IHB0090
Kmeans in Matlab
Assignment

Moritz Hangen

2024/04/24

Contents

1 Task

- Taking into account all you have learned about clustering, perform the k-means clustering and evaluate the results.
- Figure out the number of clusters and the selection of parameters most suitable for the assignment, justify your decision. Cluster the subjects as well as you can.
- Give a short overview of your used methodology and the results, describe the composed clusters considering their parameter values.
- Were there some parameters, which did not improve the clustering and were rejected? Can you assume, what were the reason for that?

2 Results

2.1 Number of Clusters

Trying out all possibilities from 2 clusters to 30, it has shown that 2 is the optimal number of clusters, leading to the highest mean Silhouette value.

2.2 Selection of Parameters

After establishing two as the optimal numbers of clusters for all parameters, it has been tried to reduce the number of parameters to check for possible improvements in the Silhouette score.

Trying out to remove one parameter, it can be observed that the removal of the age improves the Silhouette score the most from 0.4851 to 0.5763. It can be further increased by removing the BMI as well, leading to a maximum mean Silhouette score of 0.6173.

To summarize, the maximum Silhouette score can be achieved by removing BMI and age which has been shown experimentally using trial and error.

A possible reason for this might be that the BMI and age are not clusterable as they could be just noise in the data and appear too "random", thus distracting the creation of clusters. Also, there might be correlation between BMI / age and other parameters such as physical health, thus making its inclusion redundant.