Assignment: Tree learning methods

17 de abril, 2020

Data set: Aqueous Solubility in Drug Discovery

In order to identify high-quality candidate drugs, pharmaceutical companies need to assess the absorption, distribution, metabolism, and excretion (ADME) characteristics of compounds, including biopharmaceutical properties such as aqueous solubility, permeability, metabolic stability, and in vivo pharmacokinetics. One of the most fundamental tests to perform is that of solubility of a compound in water (or a solvent mixture), which now takes place routinely prior to biological testing. In fact, "aqueoussolubility" testing now usually occurs very early within the drug discovery and development process. Moreover, the Biopharmaceutics Classification System classifies compounds based upon their solubility and other properties. Because patients tend to prefer oral medication, the commercial viability of a candidate drug would be greatly improved if the drug were soluble in water and could be delivered orally. For compounds that are not water soluble, results from experimental in vitro screening assays (which test the ability of a compound to dissolve in water) may not be reliable or reproducible and can lead to biological problems and increased drug-development costs. In recent years, the pharmaceutical industry has seen more candidate drugs that are highly insoluble, and this has become a real problem in drug development.

This example examines a data set involving 5631 compounds on which an in-house solubility screen was performed. Based upon this screen, compounds were categorized as either "insoluble" (3493 compounds) or "soluble" (2138 compounds).

Then, for each compound, 72 continuous, noisy structural variables were recorded. For proprietary reasons, the identities of the variables and compounds were not made publicly available. The data is in `soldat.csv` file and the `y` variable indicates the 'solubility'.

Questions:

1. Do a short exploratory data analysis in order to know some characteristics of each variable

   - variability,
   - percentage missing values
   - ...

   Also, you can apply a multivariate techniques such as PCA, clustering, ...

   Would you eliminate some variable before to do the classification study?

2. Separate the data into 2 balanced partitions: a training set (2,815 compounds) and a test set (2,816 compounds). Use this same 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 a pruned single tree classifier to predict the aqueous solubility. Assess the performance of the tree by using suitable metrics.

4. Fit a Random Forest (RF) classifier to predict the aqueous solubility. Tune the parameters: number of trees and number of variables in 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 solubility prediction.

5. Taking into account above metrics, compare the classifiers in 3) and 4).

6. Apply the gradient boosting algorithm with `adaboost` specification:

   6.1. Using stumps as classification trees compute the misclassification rates of both the learning set and the test set across 2,000 iterations of gbm. Represent graphically the error as a function of the number of boosting iterations.

   6.2. 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.

Important remarks

- Answer the questions in a reasoned way, adding the necessary comments, not just the code.
- A R markdown (or R latex) report as dynamic as you can. (As an example see `Tree_Classification.Rmd` file on atenea). Try to use R code into paragraph.
- You can use caret functions or not but is better use them.
- 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

A zip file including:

- the data set,
- the Rmd (or Rnw) file used as template for the report,
- the output reports in pdf and html files,

Deadline: 4th of May, 2020