

CSCI 720: Big Data AnalyticsHomework 7**Part A: Using Cross – Correlation for Feature Rejection and Selection**Question 1

## Implementation

Here the Record ID is removed, and then the cross correlation is performed. The values are rounded at 2 decimal places.

```
C:\Users\rutvi\hw7_bda\.venv\Scripts\python.exe C:\Users\rutvi\hw7_bda\agglo.py
```

	Fiction	Sci-Fict	Baby_Toddler	...	Poetry	Romance	Horror
Fiction	1.00	0.51	-0.04	...	0.02	0.31	0.37
Sci-Fict	0.51	1.00	0.39	...	0.01	-0.04	0.23
Baby_Toddler	-0.04	0.39	1.00	...	-0.01	-0.04	0.34
Teen	0.12	-0.37	-0.64	...	0.01	0.54	0.24
Manga	0.35	0.03	-0.68	...	-0.01	-0.08	-0.30
Art&Hist	0.01	0.02	0.01	...	-0.00	0.01	-0.03
SelfImprov	-0.25	-0.64	-0.54	...	-0.02	0.40	0.08
Cooking	-0.52	-0.33	0.33	...	-0.02	-0.09	-0.00
Games	0.48	0.64	0.62	...	-0.00	0.26	0.58
Gifts	0.01	0.00	0.01	...	0.04	0.00	-0.02
Journals	0.62	0.44	-0.21	...	0.03	0.16	0.14
News	-0.27	0.07	0.18	...	-0.03	-0.66	-0.56
NonFict	-0.39	-0.01	0.17	...	-0.01	-0.82	-0.72
HairyPottery	0.23	0.27	-0.33	...	0.01	-0.49	-0.55
Mysteries	-0.42	-0.64	-0.50	...	0.01	-0.02	-0.30
Thrillers	-0.30	-0.50	-0.51	...	-0.01	-0.04	-0.31
Classics	-0.67	-0.49	-0.14	...	-0.01	-0.52	-0.64
Poetry	0.02	0.01	-0.01	...	1.00	0.00	-0.01
Romance	0.31	-0.04	-0.04	...	0.00	1.00	0.77
Horror	0.37	0.23	0.34	...	-0.01	0.77	1.00

[20 rows x 20 columns]

Question 2

**a. Which two attributes are most strongly cross-correlated with each other?**

The two attributes that are most strongly related are non fiction and romance with a cross correlation of 0.82

**b. If someone buys lots of Manga books, else are they likely to buy (or not buy, depending on what is strongest.). In other words, what is the strongest Absolute value of CC with any other category? Are they also likely to buy Horror or Gifts? Or, are they NOT likely to buy Thrillers?**

The attribute strongly related to Manga is Baby Thriller with a cross correlation magnitude of 0.68.

Manga and Horror have a -0.3 cross correlation which indicates a weak correlation and the negative sign indicates that when Manga will decrease then Horror will increase.

Manga and Gifts have a 0 magnitude which indicates there is no linear relationship between Manga and Gift books

Next is Manga and Thrillers with a CC of 0.21 which shows a weak positive correlation slight tendency for the attributes to move in the same direction, but the relationship is not strong enough to make reliable predictions based on one attribute's value alone.

Therefore, while there may be some relationship between the attributes, it's not considered a strong or robust correlation.

**c. What other category is Fiction most strongly correlated with?**

Fiction is strongly related to Classics with a cross correlation of : -0.67 which is moderately strong, meaning when one would increase other would decrease. There is a noticeable tendency for both the attributes to move in opposite directions. In other words, knowing the value of one attribute provides a reasonable indication of the value of the other attribute.

**d. What other category is Self Improvement most strongly correlated with?**

Self - Improvement is strongly related to Teen with 0.71

**e. If someone buys cookbooks, what can you tell about them?**

It has a moderately strong correlation with Journals, of -0.60, which means that if someone buys a cookbook, it is very less likely that he will buy a journal. They both are opposite due to the negative sign.

It also means that cookbooks and gift books have no relationship at all as the correlation is zero, indicating that knowing the value of the cooking attribute will provide me no information about the gift books.

**f. If someone buys lots of classic novels, what can you tell about them?**

The cc between horror and classic is -0.64 indicating that on buying a classic novel, a person would rarely buy a horror book. The magnitude says nearly strong correlation and negative sign means they are opposite, hence very less likely that both will be bought together.

The same could be said with respect to fiction and classics with a CC of -0.67

**g. If someone buys NEWS, what can you tell about them?**

The cross correlation of news and non-fiction is 0.69. It means they are strongly related hence on buying news, it is likely that nonfiction will be bought, as compared to the other books.

ArtHist and news have 0.01 cc meaning they both have no influence on each other.

#### **h. What do you know about people who buy Hairy Pottery?**

There is no relation of thriller with hp as cc of 0.0. Most of the other attributes are moderately correlated, meaning they might or might not buy.

