Stable LASSO for High-Dimensional Feature Selection through Proximal Optimization

Roman Zakharov, Pierre Dupont
Machine Learning Group, ICTEAM Institute,
Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium
{roman.zakharov,pierre.dupont}@uclouvain.be

Abstract: The l_1-norm regularization is commonly used when estimating (generalized) linear models while enforcing sparsity. The automatic feature selection embedded in such an estimation is however known to be highly unstable since, among correlated features, an l_1 penalty tends to favor the selection of a single feature, essentially picked at random.
This paper introduces a modified optimization objective to stabilize LASSO or similar approaches. The solution to this modified problem is constrained by a norm ball rescaled according to the variances of the predictor variables. We further describe how such problems can be efficiently solved through proximal optimization.
Classification experiments conducted on several microarray datasets show the benefits of the proposed approach, both in terms of stability and predictive performances, as compared to the original LASSO, Elastic Net, Trace LASSO and a simple variance based filtering.

Keywords: feature selection, regularization, stability, LASSO, proximal optimization

1 Introduction

Feature selection aims at improving the interpretability of predictive models and at reducing the computational cost when predicting from new observations. Such a selection is also desirable when it is *a priori* known that the model should be sparse or to prevent overfitting. This is especially relevant when the number p of input features, or predictor variables, largely exceeds the number n of training observations. In such contexts, feature selection can also increase the predictive performances.

The l_1-norm is often used to regularize (generalized) linear models while performing an automatic feature selection by driving most model coefficients towards zero. The LASSO method [2] precisely combines such a regularization with a least square loss for estimating a regression model.

Predictive models estimated with a LASSO penalty are however known to be highly unstable, which means that small data perturbations can imply drastic changes in the subset of automatically selected variables. The lack of stability of the LASSO is generally attributed to the fact that, among several correlated features, an l_1 penalty tends to favor the selection of a single feature, essentially picked at random. In contrast, univariate filter methods, such as a t-test feature ranking, rely on general statistical characteristics of the data, which are much less sensitive to small data perturbations. Such simple selection methods are typically more stable but ignore the possible correlations between features and are not embedded into the estimation of a predictive model.

The S-LASSO method detailed in section 2 relies on a modified optimization objective to stabilize the LASSO. The solution to this modified problem is constrained by a norm ball rescaled according to the variances of the predictor variables. In contrast to the Elastic Net [5] and Trace LASSO [1] approaches, which favor the joined selection of correlated features, S-LASSO tends to discard low variance features because they are expected to be less informative.

2 A scaled proximal method for feature selection

Let $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_n})^T \in \mathbb{R}^{n \times p}$ be the design matrix made of n training observations in $\mathbb{R}^p$, and $\mathbf{y} \in \mathbb{R}^n$ be the response vector. Learning the weight vector $\mathbf{w}$ of a simple linear model $y = \mathbf{w}^T\mathbf{x} + \varepsilon$, where ε denotes a Gaussian noise with 0 mean and variance σ^2, is commonly phrased as a convex optimization problem of the form

$$\min_{\mathbf{w} \in \mathbb{R}^p} f(\mathbf{w}) + \Omega(\mathbf{w}), \tag{1}$$

where $f : \mathbb{R}^p \to \mathbb{R}$ is a convex differentiable loss function and $\Omega : \mathbb{R}^p \to R$ is a convex norm, not necessarily smooth or Euclidean. The Ω regularization term aims at reducing over-fitting by penalizing large absolute weight values, for which small input changes would have a significant impact on the predicted output.

In this work we propose to modify the general optimization problem (1) while rescaling the norm penalty according to individual feature variances:

$$\min_{\mathbf{w}\in R^p} f(\mathbf{w}) + \lambda \sum_{j=1}^{p} \frac{1}{r_j} \Omega(w_j), \qquad (2)$$

where vector $\mathbf{r} \in \mathbb{R}^p$ is proportional to the feature variances. We note that those variances are estimated *before* centering and normalizing the data to unit variance as usual when estimating a LASSO model. The proposed method also offers a general framework beyond this specific choice of variance weighting. In practice, any vector r can be used to favor the selection of some variables a priori believed to be more relevant. This modified objective can be straightforwardly used with any penalty that can be decomposed component-wise. We focus on the l_1-norm as a special case of interest but we also report positive experimental results with an Elastic Net penalty. As compared to (1), the regularization ball associated to $\Omega(\mathbf{w})$ now takes the form of an ellipsoid elongated along the directions of higher variance. In other words, the regularization constant λ is rescaled in a component-wise fashion.

The proposed method is related to the Adaptive LASSO [4] which also penalizes the l_1 penalty component-wise with adaptive weights. These weights are initially equal to the ordinary least square estimates and iteratively updated under the control of an additional tuning parameter. In contrast, S-LASSO does not require such an additional parameter and iterative reweighting as it relies on the observed variances along each dimension. The proposed framework is also not restricted to the l_1 penalty. Modification to the LASSO by some form of variance weighting has already been proposed in [3]. This related work describes bounds on the prediction error and oracle properties while we focus on the improved stability of the embedded feature selection as a result of such variance weighting. We also show how the S-LASSO modified objective can be efficiently solved through proximal optimization.

The modified objective (2) promotes stability since the identity of high variance features is expected not to change much while varying the data sampling. As shown experimentally, the predictive performances may also be improved since low variance features across learning observations are expected to be less informative for prediction. We also show that solving (2) offer better results than simply pre-filtering features based on their variances. We further detail how this modified objective can be efficiently solved through proximal optimization.

3 Results

We report experimental performances of the proposed method with a LASSO or Elastic Net penalty, and refer to those approaches as S-LASSO and S-ENET respectively. The competing approaches are the original LASSO or ENET with a logistic loss. We also report the performances obtained with TRACE LASSO adapted to a classification problem. Since S-LASSO and S-ENET use the individual feature variances to modify the optimization objective, we also compare to VARIANCE_RANKING, which is a filter method keeping only a desired number of features with the largest variances. Our experiments conducted on 5 microarray datasets illustrate the benefits of the proposed approach both in terms of stability of the gene selection and the classification performance, as compared to the original LASSO, ELASTIC NET or TRACE LASSO. In contrast, the stability of VARIANCE_RANKING is always very high but the predictive performances drop drastically when reducing the number of selected features.

Acknowledgments
Roman Zakharov is supported by a FRIA grant (5.1.191.10.F).

References

[1] Edouard Grave, Guillaume Obozinski, and Francis Bach. Trace Lasso: a trace norm regularization for correlated designs. In *Advances in Neural Information Processing Systems*, Granada, Spain, 2011.

[2] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[3] Sara van de Geer, Peter Buehlmann, and Shuheng Zhu. The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). *Electronic Journal of Statistics*, 5:688–749, 2011.

[4] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, Vol. 101, No. 476, Theory and Methods:1418–1429, 2006.

[5] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.