Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

FEATURE SELECTION IN GENERALIZED LINEAR MODELS VIA THE LASSO: TO SCALE OR NOT TO SCALE?

Anant Mathur

School of Mathematics and Statistics
University of New South Wales
High Street
Kensington Sydney, NSW 2052, AUSTRALIA

ABSTRACT

The Lasso regression is a popular regularization method for feature selection in statistics. Prior to computing the Lasso estimator in both linear and generalized linear models, it is common to conduct a preliminary rescaling of the feature matrix to ensure that all the features are standardized. Without this standardization, it is argued, the Lasso estimate will, unfortunately, depend on the units used to measure the features. We propose a new type of iterative rescaling of the features in the context of generalized linear models. Whilst existing Lasso algorithms perform a single scaling as a preprocessing step, the proposed rescaling is applied iteratively throughout the Lasso computation until convergence. We provide numerical examples, with both real and simulated data, illustrating that the proposed iterative rescaling can significantly improve the statistical performance of the Lasso estimator without incurring any significant additional computational cost.

1 INTRODUCTION AND BACKGROUND

Recall that in a linear model, we denote the $n \times p$ regression matrix (or feature matrix) containing the p features $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ as $\mathbf{X} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p]$ and the corresponding regression response vector as $\boldsymbol{Y} \in \mathbb{R}^n$ (with $\boldsymbol{y}$ being the realization of the random vector $\boldsymbol{Y}$). We assume that $\mathbb{E}_{\mathbf{X}}[\boldsymbol{Y}] = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta}$ is a linear function of some model coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\beta_0 \in \mathbb{R}$, the last one being called the intercept and corresponding to the constant feature $\mathbf{1} \in \mathbb{R}^n$. Define the projection (idempotent) matrix $\mathbf{C} := \mathbf{I}_n - \mathbf{1}\mathbf{1}^\top/n$ and let

$$\eta_i^2 := \frac{\|\mathbf{C}\boldsymbol{v}_i\|^2}{n} = \frac{\boldsymbol{v}_i^\top \mathbf{C}\boldsymbol{v}_i}{n}, \quad i = 1, \ldots, p$$

be the empirical variance of the components of the i-th feature vector. We define the Lasso estimate (Tibshirani 1996) of $\boldsymbol{\beta}$ as the solution to the penalized least squares:

$$(\hat{\beta}_{0,\lambda}, \hat{\boldsymbol{\beta}}_\lambda) = \underset{b_0, \boldsymbol{b}}{\operatorname{argmin}} \frac{\|\boldsymbol{y} - \mathbf{1}b_0 - \mathbf{X}\boldsymbol{b}\|^2}{2n} + \lambda \sum_{i=1}^{p} \eta_i \times |b_i|, \tag{1}$$

where $\lambda > 0$ is a suitably chosen regularization parameter and the intercept b_0 is not penalized.

Feature Standardization. As mentioned in the abstract, it is common practice to standardize the features $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ so that the variance of each $\boldsymbol{v}_i$ is unity (Hastie et al. 2015; Tibshirani 1996). This standardization ensures that the Lasso estimate $\hat{\boldsymbol{\beta}}_\lambda$ is not affected by the units in which the features are measured, and in general, improves the performance of the estimator (Hastie et al. 2001). The standardization can be accomplished by working with the matrix $\mathbf{XS}$, rather than $\mathbf{X}$, where $\mathbf{S}$ is the rescaling matrix

$$\mathbf{S} := \operatorname{diag}(\eta_1^{-1}, \ldots, \eta_p^{-1}).$$

Extensions to GLMs. Suppose that the joint density of the response vector $\boldsymbol{Y}$ given $\beta_0, \boldsymbol{\beta}, \mathbf{X}$ is $g(\boldsymbol{y} \,|\, \beta_0, \boldsymbol{\beta}, \mathbf{X})$, where the dependence on $\beta_0, \boldsymbol{\beta}, \mathbf{X}$ is through the linear map $(\beta_0, \boldsymbol{\beta}) \mapsto \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta}$. Here, the cross-entropy training loss (Kroese et al. 2019) (negative average log-likelihood) is $-\frac{1}{n} \ln g(\boldsymbol{y} \,|\, \beta_0, \boldsymbol{\beta}, \mathbf{X})$ and the extension of the Lasso estimator to the setting of generalized linear models is then given by :

