Final Assignment: Supervised Machine Learning

Patrick J.F. Groenen & Pieter C. Schoonees

3 December 2020

The **deadline** is 23 December 2020 at 23:59. Submit your report as a single PDF file on Canvas. You can add code files, but only the PDF will be graded. Therefore make sure the PDF includes your code as well.

Instructions

Write a report in which you solve a substantive research question using a data set of your choosing. The report should be in the form of a small article (introduction with substantive research question, description of the data, methods, results, discussion and conclusions), similar to that of the group assignments. It should be at most 5 pages (excluding appendices, 12pt font, single column, 1.5 line spacing).

Please follow the instructions provided carefully:

- Formulate a substantive research question that can be answered using your data, answer that question using at least two techniques from the course (see below), and report on it in the form of a short academic article. Show that you understand how to use the techniques sensibly by giving a concise and intuitive explanation of them. Do not copy the methods section from your group assignment (plagiarism): write it in your own words.
- Search for your own data set. Do not use a data set that you have used in this course before. Give a reference for the data source. The data should be available so that the results in the report can be replicated, if required.
- You should answer the research question by using at least two techniques from the course. Choose and use at least one technique marked (a) and one technique marked (b) in the list below. The techniques covered are:

(a) regularized regression (ridge, lasso, elastic net)

(a) better subsets regression

(a) binary support vector machine

(a) kernel ridge regression

(b) classification and regression trees

(b) bagging

(b) random forests

(b) boosting

- Besides these two techniques, you have to provide and use your own implementation of one of following options.

  (a) Implement either the *Real Adaboost* or *LogitBoost*[1] algorithm listed in Friedman et al. (2000) in R. You can use either the binary or multiclass version. You can use a different implementation of boosting for the results in the main part of your report, but compare your implementation to a standard one in the appendix.[2] Present your code as well as the script that runs the analysis in an appendix, and demonstrate – using appropriate output – that your own code provides reasonable results.

- Make sure that the techniques used are tested using the same test data, so that they can be compared meaningfully. This might require you to write your own code for cross-validation instead of using built-in functions from packages, and it will certainly require you to reset the seed of the random number generator as appropriate.

References

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., and Li, Y. (2020). *xgboost: Extreme Gradient Boosting*. R package version 1.2.0.1.

[1] Note the different notation y^* and its meaning. Step 2 (b) can use any regression method, such as a regression tree. Take note of the computational recommendations starting at the bottom of page 352.

[2] The `gbm` Greenwell et al. (2020), `xgboost` Chen et al. (2020) and `mboost` Hothorn et al. (2020) packages could used as reference.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337–407.

Greenwell, B., Boehmke, B., Cunningham, J., and GBM Developers (2020). *gbm: Generalized Boosted Regression Models*. R package version 2.1.8.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2020). *mboost: Model-Based Boosting*. R package version 2.9-3.