#### **i. What are Thrillers most strongly associated with, or not associated with?**

Thrillers is most strongly related with Games out of all the attributes with a cross correlation of -0.58 which indicates that on buying thrillers, it is very less likely to buy games books.

Another most strongly associated attribute is mystery with CC of 0.54 meaning on buying thrillers more likely to buy mystery.

Also, the attribute that is not associated at all is Hairy Potter as cc is 0.

#### **j. What can we infer about people who buy Art & History books?**

It has an extremely weak correlation with Fiction, Baby Toddler, Gifts, News, NonFiction, Hairy Potter, Romance as the cross correlation is 0.01.

Also, it has no relation with Cooking, Games, Journals and Poetry.

### Question 3

The three attributes to be deleted would be:

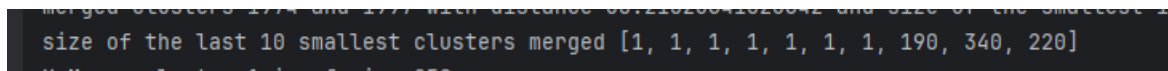
Art&Hist, Poetry and Gifts

These attributes can be deleted because the cross correlation of these attributes with all other attributes has a very less magnitude hence not giving much information. Therefore, they can be removed.

## Part B: Agglomeration

### Question 4

The size of the last 10 smallest clusters merged are [1, 1, 1, 1, 1, 1, 1, 190, 340, 220] as seen below.



```
merged clusters 1774 and 1777 with distance 0.21020041020042 and size of the smallest 1
size of the last 10 smallest clusters merged [1, 1, 1, 1, 1, 1, 1, 190, 340, 220]
K-Means cluster 4 is of size 250
```

### Question 5

Report the size of each suspected cluster, from smallest to largest size.

There are four clusters that are formed overall. This can be seen from the dendrogram that is attached below. Also, from the above image, the clusters formed are of size 190, 340 and 220. Hence the smallest size is 190 followed by the next cluster of size 220 and then followed 340. Hence the last cluster that will be formed since there are four that are formed will be 250. This cluster will be the largest among all the other clusters. For agglomerative clustering, at the end we get one cluster with all datapoints which is basically the complete dataset of size 1000.

### Question 6

Report the average prototype of each of these the clusters.

The clustering starts with each datapoint and then is combined as per the distance between them. On average, initially all the single clusters are merged majorly. With each merger, the average prototype of the newly formed cluster gives a depiction of the group's nature. We get three of the cluster sizes towards the end of the clustering. Most of the times the cluster of 1 or maybe a little more are merged to get a larger one.

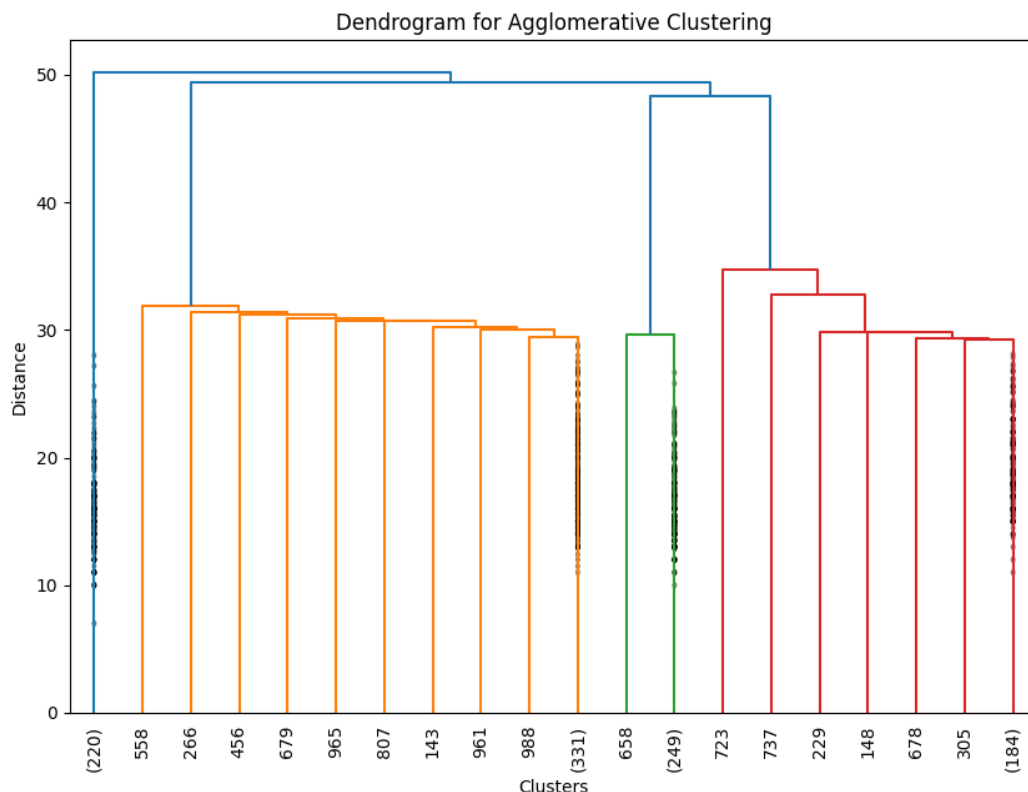
## Question 7

What typifies each of the clusters? What typical names should we give each of these prototypes? Is there a family group? Is there a gift-giving group? What typifies each group?

