MDM-2021 Assignment 2

Jeyhun Yagublu (SN:1011331)

October 23, 2021

1 Task 1

This tasks continues Exercise session 2, clustering the rat data. Use the same data set ratdataNormChecked.csv. Use only features liverind, heartind, appind, batind, tailind, ADWBind, gonind, BMI. The distance function is Euclidean

1.1 A)

Task: Cluster the data with agglomerative hierarchical clustering and compare four linkage metrics:
– single link
– complete link
– average link
– Ward's method
Test values of K = 2, . . . , 8 and choose the best clustering with Silhouette index (SI). Report the best clustering with each linkage metric (its K and SI value).

Answer: I have calculated the Agglomerative clustering using sklearn.cluster library and calculated SI index using sklearn metrics library.

- Single Link: Best one is K=2 when SI=0.6572
- Complete Link: Best one is K=2 when SI=0.59692
- Average Link: Best one is K=2 when SI=0.59692 which was same as Complete Link.
- Ward's Link: Best one is K=3 when SI=0.545776

1.2 B)

Task:Perform PCA and present the data using only the first principal component. Repeat the same tests as in a) and report the best clusterings.

Answer: I have performed the PCA using sklearn library. Resulting first 5 indices of 1 dimensional data are:

1. -0.98718407
2. 0.6866718
3. -0.99371096
4. 1.67443154
5. -0.98495406

The SI indexes for best number of K's are :

- Single Link: Best one is K=2 when SI=0.674636
- Complete Link: Best one is K=2 when SI=0.622061
- Average Link: Best one is K=4 when SI=0.6049
- Ward's Link: Best one is K=3 when SI=0.620102

1.3 C)

Task:Compare briefly the best clusterings with each linkage metric (for each, choose the better from a) or b)). Did the methods find the same number of clusters? Are cluster sizes similar? If any of the methods found outlier clusters (containing only one or at most a few rats) check them and try to find reasons (like extreme values in individual features).

Answer: Before performing PCA we can see that almost all clustering methods show that K=2 is the best clustering but when we draw the histograms of clusters after single link method we see that almost all rats are always in 1 cluster and the others create singleton clusters. Maybe this is because of the outlier rats so I decided to do scatter plot after performing PCA which gives us following plot:

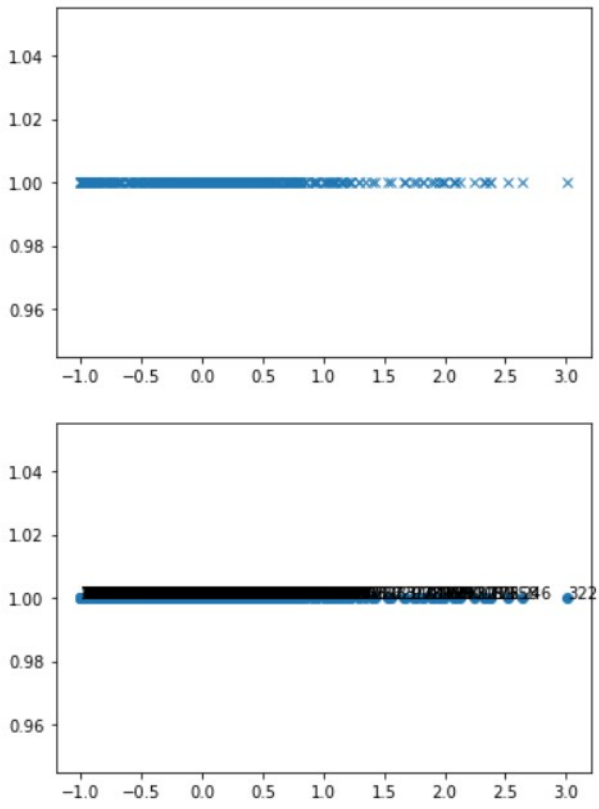

Figure 1: 1 dimensional Rat data after PCA

We can see that after performing PCA some rats have more distance from the nearest rats and rat with id 322 (rat332) is the farthest of all so when

we choose K=2 it puts all the rats to first cluster and puts rat id 322 to the second singleton cluster.
Similarly it continues to put second farthest point to the 3rd cluster when we choose K=3 and so on.
The reason that rat332 is outlier lies in its value of gonind which is the biggest of all with value 3.9453 while the mean value for that feature over all rat data is 0.954.

- Single Link: As we have discussed about outlier we can see that Single Link method doesn't perform well for this data as outlier affects too much and so we got many singleton clusters for both original data and also after performing PCA.
- Complete link: The cluster sizes for K more than 2 were different before and after performing PCA but when K was 2 clusters were same for both data and we got highest SI index score for when K was 2.
- Average Link: before PCA best K was 2 and after best K was 4. Better one is after PCA cluster because it detects the outlier and puts it in singleton cluster and then divides data in 3 clusters.
- Ward's Link: Cluster sizes in both after PCA and before PCA were similar and they didn't detect the outlier rat. In this case before PCA clustering would be favored in my opinion because it considers all the features without any loss.