DMRnet - wybór zmiennych ciągłych i łączenie poziomów czynników w uogólnionych modelach liniowych dla danych wysokiego wymiaru

Agnieszka Prochenka i Piotr Pokarowski

[1] Instytut Podstaw Informatyki Polskiej Akademii Nauk
[2] Wydział Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego

Streszczenie

Wybór modelu regresyjnego jest często rozumiany jako wybór podzbioru ciągłych predyktorów. Gdy jednak w zbiorze danych mamy zarówno zmienne ciągłe jak również czynniki, to zmniejszenie modelu może polegać na usunięciu zmiennej ciągłej lub połączeniu poziomów czynnika. W pracy [2] zaproponowaliśmy algorytm DMR (Delete or Merge Regressors), który dzięki zastosowaniu klasteryzacji hierarchicznej i statystyk Walda ogranicza przeszukiwanie do zagnieżdzonej rodziny modeli, a następnie wybiera model na podstawie minimalizacji GIC (Generalized Information Criterion). DMR z powodzeniem konkuruje z metodami opartymi na regularyzacji lasso [1], [5] dla "klasycznych" macierzy planu ($p < n$). W referacie przedstawiony zostanie algorytm DMRnet, czyli rozszerzenie DMR dla regresji liniowej i logistycznej wysokiego wymiaru ($p >> n$). Schemat algorytmu DMRnet:

1. Dla siatki parametrów grupowego progowanego lasso $j = \lambda_1, ..., \lambda_k$:
   (a) Przesiej zbiór predyktorów ciągłych i czynników (bez łączenia poziomów) za pomocą grupowego lasso [3], [4].
   (b) Zbuduj rodzinę zagnieżdżonych modeli G_{λ_j} za pomocą klasteryzacji hierarchicznej opartej na statystykach Walda.
   (c) Wybierz model T_j w rodzinie G_{λ_j} za pomocą GIC.
2. Wybierz ostateczny model za pomocą GIC spośród $T_1, ..., T_k$.

Pokażemy wyniki eksperymentów selekcji i predykcji dla realnych i symulowanych danych wysokiego wymiaru, w których porównywaliśmy DMRnet z regularyzacjami opartymi na karach wypukłych (grupowe lasso) oraz niewypukłych (grupowe MPC - Minimax Concave Penalty - [6]).

Literatura

Bondell, H.D. and Reich, B.J. (2009). *Simultaneous factor selection and collapsing levels in ANOVA*. Biometrics, 65(1), 169-177.

Maj-Kańska, A., Pokarowski, P. and Prochenka, A. (2015). *Delete or merge regressors for linear model selection*. Electron. J. Statist., 9(2), 1749-1778.

Yuan, M. and Lin, Y. (2006). *Model selection and estimation in regression with grouped variables*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1), 49-67.

Meier, L., Van De Geer, S. and Bühlmann, P. (2008). *The group lasso for logistic regression*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(1), 53-71.

Gertheiss, J. and Tutz, G. (2010). *Sparse modeling of categorial explanatory variables*. The Annals of Applied Statistics, 2150-2180.

Breheny, P. and Huang, J. (2015). *Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors*. Statistics and Computing, 25: 173-187.