Stokell, Benjamin G.; Shah, Rajen D.; Tibshirani, Ryan J.
Modelling high-dimensional categorical data using nonconvex fusion penalties. (English)
Zbl 07555497

Summary: We propose a method for estimation in high-dimensional linear models with nominal categorical data. Our estimator, called SCOPE, fuses levels together by making their corresponding coefficients exactly equal. This is achieved using the minimax concave penalty on differences between the order statistics of the coefficients for a categorical variable, thereby clustering the coefficients. We provide an algorithm for exact and efficient computation of the global minimum of the resulting nonconvex objective in the case with a single variable with potentially many levels, and use this within a block coordinate descent procedure in the multivariate case. We show that an oracle least squares solution that exploits the unknown level fusions is a limit point of the coordinate descent with high probability, provided the true levels have a certain minimum separation; these conditions are known to be minimal in the univariate case. We demonstrate the favourable performance of SCOPE across a range of real and simulated datasets.

An R package CatReg implementing SCOPE for linear models and also a version for logistic regression is available on CRAN.

MSC:
62-XX Statistics

Keywords:
categorical data; dynamic programming; high-dimensional regression; nonconvex penalty

Full Text: DOI