Statistical Learning, Homework #3

Veronica Vinciotti, Marco Chierici

Released: 15/05/2023. Due: 28/05/2023

This homework deals with Support Vector Machines.

You should submit a PDF file of the report via Moodle, rendered directly from an RMarkdown source file.

The **maximum allowed number of pages is 12**. 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;
- Show only what is relevant to the analysis and the results that you obtain:
- Include the code to reproduce your analyses;
- Visualize/summarize the results with a selection of informative tables and figures.

Exercise

You will be working on a gene expression data set of 79 patients with leukemia belonging to two subgroups: patients with a chromosomal translocation ("1") and patients cytogenetically normal ("-1"). The data are provided in the attached `gene_expr.tsv` file, containing expression for 2,000 genes and an additional column with patient labels. You will perform a supervised analysis for prediction of the subgroups 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 both all the analyses that you have conducted.