Statistical Learning, Homework #3

Veronica Vinciotti, Marco Chierici

Released: 18/05/2021. Due: 28/05/2021

This homework deals with PCA and clustering. You should submit an RMarkdown file and a pdf file of the report. The RMarkdown file should reproduce exactly the pdf file that you will submit. The pdf file should be rendered directly from the RMarkdown (e.g. `output: pdf_document`) and not converted from any other output format.

Note that:

- your code should run without errors (except for minor adjustments such as file paths);
- you should discuss/justify each choice that you make and provide comments on the results that you obtain.

You will be working on a gene expression data set of 128 diseased patients belonging to two subtypes. The data are provided in the attached `gene_expr.tsv` file, containing expression for 12,625 genes and an additional column with patient subtypes. You will perform an unsupervised analysis, where you will assume no knowledge about these subtypes, so that you can use this information at the validation stage of the unsupervised learning techniques. To this aim:

- Load the data and perform a PCA. Produce a plot of the variance explained by the first components, and a scree plot.
- Make a few scatterplots of the first principal component score vectors. Plot the observations according to the given subtypes and assess to what extent the subtypes are similar to each other and are captured by the PCA.
- Hierarchically cluster the patients using complete linkage and Euclidean distance as dissimilarity measure. Cut the dendrogram so as to have two groups. Evaluate the goodness of the clustering by comparing the groups with the given subtypes. Provide a numerical similarity measure.
- See if you can improve the results. For example, repeat the procedure using correlation as dissimilarity measure, and with single or average linkage. Discuss the results and pick the combination of dissimilarity measure and linkage method that works best.
- Another way to possibly improve the results is through gene filtering:
  - For example, from the full dataset select the top N (e.g., $N = 200$) genes that differ the most across all samples, based on PCA; repeat the hierarchical clustering; cut the dendrogram to have two groups; evaluate the results.
  - A different approach that is popular in gene expression analysis is to keep only the most variable genes for downstream analysis. Since most of the $\approx 10K$ genes have low expression or do not vary much across the experiments, this step usually minimizes the contribution of noise. An unsupervised technique would then aim to identify what explains this variance. Start again from the full dataset and keep only genes whose standard deviation is among the top 5%; perform PCA and produce a scatterplot of the first two principal components scores, coloring observations by subtype; repeat the clustering/cutting/evaluation procedure. What do you observe and what conclusions do you make?