Assignment no. 5

Machine Learning:
Supervised Techniques
365.076 (1UE) WS 2017

Exercise 7 (30 points) Analyse the sequence data set from Exercise 6 (with one-hot encoded features) as well as the Wine Data Set and the Wisconsin Breast Cancer Data Set from the UCI Machine Learning Repository using support vector machines and random forests. For all data sets, use 10-fold cross validation for both methods as well as out-of-bag estimates on the entire data sets for random forests. When using random forests, start with 10 trees and increase the number to 100, 1000, and 10000. For all data sets, compare the resulting cross validation results with the out-of-bag estimates. Compute and visualize both types of variable importances. Summarize and interpret your results in a report. You should particularly address the following questions:

- How do SVMs and random forests compare to each other for these tasks in terms of hyperparameter selection, training times, and classification accuracy?
- For each data set individually, consider variable importances and judge whether they provide valuable insight.
- For each data set individually, try to decide which method you would prefer.

Hints: (1) an example how to read the data directly using R is provided in `readUCIdata.R`; (2) you can re-use your cross validation code from Exercise 1 (if your own code was not correct, you may use the sample solution) for wrapping cross validation around random forests; (3) you need not re-run the sequence classification with SVMs, you can copy your solution from Exercise 6 or compare to the sample solution.

Submission: electronically via Moodle:

`https://moodle.jku.at/jku2015/course/view.php?id=2634`

Please take the submission instructions into account! Deadline: Monday, January 15, 2018, 1:00pm.