Data Analysis

Instructor: prof. Panagiotidou Sofia

4th assignment

Aristotle University of Thessaloniki
Department of Mechanical Engineering

Dimitris Nentidis

Contents

1 Introduction

Clustering of data differs quite a bit from classification, which was presented in the previous two assignments. While the goal before was to classify the observations into sets for the purpose of predicting the response, here the grouping of observations is done based on how similar they are to each other. The result is the creation of some clusters with their centers and the calculation of the distances of the data points from them.

In this analysis, the dependence of life expectancy on population vaccination against specific diseases is examined — Hepatitis B, Polio, and Diphtheria.

Initially, the data are grouped into 3 and 4 clusters, and the centroids of each group are presented, as well as the number of observations that make up each group. Then, using the elbow method, an attempt is made to find the optimal number of clusters.

In the second part, hierarchical clustering of the data is performed and the corresponding dendrograms are produced using complete linkage and single linkage.

2 K-means clustering

The K-means clustering algorithm is an unsupervised learning method for grouping data into K clusters. It initializes K centroids and then, through an iterative process, assigns data points to the closest centroid. It then updates the centroids by computing the distance of each point from its current centroid, continuing until the centroids stabilize.

2.1 K = 3

For three clusters, the results are shown below.

Cluster	Hepatitis B	Polio	Diphtheria
1	0.51317155	0.44674693	0.44303206
2	-1.2571265	-2.15908747	-2.09250953
3	-1.01764069	-0.27598103	-0.3015421

Table 1: Cluster centroids, k=3.

Cluster	Observations
1	999
2	295
3	169

Table 2: Number of observations per cluster, k=3.

2.2 K = 4

For four clusters, the centroids and number of observations per cluster are shown in the following tables.

Cluster	Hepatitis B	Polio	Diphtheria
1	0.50997413	0.45285415	0.4448392
2	-1.69078705	-1.63615923	-3.1096619
3	-2.25726238	0.18956665	0.17588103
4	-0.39306467	-1.14047574	-0.49559447

Table 3: Cluster centroids, k=4.

It is noteworthy that in both cases, for k=3 and k=4, the first cluster, which also contains the most observations, retains nearly the same centroid.

Cluster	Observations
1	997
2	259
3	107
4	100

Table 4: Number of observations per cluster, k=4.

3 Optimal K

The optimal number of clusters is determined by calculating the sum of the sum of distances of all points in each cluster from its centroid — the within-cluster sum of squares or inertia.

For a cluster C_i with centroid μ_i, the inertia is calculated as:

$$\text{Inertia}_i = \sum_{x \in C_i} \|x - \mu_i\|^2$$

Where: - x is a data point in cluster C_i - μ_i is the centroid of cluster C_i - $\| \cdot \|$ denotes the Euclidean distance

The total inertia for the entire clustering is the sum of the inertia values for all clusters:

$$\text{Total Inertia} = \sum_{i=1}^{K} \text{Inertia}_i = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2$$

Inertia is an important metric for evaluating clustering quality as it gives a quantitative estimate of how compact the clusters are. A small value indicates that the points are fairly close to each other within each cluster, meaning the clusters are well-formed. A large value indicates the points are quite spread out.

For the application of the elbow method, inertia is computed for various values of K. Then the "elbow" of the graph is located — the point at which the "rate of improvement" (not actually a continuous rate) decreases. According to the elbow method, the optimal number of clusters here is 6.

As before, the first cluster, which contains the most observations, has almost the same centroid as previously.

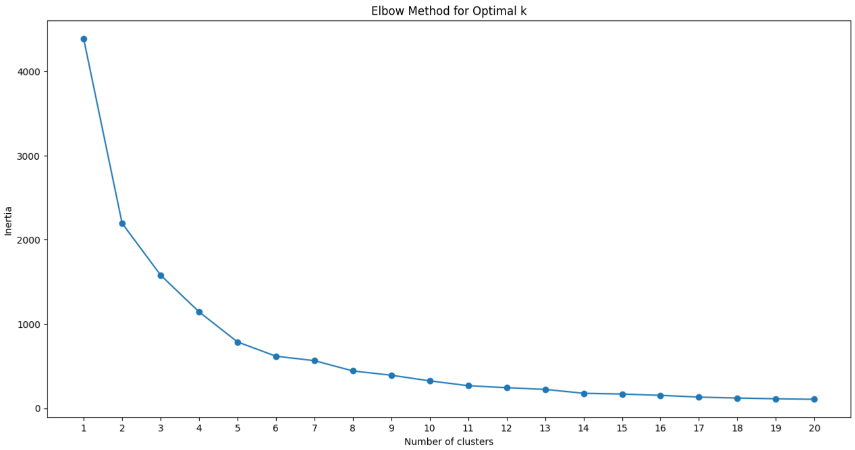

Figure 1: Inertia by number of K.

