Dendogram Application

Team 30
Dekel Viner(s2612925) Andreea Glavan (s3083691)

October 2018

Assignment 1

The decision tree we came up with is illustrated below.
It was immediate that using size as the root decision would not satisfy the requirements, as it would not lead to a binary tree. What is more, size is not available for all the given whales(specifically for the Blue whale). Thus, we have chosen "fluke=visible" as the root decision and "has dorsal fin" as the secondary decision(for both root's children nodes). These could be interchanged and lead to the same classification. However, we chose to keep "fluke=visible" as the root because it appeared to have more emphasis in the given descriptions. What is more, this was under the assumption that whale spotters would first notice whether or not the whale's fluke is visible, as opposed to its dorsal fin.
Using these two factors, everything can be classified with the exception of Narwhal whales and Bowhead whales. To add these to the decision tree, we added a node based on size. Even though they share the same fluke and dorsal fin characteristics, knowing the vast difference in their size allowed us to correctly classify them. This does, however, increase the height of the decision tree by 1.
An alternative for using the size as the deciding factor could be asking whether or not the whale has a tusk(as Narwhal whales are the only ones who have tusks). However, the tree would look the same even if the node query were changed.

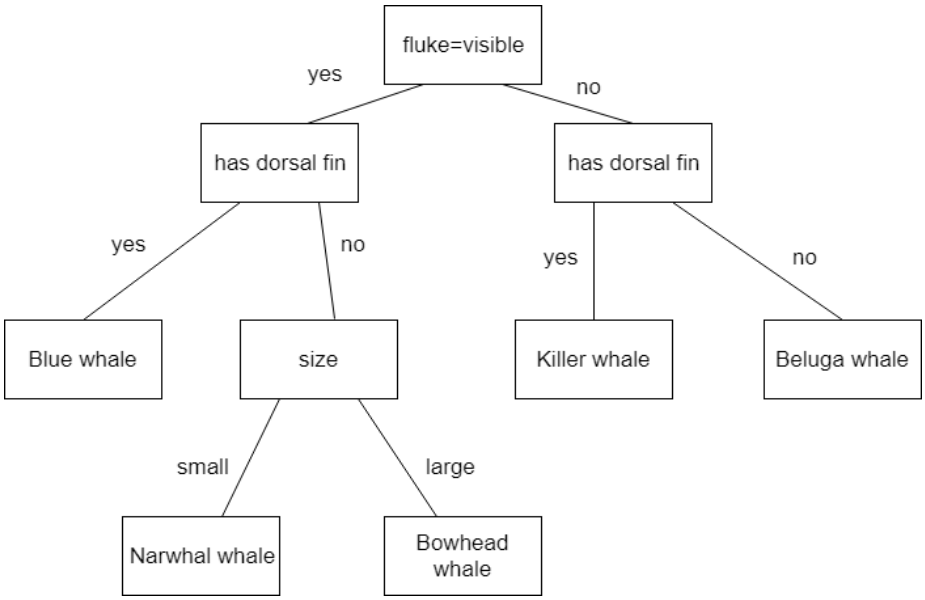

Figure 1: Binary decision tree for whale classification

Assignment 2

The plot is illustrated below, and the code used to generate it is available in the Appendix.
We used Euclidean distance as the measure for the distance between the time series, and the complete linkage algorithm which follows the formula $D(X, Y) = max(d(x, y)), x \in X, y \in Y$.
The resulting dendrogram indicates that the stock market is correlated, due to the direction of the plot and the many simiar stocks on the right hand side. However, there are also outliers mostly present on the left hand side, which have a very high dissimilarity in connection with the other stocks. It is also notable that, the closer to the right hand side, the more similar the stocks are (i.e. RDS and AFX have a very low dissimilarity). The dissimilarity becomes higher directly proportional with moving to the left hand side of the dendrogram.
It is also noticeable that there are 2 clusters(RDS, AEX, AH, HEIN, UNIL, AKZO and UNIB, ELS, WKL, DSM, WKL, DSM, COR, BOSK, PHI, FUG) which are at approximately the same value of dissimilarity. For the second one of the clusters, recursion is present in the sense that the left hand side of each 'branch' ends, while the right hand side splits further.
What is more, the dendrogram is not symmetric, having a slope lower on the right hand side and taller on the left hand side. This may be due to the stock market being correlated.

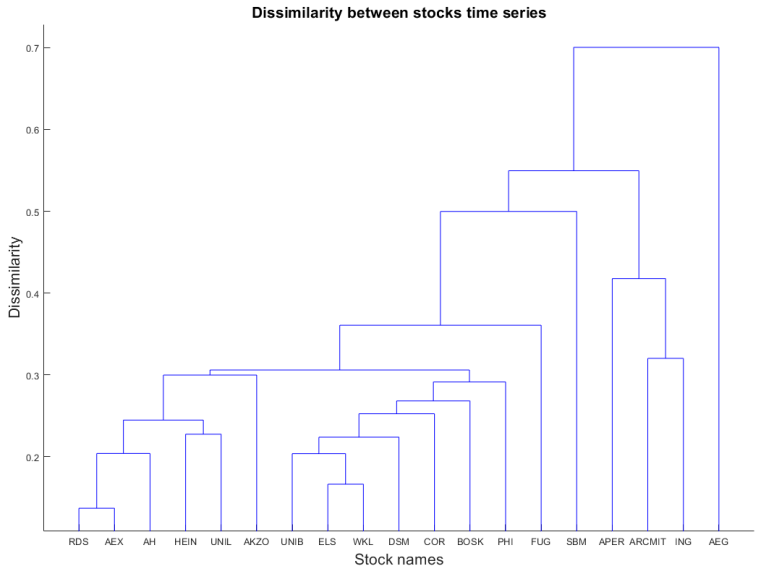

Figure 2: Dendrogram representing the dissimilarities of the different stock time series from the dataAEX data set

Appendix

Listing 1: `Source code of assignment 2`

```
%load data
load('dataAEX.mat')
load('labelsAEX.mat')

%euclidean pairwise distance computation
Y=pdist(data,'euclidean');

%apply complete linkage algorithm
Z=linkage(Y, 'complete');

%compute and plot dendrogram
figure;
dendrogram(Z,'Labels',labels);
title('Dissimilarity between stocks time series');
xlabel('Stock names');
ylabel('Dissimilarity');
```