Computational Statistics and Data Analysis 71 (2014) 771–786

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

# Sparse group lasso and high dimensional multinomial classification

Martin Vincent *, Niels Richard Hansen

*University of Copenhagen, Department of Mathematical Sciences, Universitetsparken 5, 2100 Copenhagen Ø, Denmark*

ARTICLE INFO

*Article history:*
Received 5 May 2012
Received in revised form 30 April 2013
Accepted 4 June 2013
Available online 11 June 2013

*Keywords:*
Sparse group lasso
Classification
High dimensional data analysis
Coordinate gradient descent
Penalized loss

ABSTRACT

The sparse group lasso optimization problem is solved using a coordinate gradient descent algorithm. The algorithm is applicable to a broad class of convex loss functions. Convergence of the algorithm is established, and the algorithm is used to investigate the performance of the multinomial sparse group lasso classifier. On three different real data examples the multinomial group lasso clearly outperforms multinomial lasso in terms of achieved classification error rate and in terms of including fewer features for the classification. An implementation of the multinomial sparse group lasso algorithm is available in the R package `msgl`. Its performance scales well with the problem size as illustrated by one of the examples considered—a 50 class classification problem with 10 k features, which amounts to estimating 500 k parameters.

© 2013 Elsevier B.V. All rights reserved.

## 1. Introduction

The sparse group lasso is a regularization method that combines the lasso (Tibshirani, 1994) and the group lasso (Meier et al., 2008). Friedman et al. (2010a) proposed a coordinate descent approach for the sparse group lasso optimization problem. Simon et al. (2013b) used a generalized gradient descent algorithm for the sparse group lasso and considered applications of this method to linear, logistic and Cox regressions. We present a sparse group lasso algorithm suitable for high dimensional problems. This algorithm is applicable to a broad class of convex loss functions. In the algorithm we combine three non-differentiable optimization methods: the coordinate gradient descent (Tseng and Yun, 2009), the block coordinate descent (Tseng, 2001) and a modified coordinate descent method.

Our main application is to multiclass classification based on the multinomial regression model. The lasso penalty has, for some time, been considered as a regularization approach for multinomial regression (Friedman et al., 2010b). The parameters in the multinomial model are, however, naturally structured, with multiple parameters corresponding to one feature, and the lasso penalty does not take this structure into account. To accommodate for this we suggest to add a group lasso term with the parameters corresponding to the same feature grouped together. The resulting penalty is known as the sparse group lasso penalty. We found that using the sparse group lasso penalty for multinomial regression generally improved the performance of the estimated classifier and reduced the number of features included in the model.

The formulation of an efficient and robust sparse group lasso algorithm is not straightforward due to non-differentiability of the penalty. First, the sparse group lasso penalty is not completely separable, which is problematic when using a standard coordinate descent scheme. To obtain a robust algorithm an adjustment is necessary. Our solution, which efficiently treats the singularity at zero that cannot be separated out, is a minor modification of the coordinate descent algorithm. Second, our algorithm is a Newton type algorithm, hence we sequentially optimize penalized quadratic approximations of the loss function. This approach raises another challenge: how to reduce the costs of computing the Hessian? In Section 3.6 we show

* Corresponding author. Tel.: +45 22860740.
*E-mail addresses:* vincent@math.ku.dk, martin.vincent.dk@gmail.com (M. Vincent).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.06.004

that an upper bound on the Hessian is sufficient to determine whether the minimum over a block of coefficients is attained at zero. This approach enables us to update a large percentage of the blocks without computing the complete Hessian. In this way we reduce the run-time, provided that the upper bound of the Hessian can be computed efficiently. We found that this approach reduces the run-time on large data sets by a factor of more than 2.

Our focus is on applications of the multinomial sparse group lasso to problems with many classes. For this purpose we have investigated three multiclass classification problems. We found that multinomial group lasso and sparse group lasso perform well on these problems. The error rates were substantially lower than the best obtained with multinomial lasso, and the low error rates were achieved for models with fewer features having non-zero coefficients. For example, we consider a text classification problem consisting of Amazon reviews with 50 classes and 10 k textual features. This problem showed a large improvement in the error rates: from approximately 40% for the lasso to less than 20% for the group lasso.

