Statistical Learning, Homework #3

Veronica Vinciotti, Marco Chierici

Released: 07/05/2025. Due: 18/05/2025

You should submit a single PDF file of the homework via Moodle, with the PDF rendered directly from a Jupyter notebook (if you choose to perform the analysis in Python) or from a RMarkdown/Quarto notebook (if you choose to use R).

You should write your report like a mini scientific paper, following the guidelines and feedback provided on the first homework. In particular, you should: introduce the analysis, discuss/justify the choices that you make, provide comments on the results that you obtain and draw some conclusions.

Please note that the **maximum allowed number of pages is 10**.

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 additional columns with patient labels and the outcome variable (`y`). 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 the analyses that you have conducted.