For one of the cluster, the genre comes to be majorly of Teen and SelfImprovement together hence that can be a group formed. Next group formed is consists of records that have a high number of purchases of Gifts hence that can be called as the gift giving group. The next cluster consists of a blend of Horror, Romance and Thriller groups hence can be classified as a group of stories.

## Question 8

### Dendrogram for Agglomeration



As seen from the above image, the number of clusters formed by agglomerative clustering is four. All clusters are shown in different colours. The y axis shows the distance and the dendrogram is created using the end 20 attributes. This is mentioned by taking the parameter within dendrogram p as 20 to plot.

## Question 9

### KMeans

On performing KMeans on the data, the result achieved is:

```
K-Means cluster 1 is of size 250
K-Means cluster 2 is of size 340
K-Means cluster 3 is of size 220
K-Means cluster 4 is of size 190

Process finished with exit code 0
```

This shows that on keeping k as 4 the cluster sizes achieved are 250, 340, 220 and 190.

On comparing this result with agglomeration, we can see that out of the four clusters formed, three of them are – 190, 340 and 220. This was recorded in getting the last 10 small size merged clusters. Hence it can be said that the fourth cluster would be 250. Hence by both agglomeration and kmeans we get a similar result.

## **Part C: Summary and Conclusion**

### Question 10

In the first part of the assignment, we studied about cross correlation and the relationship of attributes with each other. The cross correlation lies between -1 and 1. Also, an important note is to remove the Record ID while calculating the cross-correlation coefficients and just have the columns of the attributes. A value of 1 says a perfectly positive

correlation, -1 being negative and 0 means no correlation. While checking the correlation, we should check the magnitude first and then the direction matters. I learnt how different attributes have various relations with all other attributes based on the values. Also, how as per the values we can delete some of the attributes that give us no information with all other attributes.

Next part of the assignment is Agglomeration. Agglomeration is basically a type of hierarchical clustering method which follows a bottom-up approach. It takes each data point as a cluster and then at each iteration it keeps merging as per the distance between them. There are various ways of calculating the distance like Euclidean, Manhattan. For this assignment, we are following the Manhattan distance to calculate the distance. The distance determines which two clusters will be merged. Initially each data point is considered as a single cluster and then the Manhattan distance between the cluster center is taken into consideration as a distance metric. At each iteration, we merge two clusters. Also, the smaller cluster is merged to the bigger one and then the distance is recalculated between the newly formed cluster and the rest of the clusters. A dendrogram is further created to understand the structure of the data by the number of clusters formed. As mentioned in the assignment, the dendrogram formed here takes the 20 clusters.

KMeans, on the other hand follows a centroid based approach. Here, we give the number of clusters initially and the partitioning takes place over the data. After the k value is stated, we next select random points and start assigning these points to the respective clusters as per their distance. After a datapoint is assigned to a cluster, the centroids are recalculated as the means to all data points within the cluster. This is continuously done until the centroids don't move else when the clusters are the same as the previous ones.



For this assignment, the dataset has 20 attributes with 1000 records. Agglomerative clustering was performed on the dataset and at each level the merged clusters are recorded with distance between them as shown below.

```
merged clusters 435 and 960 with distance 7 and size of the smallest is 1
merged clusters 35 and 592 with distance 10 and size of the smallest is 1
merged clusters 102 and 243 with distance 10 and size of the smallest is 1
merged clusters 224 and 465 with distance 10 and size of the smallest is 1
merged clusters 437 and 775 with distance 10 and size of the smallest is 1
merged clusters 472 and 562 with distance 10 and size of the smallest is 1
merged clusters 23 and 54 with distance 11 and size of the smallest is 1
merged clusters 24 and 615 with distance 11 and size of the smallest is 1
```

The size of the last 10 smallest clusters merged is also recorded as shown in one of the above questions. It gives us an insight of the size of the clusters that are being formed. The distance is further recalculated as new clusters keep forming. At the end of agglomerative clustering, a final cluster is formed that consists of all the data points, as at the end all data points merge into one cluster. Further, KMeans is applied to get more insights of the cluster sizes formed.

Overall, this assignment shows how agglomeration works along with a dendrogram. We also learn how KMeans is applied. In addition to it, the importance of cross correlation is understood.