Statistical Learning, Homework #3

Veronica Vinciotti, Marco Chierici

Released: 18/05/2022. Due: 29/05/2022

This homework deals with Support Vector Machines.

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);
- in your report, you should: introduce the analysis, discuss/justify each choice that you make, provide comments on the results that you obtain and draw some conclusions.

Exercise

You will be working on a gene expression data set of 79 patients with leukemia belonging to two groups: patients with a chromosomal translocation ("1") and patients cytogenetically normal ("-1"). The data are provided in the attached `gene_expr.tsv` file, containing the expression for 2,000 genes and an additional column with the subtype. You will perform a supervised analysis for prediction of the two groups using support vector machines.

To this aim:

- Load the data and select a support vector machine for the task at hand. Evaluate different models and justify your final choice.
- A popular approach in gene expression analysis is to keep only the most variable genes for downstream analysis. Since most of the $2K$ genes have low expression or do not vary much across the experiments, this step usually minimizes the contribution of noise. Select then only genes whose standard deviation is among the top 5% and repeat the analyses performed in the previous task on the filtered data set.
- Draw some conclusions from the analyses that you have conducted.