MDM-2021 Assignment 2

Jeyhun Yagublu (SN:1011331)

October 23, 2021

1 Task 2

In this task you will study hierarchical clustering of set type data and the effect of data order on the clustering results. Consider the following market basket data of 8 transactions (t1, . . . , t8):
t1: coffee, milk, sugar, eggs, bread
t2: bread, coffee, butter, milk, eggs
t3: sugar, cheese, cream, ham, salt
t4: eggs, cheese, apples, bread, butter
t5: apples, bread, eggs, butter, tea
t6: cheese, bread, coffee, milk, tea
t7: apples, salt, butter, ham, coffee
t8: salt, butter, bread, ham, apples

1.1 A)

Task: Calculate pairwise Jaccard distances for each pair of transactions.

Answer: I have used my own defined function in order to calculate Jaccard distance and fed it to sklearn library's pairwise_distances function:

Table 1: Pairwise distances between transactions

	T1	T2	T3	T4	T5	T6	T7	T8
T1	0.0	0.3333	0.8889	0.75	0.75	0.5714	0.8889	0.8889
T2	0.3333	0.0	1.0	0.5714	0.5714	0.5714	0.75	0.75
T3	0.8889	1.0	0.0	0.8889	1.0	0.8889	0.75	0.75
T4	0.75	0.5714	0.8889	0.0	0.3333	0.75	0.75	0.5714
T5	0.75	0.5714	1.0	0.3333	0.0	0.75	0.75	0.5714
T6	0.5714	0.5714	0.8889	0.75	0.75	0.0	0.8889	0.8889
T7	0.8889	0.75	0.75	0.75	0.75	0.8889	0.0	0.3333
T8	0.8889	0.75	0.75	0.5714	0.5714	0.8889	0.3333	0.0

1.2 B)

Task: Simulate the agglomerative hierarchical clustering algorithm with the complete linkage metric until transactions are divided into two clusters. The distance function is Jaccard distance. Show that it is possible to yield two different clusterings (into two clusters) depending on the data order. You can present the simulation by updating the distance matrix or, alternatively, draw the corresponding dendrogram, if you provide the required inter-cluster distances. Explain the steps (why certain clusters are merged).

Answer: I have used sklearn library's Agglomerative Clustering :

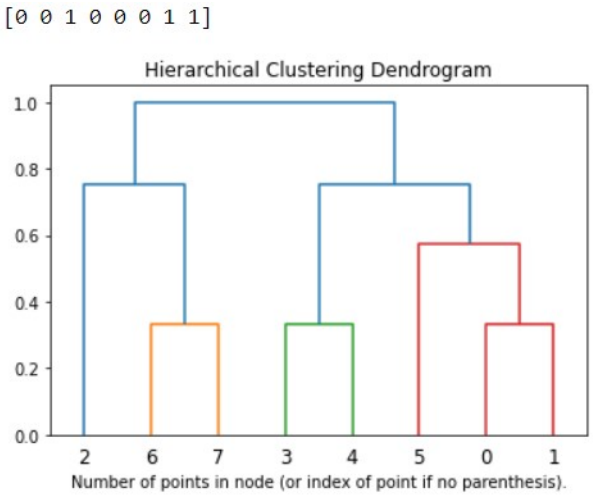

Figure 1: Agglomerative clustering simulation and clusters of transactions

if we order data as T2,T5,T6,T7,T8,T4,T3,T1 we will get that T3 will

be singleton cluster and all the others will be in one cluster. As we know Agglomerative Hierarchical clustering with complete link method depends on the order of data because when he calculates the next merge of clusters it calculates the minimum distances and this depends on initial positions of neighboring datas:

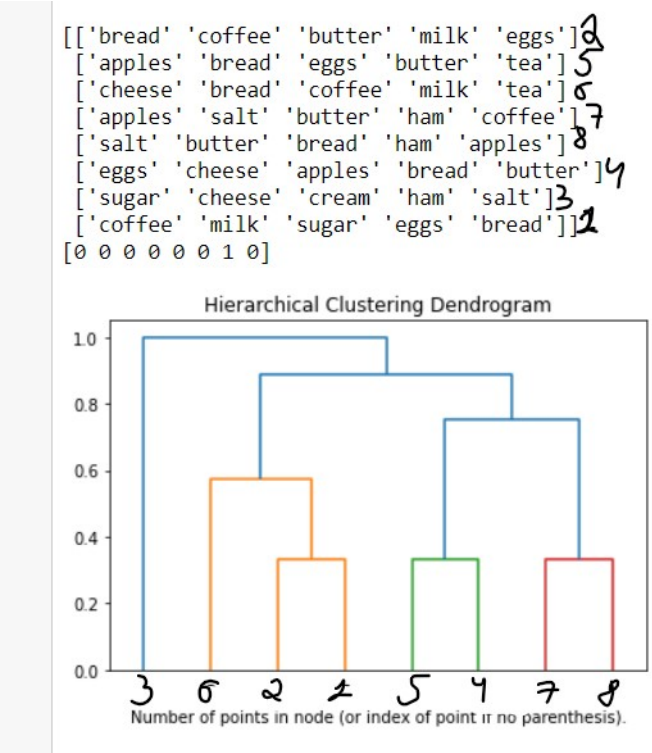

Figure 2:

1.3 C)

Task: Repeat part b) with the single linkage metric. Are the results dependent on the data order?

Answer: when we use single link method:

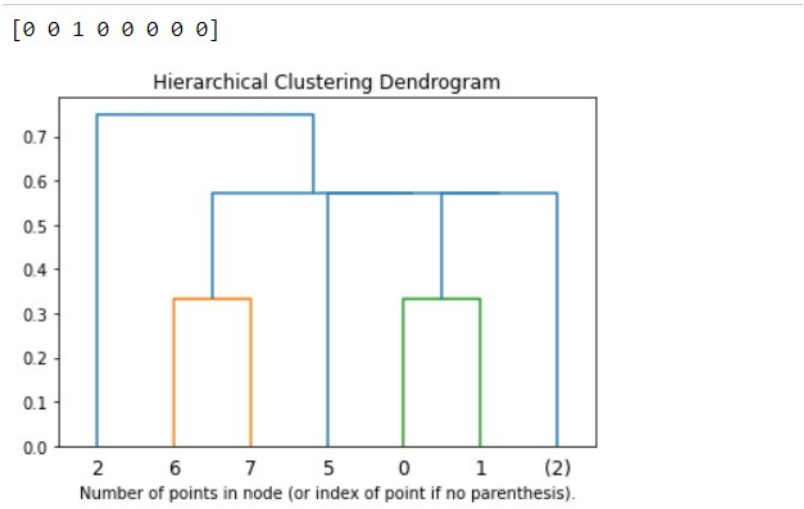

Figure 3:

And no matter how much we shuffle the data it always gives the same clustering since it is not sensitive to the order of data.