NPFL054 Introduction to Machine Learning
Charles University, December 2018

Final Homework Exercise – Specific instructions related to Part 3

In Part 3 you will work with Random Forests learning algorithm. Here are more details related to your tasks 3a) and 3b).

- Mainly, you should determine a suitable number of trees in the ensemble (`ntree`), and also you should experiment with parameter `mtry`.
- Due to relatively high computational complexity cross-validation is *not obligatory*. Train on the whole D1 and test on the whole D2. Assume that D2 is representative enough.
- First, choose a moderate `ntree` (e.g. 100) and observe the $AUC_{0.1}$ value for several different values of `mtry`. The default `mtry` is 10–11, try also e.g. 7, 15, and 20. Then try to increase `ntree`. Recommended values are `ntree` = 50, 100, 300, 500. Depends on your computational power. The more numbers, the better picture you get.
- Report your results regarding `mtry`. Then draw a plot with dependency of the $AUC_{0.1}$ on `ntree`. You can draw two sets of values into one plot, for each `ntree` value one $AUC_{0.1}$ measured on D1 (training performance), and the other $AUC_{0.1}$ measured on D2 (validation).
- Finally, choose “optimal” values of `mtry` and `ntree`.