Problem Solving Environments and Applications in Data Science ECE525

Report

Saturday 14 January 2023

Fall Semester

Lab 7
Clustering Exercise

Christos Arseniou 2730
Anastasia Psarou 2860

Supervisor: Elias Houstis

1 Clustering

Clustering is the process of dividing the data points into some groups so that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. Clustering is an unsupervised method, as we draw references from datasets consisting of input data without labeled responses.

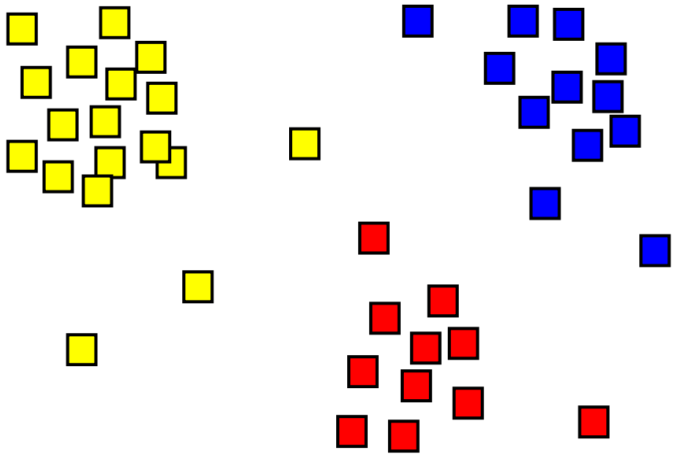

Figure 1: Clustering

1.1 K-means

K-means is a clustering method that segments data into k clusters in which each observation belongs to the cluster with the nearest mean. At the beginning of this method, the number K of clusters to be used is initialized and the K centroid points are randomly assigned. Afterwards, its data point is assigned to to its nearest centroid in order to create K clusters. We used k=2, based on the results of the Elbow method.

1.1.1 Elbow Method

The elbow method is a graphical representation of finding the optimal 'K'. It is implemented by finding the sum of the square distance between points in a cluster and the cluster centroid. Hence, the optimal K value is being picked from the graph, when there is an elbow shape in it. In this case, it looks from the graph that K=2 is the optimal value, and

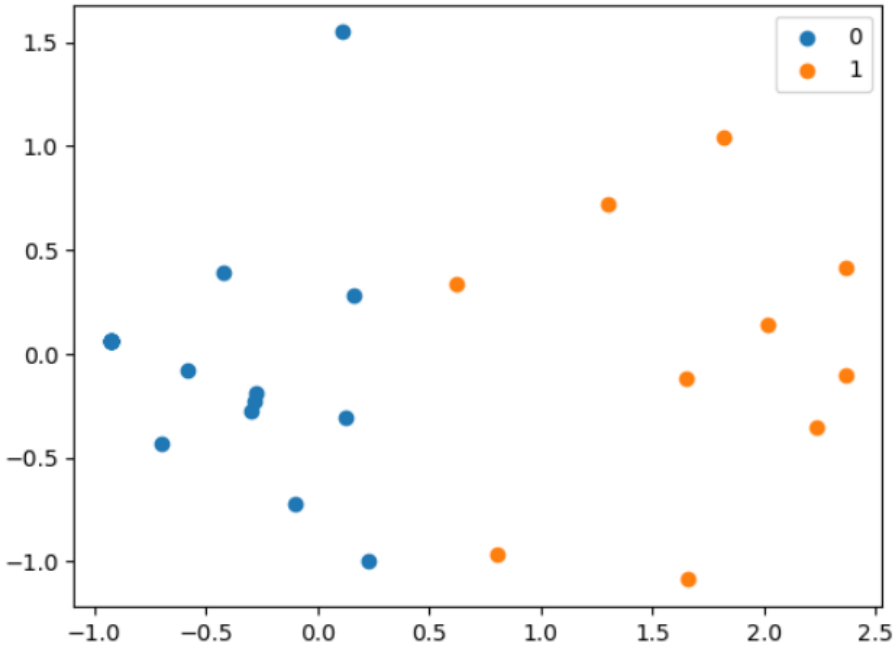

Figure 2: K-means, k=2

this is the number of clusters being used for the implementation of the k-means function. The basic problem with this method is that usually the elbow point is not always distinct from the graph, but this is not the case here.