Cluster	Hepatitis B	Polio	Diphtheria
1	0.5525116	0.49859809	0.49356559
2	-1.47496334	-0.72489302	-2.91141582
3	-2.38359224	0.18094352	0.17547732
4	-0.25931646	-0.36351979	-0.38278384
5	-0.43179295	-3.31526521	-0.21428379
6	-2.02946244	-3.23728624	-3.3798574

Table 5: Cluster centroids, k=6.

Cluster	Observations
1	908
2	298
3	91
4	72
5	56
6	38

Table 6: Number of observations per cluster, k=6.

4 Hierarchical clustering

Whereas earlier we had a predefined number of clusters for grouping the data, now the grouping appears in a dendrogram. At the very bottom, all n observations are shown, and each branch that forms is a grouping of some observations.

4.1 Complete linkage

The Complete linkage method calculates all pairwise dissimilarities between observations in cluster A and cluster B and records the largest of these.

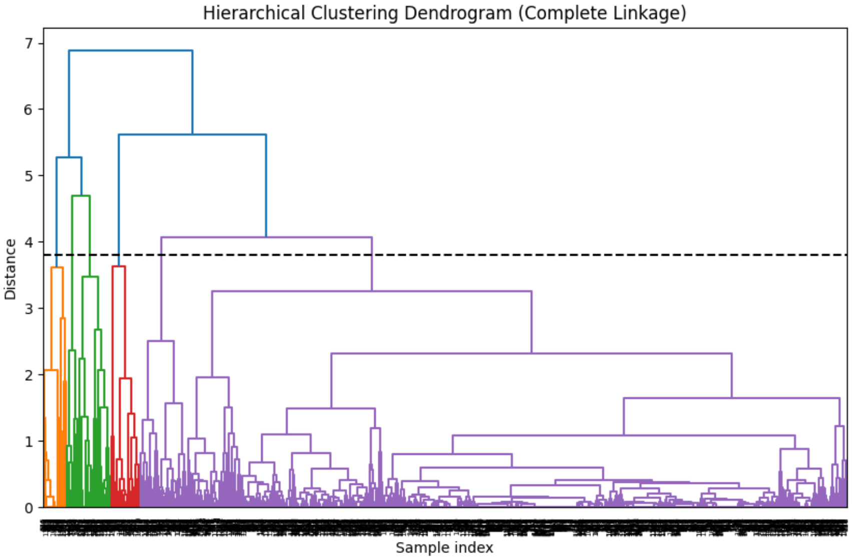

Figure 2: Complete Linkage.

As can be seen from the dendrogram, by selecting a distance of 3.8, a total of 6 clusters are formed — the same number as in the previous K-clustering determined by the elbow method. We could also select a distance roughly between 2 and 4 and still have enough space to avoid intersecting with any branching point.

4.2 Single linkage

By contrast, Single linkage uses the smallest dissimilarity. The downside of this method can be seen in the dendrogram below, with observations joining one by one each time.

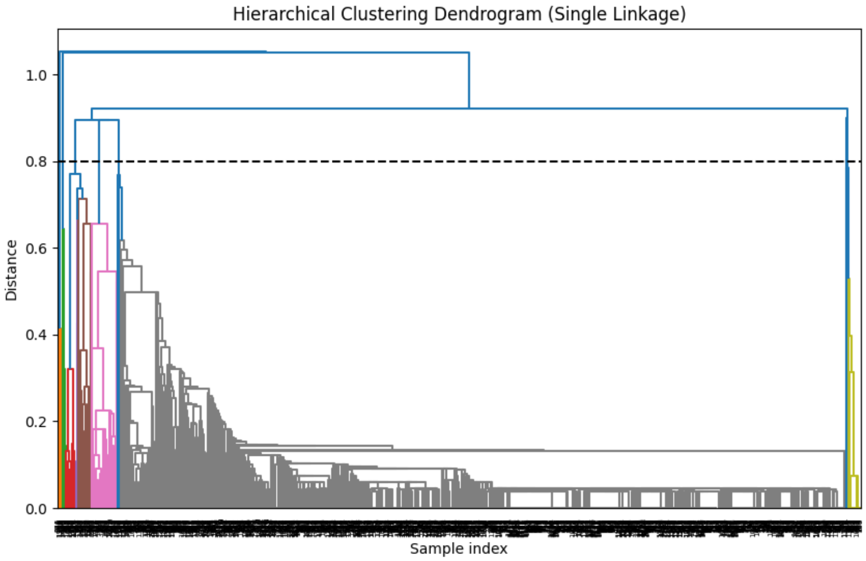

Figure 3: Single Linkage.

Similarly here, a distance of 0.8 is selected, with the data being divided into 7 clusters.