Task 2

Ensemble classifiers

2024-04-25

Table of contents

Delivery deadline: May 9th 2024

Building and comparing Ensemble predictors

The goal of this task is to test your knowledge on building predictive models using ensemble based tree models, as well as to help you consolidate th whole process of building a predictive pipeline.

You will work with a real cancer dataset where researchers aimed at predicting recurrence (re-appearance of disease some time after treatment) using the concentrations of 101 distinct peptides that had been related with this process (so they were putative recurrence biomarkers). Data ara available from the `cancerData.csv`. `cancerInfo.csv` contains information on the groups defined by the recurrence and the site the samples came from.

The researchers had previously attempted to use some statistical models that did not perform very well so you are asked to try using an ensemble biomarker, the best of a random forest or a boosted classifier.

Questions

1. Do a short exploratory data analysis in order to know some characteristics of each variable
2. Separate the data into 2 sets: training set (2/3) and test set (1/3). Use this partition in the training phase (and validation phase if necessary) and the test phase of each of the sections that are presented below. Use the value `1234` as random seed to do the partition.
3. Fit each of the models described below. For each predictor do the required tuning, train the model and do a proper test-based evaluation.

- A Random Forest (RF) classifier.
  - Tune the parameters: number of trees and number of variables per node, by implementing a grid search procedure.
  - Assess the performance of RF using suitable metrics.
  - Determine which variables are the most relevant in the prediction.
- A gradient boosting classifer.
  - Using stumps as classification trees for the response variable, compute the misclassification rates of both the learning set and the test set across 2,000 iterations. Represent graphically the error as a function of the number of boosting iterations.
  - Compare the test-set misclassification rates attained by different ensemble classifiers based on trees with maximum depth: stumps, 4-node trees, 8-node trees, and 16-node trees.
  - Eventually you can try different boosting flavours such as `xgboost` or `lightgbm` (or other)

4. Compare the predictors based on the adequate metrics and propose a classifier and not more than 10 peptides as the most relevant recurrence biomarkers (the most important variables for your classifier).

Important remarks

- Answer the questions in a reasoned way, adding the necessary comments, not just only the code. Notice that some questions may seem ambiguous. They are indeed so that you have space for creativity while answering the questions.
- Provide your report in a reproducible manner
  - A readable report in pdf with explanations, results and discussions. Explain minimally what you do and use references to complement your explanations.
  - The Rmarkdown document or the Python notebook used to generate it.

- Use relative paths instead of absolute paths to read / write files, to make it easier to run the code outside of your computer.

Delivery / Deadline

Upload a zip file to Atenea before the deadline ends. This file should have no sub-folders and contain only

- the R/Rmd/python/pynb file used as template for the report,
- the output reports in pdf or html files.

Tthe reports should have the same name with the following pattern:

`Group_XX-LastName1-LastName2-LastName3.extension`