RESCON 2025 CMS 25

GENERALISED LARS FRAMEWORK FOR VARIABLE SELECTION IN HIGH-DIMENSIONAL BINARY CLASSIFICATION

T. Kayathiri[1]*, M. Kayanan[2] and P. Wijekoon[3]

[1]*Postgraduate Institute of Science, University of Peradeniya, Peradeniya, Sri Lanka.*
[2]*Department of Physical Science, University of Vavuniya, Vavuniya, Sri Lanka.*
[3]*Department of Statistics and Computer Science, University of Peradeniya, Peradeniya, Sri Lanka.*
*skayathiri1994@gmail.com

High-dimensional logistic regression refers to situations where the number of predictor variables exceeds the number of observations in binary classification. This technique is particularly valuable in domains such as genomics, biomedical imaging, social sciences, ecology and finance. Despite its advantages, high-dimensional logistic regression presents several challenges, including the risk of overfitting, instability in parameter estimation, increased computational demands, multicollinearity among predictors and complexities, in selecting the most relevant variables. To address these issues, various penalised methods have been developed, including the Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net (ENet). The ENet method combines the strengths of both LASSO and the Logistic Ridge Estimator (LRE), providing a more flexible regularisation approach. In this study, a generalised version of the Least Angle Regression (GLARS) algorithm is proposed for variable selection, aiming at mitigating multicollinearity among predictor variables in high-dimensional logistic regression. This method combines the LARS algorithm and LASSO with existing estimators: Maximum Likelihood Estimator (MLE), Logistic Ridge Estimator (LRE), Logistic Liu Estimator (LLE), Modified Almost Unbiased Ridge Logistic Estimator (MAURLE), Modified Almost Unbiased Logistic Liu Estimator (MAULLE), Principal Component Logistic Estimator (PCLE), r-k class, and r-d class estimators. GLARS updates coefficients iteratively using least-angle directions derived from these biased estimators. Furthermore, the performance of each biased estimator integrated within the LARS algorithm is evaluated using log-loss on both empirical and real datasets, including applications to colon tumour and diffuse large B-cell lymphoma (DLBCL) data. Findings indicate that LARS-PCLE performes the best for the given empirical dataset, LARS-r-d for the colon data, and LARS-LLE for the DLBCL dataset, with corresponding log-loss values of 0.2188, 0.8425, and 0.2949, respectively. These results highlight that the effectiveness of biased estimators within the LARS framework varies with dataset characteristics. Future work will focus on developing an *R* package to assist in selecting the appropriate estimator and computing coefficients for various data types.

Keywords: High dimension, Least angle regression, Logistic regression, Log-loss evaluation, Penalised estimators