CS-E4650 - Assignment 2

Arttu Häkkinen - 596077

October 23, 2021

Task 1

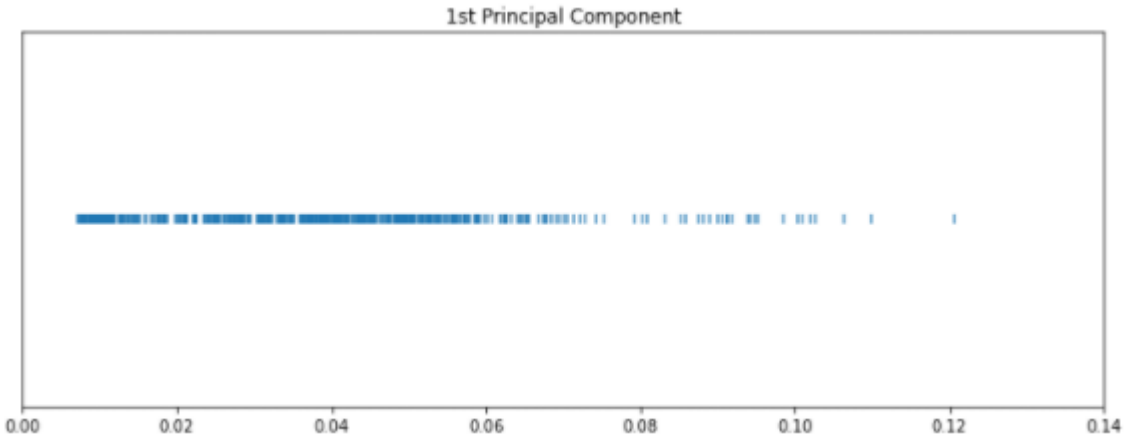
a)

First, I downloaded the data any selected the necessary features. Then I programmed a loop that clusters the data [1] using all K values, computes the SI value [2], compares it to the best SI value for that linkage metric so far, and stores the best SI value and K value if the previous best gets topped. Best clusterings for each linkage metrics:

- single link: $K = 2, SI = 0.657$
- complete link: $K = 2, SI = 0.597$
- average link: $K = 2, SI = 0.597$
- Ward's method: $K = 3, SI = 0.546$

b)

The first principal component after performing PCA for the data:



Then I repeated the same tests as I did in the part a, but now using the computed first principal component [3] as data. The best clusterings for each linkage metrics:

- single link: $K = 2, SI = 0.679$
- complete link: $K = 2, SI = 0.626$
- average link: $K = 2, SI = 0.626$
- Ward's method: $K = 4, SI = 0.594$

c)

For all linkage metrics I would choose the clusterings with the first PCA component (b), since they all have better Silhoutte Index compared to clusterings with original data (a). Single link, complete link and average link found the same amount of clusters (2) in both a and b. Ward's method on the other hand, found 3 clusters in a and 4 clusters in b. The single link method found a outlier cluster containing only 1 rat in both a and b. As we can see from the first principal component plot above, there is clearly one outlier rat parallel to the direction of the highest variance in the data. This is enough evidence to state that there has to be one rat which has extreme values in its individual features.

References

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html`

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html`

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`

[4] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`

[5] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html`

[6] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html`

[7] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html`