Multivariate Statistics: Exercise 2

October 23, 2014

Cluster analysis:

Load the data `olives` from the package `classifly`. The data contain measurements on fatty acids in Italian olive oils. The oils originate from tree different regions in Italy (`$Region`) which are again split into subregions (`$Area`). The remaining variables represent the fatty acids.

The data should be clustered, where the resulting clusters should ideally represent the regions or even the subregions. Use for the further analysis only the first seven fatty acids (why not *eicosenoic*?). Why do you first need to scale the data (`scale()`)?

- Apply k-means clustering (`kmeans()`) with $k = 3$. How many objects are misclassified (i.e. not correctly assigned to the 3 regions)? How sensitive is the result with respect to the random initialization within the algorithm?
- Apply the algorithm `pam()` from the package `cluster` and plot the result. How can you interpret the plot?
- Apply hierarchical cluster analysis (`hclust()`), using the methods *complete linkage*, *single linkage*, and *average linkage*. Here you cannot directly use the data matrix as an input, but you need to provide a distance matrix (`dist()`). Visualize the results with a dendrogram (`plot()` the resulting object). Select the 3-cluster solution using `cutree()` with option `k=3`. How many observations are misclassified? How sensitive is the result with respect to the distance measure used?
- In the package *mclust* you can find procedures for model-based clustering. Use the function `Mclust()`, provide the data matrix, and possibly a vector with the desired numbers of clusters, see helpfile. Compare the results to the methods from above.
- Apply fuzzy-k-means clustering with the function `fanny()` from `library(cluster)`, and with the function `cmeans()` from `library(e1071)`. In both cases, the result objects contain the list element `$cluster` with the cluster assignments, and `$membership` with the cluster memberships as proportions between 0 and 1. Use the function `ternary()` from the package `StatDA` to visualize the memberships. Which algorithm works better?

Save your (successful) R code together with short documentations and interpretations of results in a text file, named as *Familyname2.R*. Send the file as an email attachment to *P.Filzmoser@tuwien.ac.at*, at latest Tuesday (21.10).