Computational Statistics and Data Analysis 64 (2013) 281–298

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Logistic regression with weight grouping priors

M. Korzeń[a], S. Jaroszewicz[b,c], P. Klęsk[a,*]

[a] Faculty of Computer Science and Information Technology, West Pomeranian University of Technology, Szczecin, Poland
[b] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
[c] National Institute of Telecommunications, Warsaw, Poland

ARTICLE INFO

Article history:
Received 7 April 2012
Received in revised form 12 March 2013
Accepted 12 March 2013
Available online 21 March 2013

Keywords:
Logistic regression
Regularization
Weight grouping
Attributes grouping

ABSTRACT

A generalization of the commonly used Maximum Likelihood based learning algorithm for the logistic regression model is considered. It is well known that using the Laplace prior (L^1 penalty) on model coefficients leads to a variable selection effect, when most of the coefficients vanish. It is argued that variable selection is not always desirable; it is often better to group correlated variables together and assign equal weights to them. Two new kinds of a priori distributions over weights are investigated: Gaussian Extremal Mixture (GEM) and Laplacian Extremal Mixture (LEM) which enforce grouping of model coefficients in a manner analogous to L^1 and L^2 regularization. An efficient learning algorithm is presented, which simultaneously finds model weights and the hyperparameters of those priors. Examples are shown in the experimental part where the proposed a priori distributions outperform Gauss and Laplace priors as well as other methods which take coefficient grouping into account, such as the elastic net. Theoretical results on parameter shrinkage and sample complexity are also included.

1. Introduction

Variable selection problem for linear models is considered one of the most important in statistical inference (Hesterberg et al., 2008). Recently, many new variable selection methods became popular, including stagewise selection (Hastie et al., 2009) and L^1-regularization techniques such as *Lasso* (Williams, 1994; Tibshirani, 1996; Mkhadri and Ouhourane, 2013) and *LARS* (Efron et al., 1996). However, variable selection is not always the best possible approach.

If the predictor variables are correlated, it is often more desirable to group correlated variables together and assign them equal or similar weights. A more detailed justification is given in Section 2, where it is argued that variable *averaging* may give much better results than variable selection.

In order to achieve such averaging, we devise a supervised learning algorithm maximizing the log-likelihood criterion with suitably chosen priors on model weights. Our priors correspond to mixtures of Gaussian or Laplace distributions which force the weights to cluster around the means of the mixture components. The priors work analogously to L^1 and L^2 regularization, such that the prior based on the Gaussian distribution forces the weights to lie close to their group averages, and the prior based on the Laplace distribution forces most weights to be *exactly equal* to their group averages. As the resulting optimization problem is nonconvex, we present an algorithm, similar to the EM approach, consisting of two repeated steps: (1) maximization of log-likelihood for the current assignment of variables to groups, and (2) re-assignment of variables with identical or similar weights to appropriate groups. Theoretical properties of the proposed method, such as parameter shrinkage and sample complexity have also been analyzed.

[*] Correspondence to: ul. Żołnierska 49, 71-210 Szczecin, Poland. Tel.: +48 91 4495556; fax: +48 91 4495990.
E-mail addresses: mkorzen@wi.zut.edu.pl (M. Korzeń), s.jaroszewicz@ipipan.waw.pl (S. Jaroszewicz), pklesk@wi.zut.edu.pl (P. Klęsk).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.03.013

The advantages of coefficient grouping and averaging have, of course, already been recognized by researchers, and several methods which allow for weight grouping in regression models have been proposed. We will now review those approaches and explain the differences from the method proposed in this paper.

Zou and Hastie (2005) introduce a method called *elastic net*, which combines L^1 and L^2 regularization. The L^1 term enforces variable selection, while the L^2 term introduces a 'grouping effect', thanks to which correlated variables tend to have similar coefficients. The grouping effect is, however, just a by-product of the regularization method used, and it is thus difficult to control its strength; typically it is not possible to enforce equal or approximately equal weights. In contrast, our method allows for direct control over the strength of the grouping effect (which we demonstrate theoretically) and coefficients of correlated variables can be forced to lie arbitrarily close to each other. Moreover, the prior based on the mixture of Laplace distributions allows for enforcing strict equality of most weights to their respective group averages.

