2023-09-21
Gustav Nolgren, Jonas Teglund, Oscar Wadestig
Group 43

Assignment 3 Submission Sheet

TASK 1: k-Means

Corresponds (more or less) to the three expected species? **No**

Number of records in each cluster:

1. **193**
2. **106**
3. **1**

TASK 2: Pre-processing

Is it better to rescale before or after detecting and filtering out the outliers?
From my testing, it was better to rescale before removing outliers for clustering.

Corresponds (more or less) to the three expected species? **Yes**

Number of records in each cluster:

1. **90**
2. **100**
3. **93**

PL	PW	SL	SW
0.04042605	0.30598291	0.21940837	0.50044444
0.01862745	0.57692308	0.03577922	0.08
0.04980439	0.39123242	0.28892613	0.78752688

TASK 3: Choice of k

Which K corresponds to the best clustering? (Using the Davies-Boulding index). **k = 2 with score: 0.476**

TASK 4: Hierarchical clustering

Using SingleLink, how many records are included in each of the two top clusters?
Cluster 1: **183**
Cluster 2: **100**

Which approaches produce a (more or less) correct clustering corresponding to the three species, if any?

SingleLink: **Doesn't produce three clusters at all. In the third cluster, only one is present.**

CompleteLink: **Three decently sized clusters are present. The two clusters to the top right are overlapping, however.**

AverageLink: **Best of the three! The clusters aren't overlapping at all, and contain a similar amount of elements.**

TASK 5: DB-Scan

How many clusters does DB-SCAN find with eps=1, min_samples=5? **One (1) cluster**

Can you give a value for epsilon leading to two clusters (plus noise)? **eps = 0.29**

K-DISTANCES

Which K did you use? **K = 3**

According to the k-distances plot, what value(s) of epsilon would you consider as a parameter to DB-Scan and why? **0.09 because it was the value closest to the biggest 'jump' in the graph which was created.**