800-plus prototypes of those clusters.

e. Use the Euclidean distance between cluster centers as the distance metric.

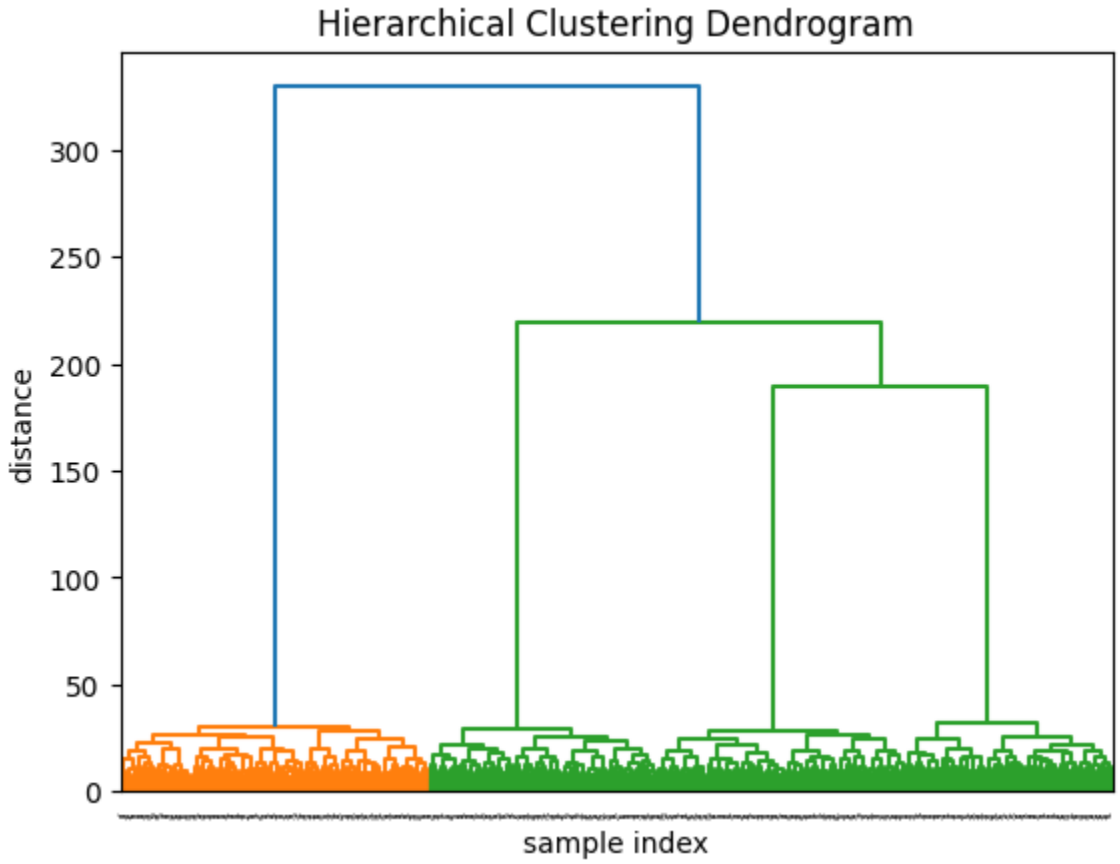
f. Use the center of mass as the prototype center, the center of mass of a set of records, to represent its center location in data space. And use the distance between these centers as the linkage method.

g. 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 10 smallest clusters merged.

[1, 1, 1, 1, 1, 1, 1, 250, 275, 375]

h. 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?



There are 4 clusters in the data, this was found by building a dendrogram based on the agglomeration. The largest gap in the dendrogram showed that there were 4 distinct clusters.

HW06
Authors: Prionti Nasir, Justin Palmer

Discussion Questions

4. Report the size of each suspected cluster, from lowest to highest.

The sizes of each suspected cluster from lowest to highest is 250, 275, 300, 375 records.

5. Report the average prototype of the clusters.

Shape: 300

Milk 5.020000 ChildBby 5.833333 Vegges 8.536667 Cereal 6.043333 Bread 0.150000 Rice 7.166667 Meat 4.460000 Eggs 4.830000 YogChs 8.046667 Chips 2.016667 Soda 0.960000 Fruit 5.456667 Corn 2.923333 Fish 2.376667 Sauce 3.020000 Beans 5.046667 Tortya 7.600000 Salt 4.676667 Scented 0.450000 Salza 3.013333

Shape: 375

Milk 1.968000 ChildBby 1.973333 Vegges 1.986667 Cereal 8.042667 Bread 2.450667 Rice 5.925333 Meat 7.952000 Eggs 5.082667 YogChs 1.002667 Chips 8.600000 Soda 7.936000 Fruit 5.384000 Corn 6.978667 Fish 2.538667 Sauce 5.957333 Beans 4.901333 Tortya 7.949333 Salt 4.698667 Scented 4.957333 Salza 6.056000

Shape: 250

Milk 4.808 ChildBby 2.956 Vegges 6.628 Cereal 7.008 Bread 6.080 Rice 6.120 Meat 6.424 Eggs 4.904 YogChs 5.192 Chips 8.016 Soda 2.004 Fruit 5.320 Corn 3.900 Fish 5.552 Sauce 3.024 Beans 5.112 Tortya 6.988 Salt 4.632 Scented 4.924 Salza 1.884

Shape: 275

Milk 9.600000 ChildBby 7.934545 Vegges 7.563636 Cereal 8.076364 Bread 8.054545 Rice 6.538182 Meat 8.109091 Eggs 5.087273 YogChs 4.821818 Chips 2.018182 Soda 1.909091 Fruit 5.338182 Corn 3.985455 Fish 2.534545 Sauce 2.072727 Beans 5.069091 Tortya 1.043636 Salt 4.643636 Scented 4.970909 Salza 2.032727

6. What typifies each of the clusters? What typical names should we give each of these 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?

The first cluster is gluten free/healthy, characterized by low bread, low soda, low chips. The second one consists of party animals, characterized by, high soda,

high salsa, high torta, high chips, low milk. The third one is pretty ambiguous but it consists of hoarders, for instance shoppers who only buy food when they cannot find free food on campus. They have high quantities of mostly everything. The fourth one is a family group characterized by high baby food, high cereal, high milk, high meat.

Part C: Conclusion

7. Write a conclusion about what you learned overall. If each of you learned different things, tell me what each of you learned.

   Justin:

   I tried doing the homework first by using arrays, but the data structure wasn't the best for agglomeration. After learning about data frames, this homework became a lot easier and the code was more readable. The use of pandas data frames and numpy arrays can really help simplify all the intermediary steps to make it easier to focus on the core logic for agglomeration.

   Prionti:

   I think by now I have a fair bit of command over dataframes so I could step in to convert our array implementation into a dataframe one. I learned to create a dendrogram using scipy and it was interesting to see the structure the dendrogram presented reflect in the prototypes of our clusters as well.