Groupwise least angle regression

Konstanca Avramovska

University of Technology and Economics

Budapest

Prediction using a linear model that contains a large number of predictor variables is problematic. Usually, it results in having large variance and often leads to poor prediction performance. Moreover, the high number of covariates makes the interpretation of the model difficult. Hence, selecting and using only the most important covariates can be quite beneficial. LARS (Least Angle Regression) is a popular and efficient stepwise algorithm for this purpose. As opposed to the classical stepwise Forward Selection, when in each step the response is updated by subtracting the actual fitted value, the response in LARS is updated in a direction equiangular between the fitted values of the previous steps.

In some applications, it can be useful to group the predictors into predictor groups. For example, dummy variables representing a categorical variable or the present and lagged values of a time series variable naturally constitute groups. The groupwise version of the LARS algorithm (GLARS) was introduced in 2006 by Ming Yuan and Yi Lin. The main chapter of the thesis (chapter 2) presents a slightly different GLARS algorithm based on the more recent (published in 2016) paper of Andreas Alfonsa, Christophe Croux and Sarah Gelper. Instead of the individual variables, the algorithm sequences the candidate predictor groups according to their importance. The key difference compared to the original LARS is the usage of R^2 measures from short regressions (regressions that use only one of the predictor groups) instead of simple correlations.

The initial chapter contains the necessary knowledge from linear algebra and statistics. The multivariate linear regression is introduced from the point of view of linear algebra. The chapter highlights important facts in the form of propositions to facilitate the understanding of the algorithm.

LARS relates to the classical model selection method Forward Selection, so it is logical to ask whether GLARS performs better than the classical Forward Selection. This comparison supplements the performance evaluation in the paper of Andreas Alfonsa, Christophe Croux and Sarah Gelper, and is present in the third chapter The original dataset used in the mentioned paper contains outliers. In addition to comparing the algorithms on the original dataset, I also compare the algorithms on the cleaned data, as well as the cleaned and randomly sampled smaller versions of the original dataset.