Homework on clustering

Author: Klára Martinásková

1. Intro

The goal of this task was to group subjects into clusters so that the clinic could later perform further EEG signal processing on these clusters. The clinic will look for objective symptoms that characterize the brain state of the subjects and can later be used to assess brain states without having to fill out lengthy subjective questionnaires. The questionnaire is likely to be used to derive parameters of physical health and psychological health.

I created the program in MATLAB 2020a. I comment the script called *"Clustering_task"* to show how I proceeded.

2. Selecting parameters

After loading the data, I selected the relevant parameters. For successful parameter selection, I followed the ultimate goal of the clinic, which is to analyze the EEG recordings to characterize the brain state. Long questionnaires were used to divide patients into groups, so I tried to use both parameters based on the questionnaire. I decided to exclude the BMI score, which is unrelated to brain activity. Similarly, I decided to exclude age, as the subjects' brain state is unlikely to be affected by age in working subjects. I also verified this fact using a correlation matrix from standardized data. This table shows that neither age nor BMI is related to physical health, psychological health, or index of depression.

Correlation matrix:

	Age	BMI	Physical health	Psychological health	Index of depression
Age	1,000	0,299	0,031	0,036	0,082
BMI	0,299	1,000	-0,130	-0,080	0,041
Physical health	0,031	-0,130	1,000	0,717	-0,667
Psychological health	0,036	-0,080	0,717	1,000	-0,625
Index of depression	0,082	0,041	-0,667	-0,625	1,000

This left three attributes: physical health (based on the Occupational Stress Questionnaire), psychological health (based on the Occupational Stress Questionnaire), and index of depression.

3. K-means

I standardized the values using z-scores (a function provided by MATLAB).

Subsequently, I used the k-means function for clustering. After many previous experiments with other cluster counts, I set the number of clusters to 3. I tried changing the settings for the initial selection of centroids and the metrics for measuring the distance from the centroids, but this had no effect on the results. I am using Squared Euclidean distance as the distance metric. I use the *"plus"* parameter

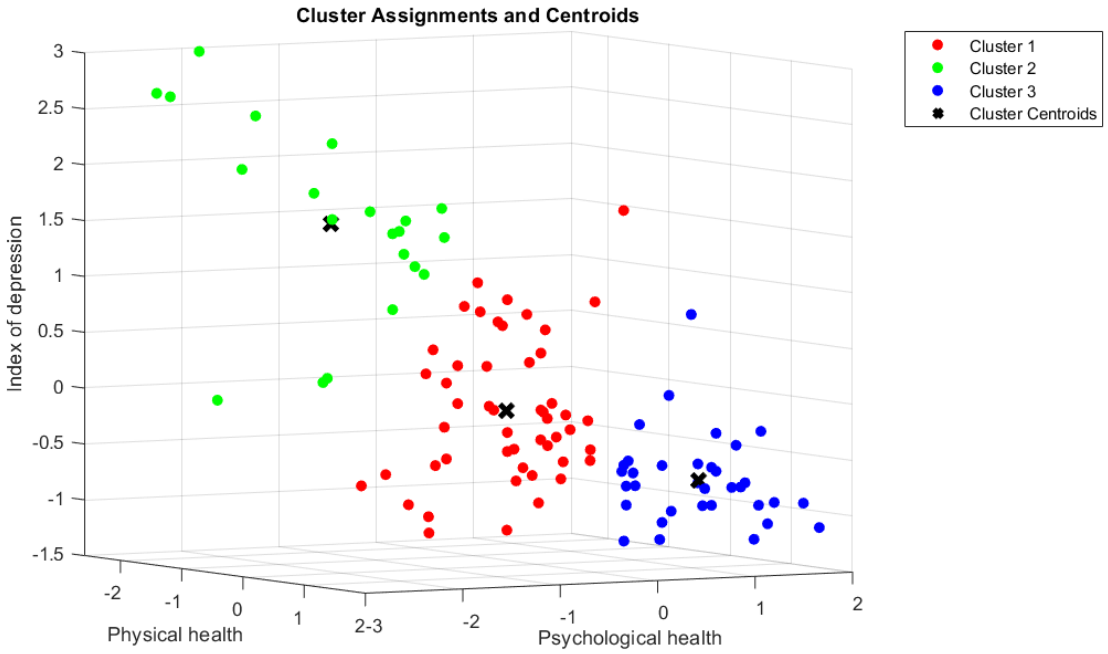
to set the initialization centroids - to "select k seeds by implementing the k-means++ algorithm for cluster center initialization" [1].

Next, the silhouette value was calculated. The silhouette value is a measure of how similar a data point within a cluster is to other clusters. I also use the Squared Euclidean distance metric for the calculation.

