Assignment 2

Gengcong Yan - 1009903
CS-E4650 Methods of Data mining

October 22, 2021

1 Task 1

1.1 Task1.a

Compare single-link, average link, Ward's method and complete-link hierarchical clustering, when $K = [2, 8]$. Calculate Silhouette index[1] validation with different K:

Comparison of single-link, complete-link clustering methods, average link and Ward's method with $K = [2, 8]$ were implemented in program ASS2_1.ipynb (attached). The best results of each clustering methods are shown in Table 1. Single-link obtained clearly the best SI. Therefore, it looks that the single-link produced best clustering. The process code are below.

Table 1: Comparison of five clustering methods ($K = [2, 8]$) in Task 1a. SI=Silhouette index.

method	Best K	SI
single-link	2	0.657
complete-link	2	0.597
average-link	2	0.597
Ward	3	0.546

```
#linkage{'ward', 'complete', 'average', 'single'}
link="single"

for numc in range(2,9):
    cluster = AgglomerativeClustering(n_clusters=numc, affinity
       ='euclidean', linkage=link)
    predicted_labels=cluster.fit_predict(ratdata_ind)
    # print("Predicted_labels:\n",predicted_labels)
    silhouette_avg = silhouette_score(ratdata_ind,
       predicted_labels)
    print("For n_clusters ={}, Linkage method={} The average
       silhouette_score is :{}".format(numc,link,silhouette_avg
       ))
```

1.2 Task1.b

We perform PCA on original data to present new data suing only the first principal component. The best results of each clustering methods using new data are shown in Table 2. Single-link with $K = 3$ obtained clearly the best SI. Therefore, it looks that the single-link produced best clustering even after PCA.

Table 2: Comparison of five clustering methods ($K = [2, 8]$) in Task 1a. SI=Silhouette index.

method	Best K	SI
single-link	3	0.675
complete-link	2	0.622
average-link	3	0.605
Ward	3	0.620

1.3 Task1.c

From Task above, we can know the best clustering in original data and data after PCA are using single-link metric with $K = 2$ and $K = 3$, respectively. The methods didn't find the same clusters. And the clusters size are not similar, because some samples' feature number in data are apparently wrong. These rat data are divided into one cluster. we know that such as the feature *gonind* of rats 246, 258, 322 are bigger than 1. This feature can not greater than 1 because it is a proportional ratio to weight. From this task, we can know that classification is used not only to distinguish different kinds of data, but also to filter out the outliers in the data at certain times.