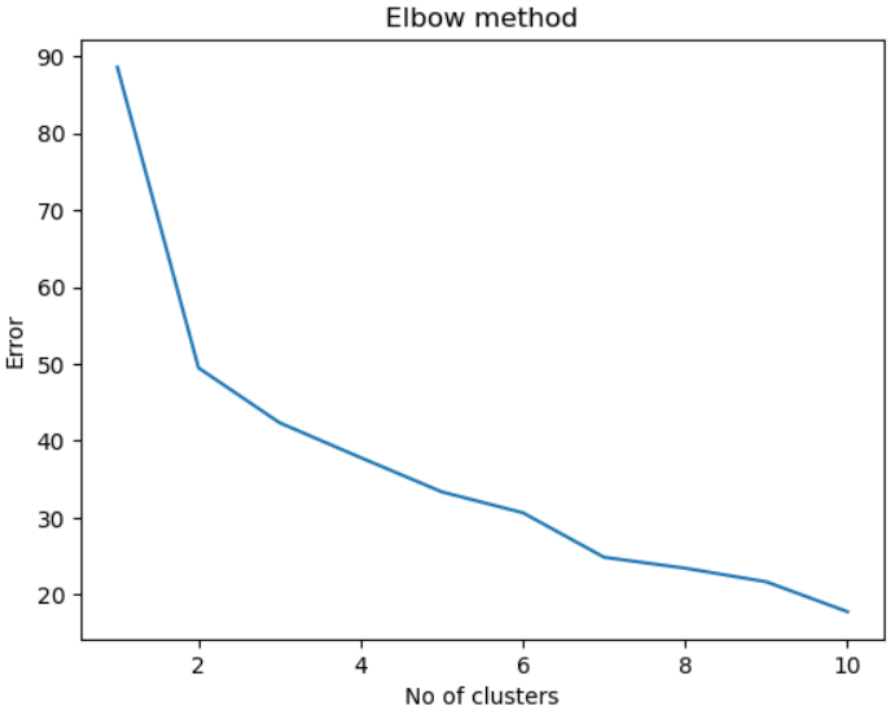

Figure 3: Elbow Method

1.2 Hierarchical Clustering

Hierarchical is also a clustering method. It begins by treating each observation as a separate cluster. Afterwards, the repeatedly execution of the following two steps takes place. Firstly, the two clusters that are closest are identified together, and then the two most similar clusters are being merge. This iterative process continues until all the clusters are merged together. The main output of Hierarchical Clustering is a dendrogram, which shows the hierarchical relationship between the clusters:

Figure 4: Hierarchical Clustering Dendogram

We can assume from the dendrogram that also after the implementation of the hierarchical clustering method we end up with 2 groups of clusters. In the below graph the results of the hierarchical clustering are being presented in a scatter plot. We notice that the scatter plot is really similar to the one produced from the K-means algorithm.

1.3 Conclusion

The basic difference between K-means and hierarchical clustering is that in Kmeans clustering, the number of clusters is pre-defined and is denoted by "K", but in hierarchical clustering, the algorithm can stop at any number of clusters, one find appropriate by interpreting the dendrogram. Even the differences of the implementations of the two algorithms we end

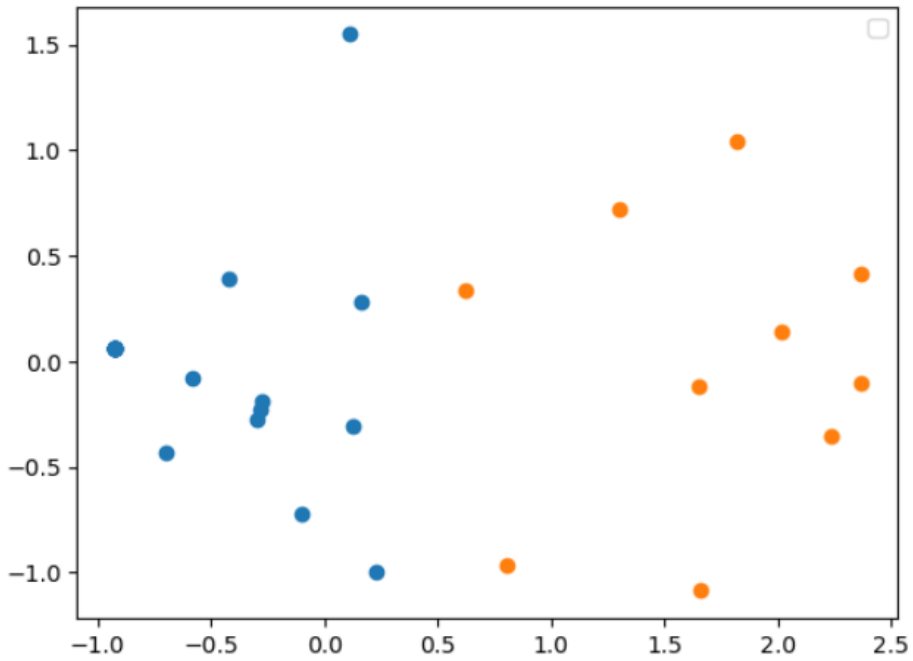

Figure 5: Hierarchical Clustering Scatter Plot

up to similar results.