MS-E2112 Multivariate Statistical Analysis Lietzén, Mellin, Ilmonen

Exercise 10

Problem 1: Hierarchical Clustering

The data set EMP.txt contains employment data in 9 different sectors in 10 different countries. The data is in percentages and it is from the year 1979.

a) Scatter plot the variables. Can you find 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? Note that there is no real answer to this.

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) Make a table of the misclassification with respect to the true classifications.

c) Change the seed number and see if it affects the results. How about the case where we want to obtain 3 clusters?

Home Exercise 10: Hierarchical Clustering

Repeat the steps from problem 1 (do not repeat (c)) to the IRIS-data. The data set can be accessed from the package MASS by writing: data(iris). Use the whole data set.