

Delete or Merge Regressors algorithm

Tuesday, August 15, 2017 10:30 AM (30 minutes)

This paper addresses a problem of linear and logistic model selection in the presence of both continuous and categorical predictors. In the literature two types of algorithms dealing with this problem can be found. The first one well known group lasso (\cite{group}) selects a subset of continuous and a subset of categorical predictors. Hence, it either deletes or not an entire factor. The second one is CAS-ANOVA (\cite{cas}) which selects a subset of continuous predictors and partitions of factors. Therefore, it merges levels within factors. Both these algorithms are based on the lasso regularization.

In the article a new algorithm called DMR (Delete or Merge Regressors) is described. Like CAS-ANOVA it selects a subset of continuous predictors and partitions of factors. However, instead of using regularization, it is based on a stepwise procedure, where in each step either one continuous variable is deleted or two levels of a factor are merged. The order of accepting consecutive hypotheses is based on sorting t-statistics or linear regression and likelihood ratio test statistics for logistic regression. The final model is chosen according to information criterion. Some of the preliminary results for DMR are described in \cite{pro}.

DMR algorithm works only for data sets where $p < n$ (number of columns in the model matrix is smaller than the number of observations). In the paper a modification of DMR called DMRnet is introduced that works also for data sets where $p \gg n$. DMRnet uses regularization in the screening step and DMR after decreasing the model matrix to $p < n$.

Theoretical results are proofs that DMR for linear and logistic regression are consistent model selection methods even when p tends to infinity with n . Furthermore, upper bounds on the error of selection are given.

Practical results are based on an analysis of real data sets and simulation setups. It is shown that DMRnet chooses smaller models with not higher prediction error than the competitive methods. Furthermore, in simulations it gives most often the highest rate of true model selection.

Primary author: Dr PROCHENKA, Agnieszka (Warsaw University)

Presenter: Dr PROCHENKA, Agnieszka (Warsaw University)