DSSV 2023
Data Science, Statistics & Visualisation
EC DA 2023
European Conference on Data Analysis

Robust and sparse logistic regression

Dries Cornilly[2], Lise Tubex[1], Stefan Van Aelst[2], and Tim Verdonck[1,2]

[1]University of Antwerp - imec, Department of Mathematics, Middelheimlaan 1, Antwerp 2020, Belgium
[2]KU Leuven, Department of Mathematics, Celestijnenlaan 200B, Leuven 3001, Belgium

Logistic regression is one of the most popular statistical techniques for solving (binary) classification problems in various applications (e.g. credit scoring, cancer detection, ad click predictions and churn classification). Typically, the maximum likelihood estimator is used, which is very sensitive to outlying observations. It also breaks down when the number of possible explanatory variables is too large. To mitigate these problems, we propose a robust and sparse logistic regression estimator where robustness is achieved by means of the γ-divergence [1]. An elastic net penalty [2] ensures sparsity in the regression coefficients such that the model is more stable and interpretable. We show that the influence function is bounded and demonstrate its robustness properties in simulations. To highlight the versatility of the approach, we simulate the case with more observations than explanatory variables. Moreover, we show that the proposed regression method also deals with the more difficult problem when the number of explanatory variables exceeds the number of observations. The good performance of the proposed estimator is also illustrated in an empirical application that deals with classifying the type of fuel used by cars.

Keywords: elastic net, γ-divergence, logistic regression, robustness, sparsity

[1] H. Fujisawa and S. Eguchi, “Robust parameter estimation with a small bias against heavy contamination,” *Journal of Multivariate Analysis*, vol. 99, no. 9, pp. 2053–2081, 2008.

[2] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.