Tutorial 8: Cluster Analysis

The package **cluster** is extremely useful and contains most of the methods (for example `diana` and `pam`) discussed in the lecture. Install this package and check it out.

1. Follow the steps outlined in the lecture for clustering the European Employment data in the file `employment.dat` on the course page. Use the first nine variables. Do the clusters correspond to the four categories? Try various linkage methods and construct dendrograms.

2. Consider the data set `satimage.txt` in the course data directory. The description is given in the lecture note.

   (a) Do not use the class variable. Standardise the other variables and perform cluster analysis using different techniques from those in the example, for example SL (single linkage), AL-average linkage, CL-complete linkage. Are they better or worse than the Ward linkage? Is there a difference if K-means is used instead of partitioning around medoids?

   (b) Construct a silhouette plot for partitioning around medioids (`pam`) with values of K different from 6, for example $K = 7$.

   (c) Construct a confusion table for `pam` clustering with $K = 7$ clusters. How does it compare with $K = 6$?

   (d) Run the clustering algorithms for the `satimage.txt` data, but only using the centre pixels (i.e. the variables CC1, CC2, CC3, CC4) of each 3×3 neighbourhood. Compare your results with those obtained from the full data set.

   (e) There are several R packages tha deal with self organising maps. I draw attention to two of them; 1) **som** which can construct Self Organising Maps. and 2) the package **class**, within which there are functions `batchSOM` and `SOM`. Compare these with the **kohonen** package used in the worked example. That is, make a 6×6 hexagonal batch-SOM plot of the Landsat satellite image data. The other packages reequire specification of parameters where **kohonen** has default settings

3. The data in `primate.scapulae.txt` (and `primate.scapulae.xlsx`) contain indices and angles that are related to scapular shape (shoulder bones of primates), but not to functional meaning. There are 8 variables in the data set. The first five (AD.BD, AD.CD, EA.CD, Dx.CD, SH.ACR) are indices and the last three (EAD, β, γ) are angles. Of the 105 measurements on each variable, 16 were taken on *Hylobates* scapulae, 15 on *Pongo* scapulae, 20 on *Pan* scapulae. 14 on *Gorilla* scapulae, and 40 on *Homo* scapulae. The angle γ was not available for *Homo.*

   (a) Apply agglomerative and divisive hierarchical methods for clustering the variables using all 5 indices and the 2 angles available for all items. Construct dendrograms with single-linkage, average-linkage, complete-linkage and Ward-linkage for the methods.

When an isolated observation appears high enough up in the dendrogram, it becomes a cluster of size one and hence plays the role of an outlier. Which linkage methods give outliers?

(b) Find the five-cluster solutions for these methods. Construct confusion tables and compute the misclassification rate. Which method gives the lowest rate? Which gives the highest rate?

4. Consider the `diabetes.txt` data set. Use all the variables other than the class variable. Try to apply the E-M algorithm. Is the E-M algorithm appropriate here?