Statistical Learning, Homework #2

Veronica Vinciotti, Marco Chierici

Released: 26/04/2023. Due: 07/05/2023

This homework deals with decision trees.

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.

For this homework, you will work on cancer data to investigate the association between the level of prostate-specific antigen (`lpsa`, in ng/ml and log scaled) and a number of clinical measures, measured in 97 men who were about to receive a radical prostatectomy. In particular, the explanatory variables are:

- `lcavol`: log(cancer volume in cm3)
- `lweight`: log(prostate weight in g)
- `age` in years
- `lbph`: log(amount of benign prostatic hyperplasia in cm2)
- `svi`: seminal vesicle invasion (1 = yes, 0 = no)
- `lcp`: log(capsular penetration in cm)
- `gleason`: Gleason score for prostate cancer (6,7,8,9)
- `pgg45`: percentage of Gleason scores 4 or 5, recorded over their visit history before their final current Gleason score

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. Consider now a random forest and let m be the number of variables to consider at each split. Set the range for m from 1 to the number of explanatory variables, say $nvar$, and define a k-fold cross-validation schema for the selection of this tuning parameter, with k of your choice. Prepare a matrix with $nvar$ rows and 2 columns and fill the first column with the average cross-validation error corresponding to each choice of m and the second column with the OOB error (from the full dataset). Are the CV and OOB error different? Do they reach the minimum at the same value of m? Interpret the optimal model (either using the CV or the OOB error).

3. Fit boosted regression trees making a selection of the number of boosting iterations (`n.trees`) by CV. 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.