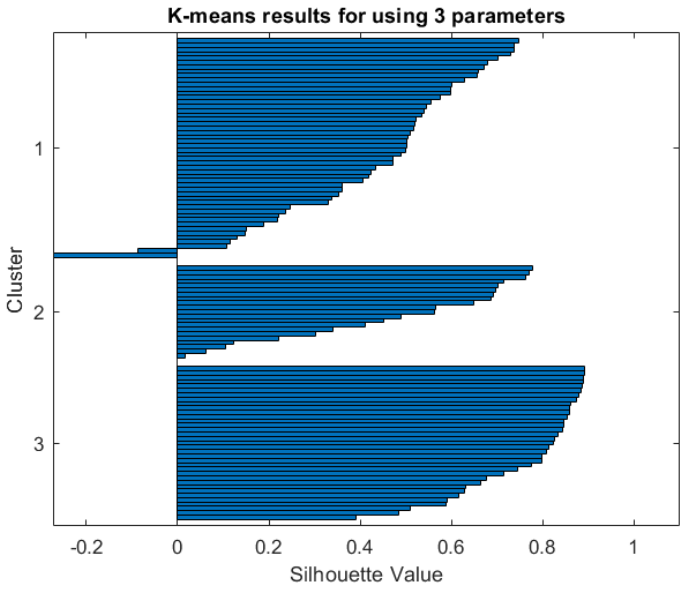
3.1. Three parameters

In the *"Clustering_task"* script, I first perform k-means for the three selected parameters. I use the *scatter3* function for rendering [2]. The silhouette score is affected by the dimensionality of the data, with higher-dimensional data being more challenging to cluster effectively. But the plot demonstrates that the algorithm can split the points well.

Clustering result:



Silhouette:



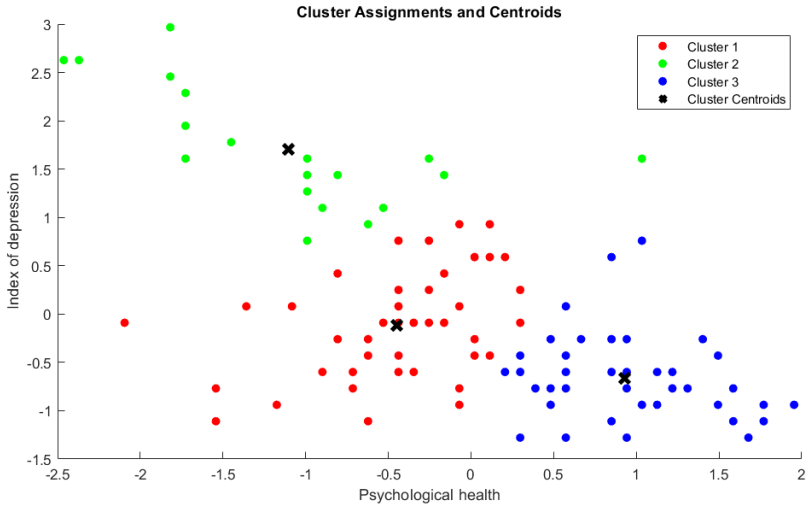
3.2. Two parameters

Then, in the same script (*"Clustering_task"*) I do k-means with only two parameters, namely psychological health and index of depression. This is because omitting the physical health parameter did not affect the clustering results in any significant way. I attribute this to the fact that the data sufficiently describe the two parameters used. Also, the two parameters from the questionnaire are strongly correlated with each other.

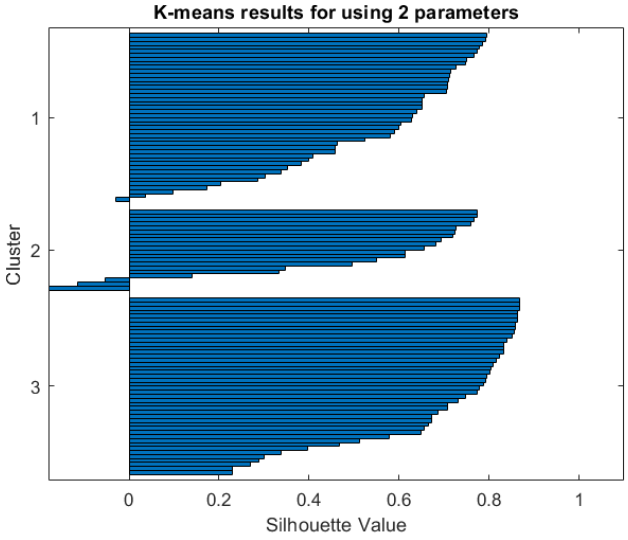
I think I can take this step because physical health is unrelated to brain states. Moreover, when we compare the silhouette values for the two methods used, it is evident that there is not much difference for the clusters.

It should be said that I also differentiate clusters by their location on the axes, the cluster number designation depends on the k-means progression, but is largely the same for both methods.

Clustering result:



Silhouette:



The final classification of patients into clusters is stored in the variable *final_results*.

[1] https://www.mathworks.com/help/stats/kmeans.html#namevaluepairarguments

[2] https://www.mathworks.com/help/matlab/ref/scatter3.html