Statistical Learning, Homework #2

Veronica Vinciotti, Marco Chierici

Released: 16/04/2025. Due: 27/04/2025

You should submit a single PDF file of the homework via Moodle, with the PDF rendered directly from a Quarto or RMarkdown source file (see Guidelines) and not converted from any other output format.

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**.

For this homework, you will work on diabetes data to investigate the association between disease progression after 1 year from a baseline (`progr`; the higher the value, the worse the progression) and a number of clinical predictors, measured in 442 diabetic patients. In particular, the explanatory variables are: age, sex, BMI (body mass index), BP (average blood pressure, in mm Hg), TC (total cholesterol, mg/dl), LDL (low-density lipoproteins, mg/dl), HDL (high-density lipoproteins, mg/dl), TCH (ratio between total cholesterlol and HDL), TG (triglycerides level, mg/dl, log-scaled), GC (blood glucose, mg/dl).

In your report you should:

1. Fit a decision tree on the whole data and plot the results. Choose the tree complexity by cross-validation and decide whether you should prune the tree based on the results. Prune the tree if applicable and interpret the fitted model.
2. Fit a random forest, making an appropriate selection of the number of variables m to consider at each split. Interpret your selected optimal model.
3. Fit boosted regression trees making an appropriate selection of the number of boosting iterations (`n.trees`). Interpret your selected optimal model.
4. Compare the performance of the three methods (cost-complexity decision trees, random forests and boosting) using cross-validation. Make sure that the model complexity is re-optimized at each choice of the training set (either using another CV or using the OOB error).
5. Draw some general conclusions about the analysis and the different methods that you have considered.