MS-E2112 Multivariate Statistical Analysis Lietzén, Voutilainen, Mellin, Ilmonen

Exercise 10

Problem 1: Hierarchical Clustering

The data set polls.txt contains voting data of 11 different regions from the 2017 municipal elections. The variables are the percentages of votes each of Finland's eight largest political parties gained in the elections. We will use clustering methods to see which regions resemble each other most closely.

a) Scatter plot the variables. Can you spot the different clusters?

b) Calculate the euclidean distances between the countries.

c) Perform the "bottom up" hierarchical clustering by hand. Aggregate two clusters using the minimum distance (single linkage).

d) Repeat (c) using the function **hclust()**.

e) Plot the classification tree (dendrogram).

f) Repeat the steps by aggregating the clusters using the average (average linkage) and the maximum (complete linkage). Compare the results.

g) Where would you cut the tree?

Problem 2: k-means clustering

Use the data BANK.txt. The first column contains the true classification.

a) Apply the k-means algorithm to obtain 2 clusters.

b) How many observations are classified to a wrong category?

c) Change the seed number and see if it affects the results.

d) Apply the k-means algorithm to obtain 3 clusters. Does the seed number affect the results here?

Homework Assignment 10: Hierarchical Clustering

Repeat steps (a)-(b) and (d)-(g) of Problem 1 for the iris data set. The data set can be accessed from the package MASS via the command: **data(iris)**. Leave out the variable species. Remember to provide figures related to (a), (e) and (f). In addition, answer the following:

h) Does the seed number affect the results in hierarchical clustering? Justify.

i) Which aggregation metric (single, complete, average) results in the best separation of the different species? Justify your answer using the dendrogram.