Yiling Chen

Hw#3 Write-Up

Mar.24, 2021

**Question 1**

Text, letter

Description automatically generated

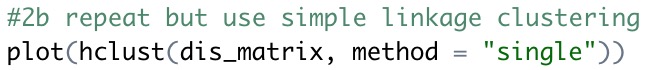
**Question 2**

The purpose of this question is to use the dissimilarity matrix of four observations to sketch a dendrogram from hierarchically clustering using complete linkage and single linkage. Determine which observations are in each cluster if cut the dendrogram obtained only two clusters using complete and single linkage. And draw a dendrogram similar to the complete linkage dendrogram but two or more of the leaves are repositioned yet the meaning of the dendrogram stays the same.

Calendar

Description automatically generated Text

Description automatically generated



My first step is to enter the dissimilarity matrix for the observations into R, then I sketch the dendrogram using complete linkage (bottom left) as well as single linkage (bottom right) hierarchically clustering.

Chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

In complete linkage, dendrogram shows us that there are two clusters where observation one and two are first clustered together and then observation three and four joined another cluster second. This means that there is a closer similarity between observation one and two as well as three and four. Base on table, we can see that the there is a dissimilarity of 0.3 compare observation one with observation two, and a dissimilarity of 0.45 compare observation three with observation four.

On the other hand, single linkage separates the observations into four clusters with observation one and two as one cluster, observation three as a single cluster, and observation four as another single cluster.

2c. In order to obtain two clusters using complete linkage, we cut the dendrogram at the value of 0.7, which leaves us two clusters (1 and 2; 3 and 4).

2d. In order to obtain two clusters using single linkage at the value of 0.42, cluster one contains (1,2,3), cluster two contains (4).



Chart, box and whisker chart

Description automatically generated

Next, the question requires us to draw a dendrogram swap the leaves in the clusters since the meaning of the dendrogram is the same, I swapped observation one with observation two, and observation three with observation four.

In conclusion, we have observed the difference in complete linkage cluster dendrogram and single linkage cluster dendrogram. With complete linkage, we have clustered the data into two clusters whereas single linkage suggests three clusters. We also notice that the dendrogram didn’t change the meaning when we swapped the leaves within the cluster.

**Question 3**

The purpose for this problem is to generate a dataset with 60 observations and 50 variables with three classes. Then, perform k-means where k=3 to see how well the clusters compare to the true labels. Finally, use silhouette and gap statistics plots to determine the optimal number of clusters.

Table

Description automatically generated with medium confidence Chart, scatter chart

Description automatically generated

My first step is to create the 60 observations with three clusters of data using rnorm() function. As we can see above from the graph, the optimal number of clusters should be two or three. Next, I use k=3 to check how well does kmeans cluster my data. Luckly, base on the table, all my data perfectly well fitted into three categories. Both the rand index and adjusted rand index shows the clusters that I obtained in K-means clustering is 100% accurate to the true labels. This is the reason because the three bundles of datas in my dataset are very different comparing with each other, therefore, it has made k means clustering clearly identified them.

A picture containing table

Description automatically generated Graphical user interface, text

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated Chart, scatter chart

Description automatically generated

Next, I use silhouette plots to determine the optimal number of clusters which shows that when k=2 the si value is the highest which is 0.58. Therefore, silhouette plot suggests to let k=2.

Chart

Description automatically generated Graphical user interface, text

Description automatically generated with medium confidence

The next step is to look at gap statistic using kmeans and find the optimal number of clusters. I set the maxium of k for the gap statistic graph as k=10, then I use the clusGap() function to generate the gap statistic. The result came back as shown below, the optimal clusters should be three.

Graphical user interface, text

Description automatically generatedChart

Description automatically generated

In conclusion, based on using means when k=3, the table accuracy categorizes the three type of clusters, but base on silhouette, the optimal number of clusters is two, and gap statistic suggest 3 clusters.