$$(\hat{\beta}_{0,\lambda}, \mathbf{S}^{-1}\hat{\boldsymbol{\beta}}_\lambda) = \underset{b_0, \boldsymbol{b}}{\operatorname{argmin}} \frac{-\ln g(\boldsymbol{y} \,|\, b_0, \boldsymbol{b}, \mathbf{XS})}{n} + \lambda \sum_{i=1}^{p} |b_i|.$$

Observe that, just like in the linear Lasso estimator, we scale the features so that their variance is unity (Hastie et al. 2015). This scaling need only be applied once on $\mathbf{X}$, possibly as a preprocessing step prior to the main optimization, and then reversed at the end of the optimization.

New Rescaling Method for GLMs. Let $r(b_0, \boldsymbol{b}) := -\ln g(\boldsymbol{y} \,|\, b_0, \boldsymbol{b}, \mathbf{X})/n$ be our shorthand notation for the cross-entropy loss. We define $\eta_i^2(\boldsymbol{\beta}), i = 1, \ldots, p$, to be the i-th diagonal element of the $p \times p$ Hessian matrix of second derivatives:

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \min_{b_0} r(b_0, \boldsymbol{\beta}).$$

This is the Hessian matrix of the cross-entropy loss, evaluated at the true parameter $\boldsymbol{\beta}$, and after the nuisance parameter β_0 is eliminated from the optimization. Then, instead of the usual rescaled Lasso estimator, we propose the following alternative *iteratively rescaled Lasso* (IRL):

$$\underset{\boldsymbol{b}, b_0}{\operatorname{argmin}} \frac{-\ln g(\boldsymbol{y} \,|\, b_0, \boldsymbol{b}, \mathbf{X})}{n} + \lambda \sum_{i=1}^{p} \eta_i(\boldsymbol{\beta}) \times |b_i|. \tag{2}$$

We now make three observations. First, since each $\eta_i(\boldsymbol{\beta})$ depends on the unknown $\boldsymbol{\beta}$, the approximate computation of (2) will be iterative and is the main reason for naming the method IRL.

Second, the linear Lasso estimator (1) is a special case of (2) when $\boldsymbol{Y}$ is a multivariate Gaussian with mean $\mathbb{E}_{\mathbf{X}}[\boldsymbol{Y}] = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta}$ and variance $\mathbb{V}\mathrm{ar}_{\mathbf{X}}(\boldsymbol{Y}) = \mathbf{I}$, because then $\eta_i^2(\boldsymbol{\beta}) = \|\mathbf{C}\boldsymbol{v}_i\|^2/n$.

Third, one may ask what is the motivation for the proposed IRL estimator. The answer is that the IRL estimator coincides with the traditional linear regression estimator (1), provided that the cross-entropy loss $r(b_0, \boldsymbol{b})$ is replaced by its quadratic approximation in the neighborhood of the true coefficients $\beta_0, \boldsymbol{\beta}$. In other words, our proposed IRL estimator uses exactly the same scaling as the linear Lasso estimator (1) when the generalized linear model is *linearized* at the true solution. Note that there is no such agreement in the scaling between the currently accepted linear estimator (1) and its GLM counterpart, that is, the current widely-used scaling is not consistent across linear and nonlinear models. Our proposal is thus motivated by the desire for consistency in the scaling applied to linear and nonlinear models.

REFERENCES

Hastie, T., R. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hastie, T., R. Tibshirani, and M. Wainwright. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. New York, NY, USA: Chapman & Hall/CRC.

Kroese, D. P., Z. Botev, T. Taimre, and R. Vaisman. 2019. *Data Science and Machine Learning: Mathematical and Statistical Methods*. New York, NY, USA: Chapman and Hall/CRC.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1):267–288.