We provide a generic implementation of the sparse group lasso algorithm in the form of a `C++` template library. The implementation for multinomial and logistic sparse group lasso regressions is available as an R package. For our implementation the time to compute the sparse group lasso solution is of the same order of magnitude as the time required for the multinomial lasso algorithm as implemented in the R package glmnet. The computation time of our implementation scales well with the problem size.

### 1.1. Sparse group lasso

Consider a convex, bounded below and twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. We say that $\hat{\beta} \in \mathbb{R}^n$ is a *sparse group lasso minimizer* if it is a solution to the unconstrained convex optimization problem

$$\text{minimize } f + \lambda\Phi \tag{1}$$

where $\Phi : \mathbb{R}^n \to \mathbb{R}$ is the *sparse group lasso penalty* (defined below) and $\lambda > 0$.

Before defining the sparse group lasso penalty some notation is needed. We decompose the search space

$$\mathbb{R}^n = \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m}$$

into $m \in \mathbb{N}$ blocks having dimensions $n_i \in \mathbb{N}$ for $i = 1, \ldots, m$, hence $n = n_1 + \cdots + n_m$. For a vector $\beta \in \mathbb{R}^n$ we write $\beta = (\beta^{(1)}, \ldots, \beta^{(m)})$ where $\beta^{(1)} \in \mathbb{R}^{n_1}, \ldots, \beta^{(m)} \in \mathbb{R}^{n_m}$. For $J = 1, \ldots, m$ we call $\beta^{(J)}$ the $J$'th *block* of $\beta$. We use the notation $\beta_i^{(J)}$ to denote the $i$'th coordinate of the $J$'th block of $\beta$, whereas $\beta_i$ is the $i$'th coordinate of $\beta$.

**Definition 1** (*Sparse Group Lasso Penalty*)**.** The sparse group lasso penalty is defined as

$$\Phi(\beta) \stackrel{\text{def}}{=} (1-\alpha)\sum_{J=1}^{m} \gamma_J \left\|\beta^{(J)}\right\|_2 + \alpha \sum_{i=1}^{n} \xi_i |\beta_i|$$

for $\alpha \in [0, 1]$, group weights $\gamma \in [0, \infty)^m$, and parameter weights $\xi = (\xi^{(1)}, \ldots, \xi^{(m)}) \in [0, \infty)^n$ where $\xi^{(1)} \in [0, \infty)^{n_1}, \ldots, \xi^{(m)} \in [0, \infty)^{n_m}$.

The sparse group lasso penalty includes the lasso penalty ($\alpha = 1$) and the group lasso penalty ($\alpha = 0$). Note also that for sufficiently large values of $\lambda$ the solution of (1) is zero. The infimum of these, denoted $\lambda_{\max}$, is computable, see Section 3.2.

We emphasize that the sparse group lasso penalty is specified by

- a grouping of the parameters $\beta = (\beta^{(1)}, \ldots, \beta^{(m)})$,
- and the weights $\alpha$, $\gamma$ and $\xi$.

It is well known that the lasso penalty results in sparse solutions to (1), while the group lasso penalty results in groupwise sparse solutions (that is, the entire group of parameters is zero or non-zero). However group lasso does not give sparsity within groups — sparse group lasso does.

In the second part of the paper we develop an algorithm for solving the optimization problem (1). The convergence of the algorithm is established for any sparse group lasso penalty, regardless of how the parameters are grouped. For multinomial regression, as considered in the next section, we restrict attention to a specific grouping of the parameters that reflects the features. In the symmetric parametrization of the multinomial regression model with $K$ classes there are $K$ parameters per feature. Our suggestion is to group these $K$ parameters together. Thus we do not group the features, only the parameters associated with each feature. For the examples we considered this particular grouping resulted in models with fewer features having non-zero parameters compared to ordinary lasso penalization. More importantly, the error rates were typically also smaller.

Our `msgl` R package supports the particular grouping for multinomial regression as well as additional groupings of the features, i.e. the number of parameters in each group is a multiple of $K$. The `sgl` `C++` template library can be configured to handle any grouping.

Download English Version:

https://daneshyari.com/en/article/6870515

Download Persian Version:

https://daneshyari.com/article/6870515

Daneshyari.com