Abstract

The thesis addresses a problem of linear and logistic model selection in the presence of both continuous and categorical predictors. In the literature two groups of algorithms dealing with this problem can be found. The first one contains the well known group lasso [5], which selects a subset of continuous and a subset of categorical predictors. Hence, it either deletes or not entire factors. Another algorithm in this group uses non-convex regularization MCP [2], which is a combination of l_1 (lasso) and l_0 (number of parameters) penalty. The second group contains CAS-ANOVA [1] which selects a subset of continuous predictors and partitions of factors. Hence, it merges levels within factors. A more effective implementation of CAS-ANOVA is gvcm [4]. All these algorithms are based on regularization.

In the thesis a new algorithm called DMR (Delete or Merge Regressors, [3]) is described. Like CAS-ANOVA it selects a subset of continuous predictors and partitions of factors. However, instead of regularization, it uses a greedy subset selection method. First, a nested family of models is created, which differ by one parameter, by either deleting one continuous variable or merging two levels of a factor. The order of accepting consecutive hypotheses is based on sorting likelihood ratio test statistics. Next, the final model is chosen according to information criterion.

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 thesis a modification of DMR called DMRnet is introduced that works also for data sets where $p \gg n$. DMRnet uses group lasso regularization in the screening step and DMR procedure after decreasing the model matrix to $p < n$.

Practical results are based on an analysis of six real data sets and twelve simulation setups. It is shown that DMRnet chooses smaller models with almost minimal prediction error in comparison to the competitive methods. Furthermore, in simulations DMRnet gives most often the highest rate of true model selection.

Theoretical results of the thesis are theorems that DMR algorithms for linear and logistic regression choose the true model with probability tending to one even when p tends to infinity with n. Furthermore, upper bounds on the error of selection are given.

References

[1] Bondell, Howard D., and Brian J. Reich. *Simultaneous factor selection and collapsing levels in ANOVA.* Biometrics 65.1 (2009): 169-177.

[2] Breheny, Patrick, and Jian Huang. *Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors.* Statistics and computing 25.2 (2015): 173-187.

[3] Maj-Kaska, Aleksandra, Piotr Pokarowski, and Agnieszka Prochenka. *Delete or merge regressors for linear model selection.* Electronic Journal of Statistics 9.2 (2015): 1749-1778.

[4] Oelker, Margret-Ruth, Jan Gertheiss, and Gerhard Tutz. *Regularization and model selection with categorical predictors and effect modifiers in generalized linear models.* Statistical Modelling 14.2 (2014): 157-177.

[5] Yuan, Ming, and Yi Lin. *Model selection and estimation in regression with grouped variables.* Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68.1 (2006): 49-67.