A technique called *group Lasso* has been described in Yuan and Lin (2004), Kim et al. (2006) (and introduced earlier by Bakin, 1999), which extends Lasso by taking into account the group structure of variables. However the groups need to be specified in advance and incorporated into the regularization term. Our method, on the other hand, groups variables automatically. Moreover, group Lasso does not allow for direct control over the relative sizes of weights within groups, while our approach gives the analyst precise control of the grouping behavior.

If attributes are ordered in some natural way (e.g. in time series data), there is an interesting approach called *fused lasso*, where both large weight values and large differences between consecutive weights are penalized (Friedman et al., 2007; Tibshirani et al., 2005). The approach has been generalized to image data by requiring similar weights for variables corresponding to adjacent pixels. Our motivation is different, as we require whole groups of attributes to have similar weights, not just consecutive or adjacent ones.

The remaining part of the paper is organized as follows: we present the motivation in Section 2, give a detailed description of the proposed method in Section 3, describe the optimization algorithm in Section 4, and analyze the approach theoretically and experimentally in Sections 5 and 6. Section 7 concludes.

2. Motivation and a simplified model

Consider a simplified model with s independent hidden variables H_k, $k = 1, \ldots, s$, which are not observed directly, but which are directly correlated with the target variable Y. Instead of H_k, we can only observe a set of variables X_j, $j = 1, \ldots, n$, which are noisy observations of the H_k, each X_j depending on a single hidden variable H_k. More precisely, for every j, $X_j = H_k + \xi_j$ for some k, where ξ_j is a random noise term with zero mean and variance σ_k^2, equal for all X_j depending on a given H_k. Assume further, that $\mathrm{Cov}(\xi_j, H_k) = \mathrm{Cov}(\xi_j, Y) = 0$ for all values of j and k, and $\mathrm{Cov}(\xi_i, \xi_j) = 0$ for all $i \neq j$, but for all k, $\mathrm{Cov}(Y, H_k) \neq 0$. Our task is to construct a linear predictor of Y based on X_j's:

$$\hat{Y} = \sum_{j=1}^{n} \alpha_j X_j.$$

We now state a lemma, which underlies the main motivation of the paper. We first introduce some additional notation. Two indexing functions $\kappa, \jmath$ will be used: $\kappa(j)$ gives an index from $\{1, \ldots, s\}$ such that X_j is a noisy observation of $H_{\kappa(j)}$; $\jmath(k, j)$ gives an index from $\{1, \ldots, n\}$ of the j-th variable dependent on H_k. Further, let n_k denote the number of variables being the noisy observations of H_k.

Lemma 2.1. *Suppose H_k, $k = 1, \ldots, s$, are the independent hidden variables in the model described above, and $X_j = H_{\kappa(j)} + \xi_j$, $j = 1, \ldots, n$, are the observed variables. Assume all noise terms ξ_j have zero mean and variance $\sigma^2_{\kappa(j)}$ (equal for all noisy observations of H_k). Then, for any constants A_k and any numbers $\alpha_{\jmath(k,j)} \geq 0$ such that $\sum_{j=1}^{n_k} \alpha_{\jmath(k,j)} = 1$ (for all k) we have:*

$$E\left(\sum_{k=1}^{s} A_k \sum_{j=1}^{n_k} \alpha_{\jmath(k,j)} X_{\jmath(k,j)} - Y\right)^2 \geq E\left(\sum_{k=1}^{s} A_k \frac{1}{n_k} \sum_{j=1}^{n_k} X_{\jmath(k,j)} - Y\right)^2. \tag{1}$$

The proof can be found in Appendix A.1. The interpretation of the lemma is as follows: if there exists a set of independent hidden variables H_k, which are related to the target variable Y by a linear relationship

$$Y = A_1 H_1 + \cdots + A_s H_s + \xi, \tag{2}$$

where ξ is a noise term with zero mean, and a set of observed variables X_j (being noisy observations of the H_k's), then the best modeling approach is to form averages of observed variables in each group $\left(\frac{1}{n_k} \sum_{j=1}^{n_k} X_{\jmath(k,j)}\right)$ and to build a linear model based on these averages. Moreover, one can see that the averages reconstruct the hidden variables H_k and that the coefficients corresponding to the averages are good estimates of A_k's.

Unfortunately in practice we do not know which X_j's correspond to which hidden variables or how close their relationship is. In the following two sections we introduce an algorithm which allows for finding groupings of variables which presumably are noisy observations of the same hidden variable, and which automatically estimates how close the coefficient of each variable in a group should be to the group's average.