Contents

1 Variable Selection with Bootstrap in GLM Model for Massive Data

1.1 Summary

Variable Selection with Bag of Little Bootstraps for Generalized Linear Regression

One method for variable selection in a linear regression setting is the LASSO. Traditional linear regression involves finding a solution, β, that minimizes $\|Y - X\beta\|_2^2$. In the LASSO, we trade a small bit of bias, determined by the choice of a regularizatieon hyperparameter, λ, to *shrink* the parameters toward zero, and even setting some to zero. The problem is equivalent to minimizing

$$\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

We can find a solution by applying the Cyclic Coordinate Descent algorithm to the above equation.

One of the drawbacks of the major drawbacks of t

he LASSO is the lack of a measure of uncertainty of the estimate, in the sense of how sensitive it is to the obs

erved data versus the true data generating mechanism. We can asses this variablity by using a bootstrap variable selection procedure. Essentially, we use bootstrap resamples and a model averaging voting scheme similar to random forests. A regression coefficent recieves a vote for each resample that estimates the parameter to be non-zero. The proportion of votes to resamples above a threshold is then used to select variables.

In the context of a massive dataset, the computational cost of even a small number of boostrap resamples becomes very large. However, given that we have a large dataset, we can use only a portion of our data for each replicate. This allows us to drastically reduce the computational cost while still obtaining a good bootstrap estimate.

2 BLBVS

2.1 Motivation

The Bag of Little Bootstraps variable selection method was introduced to give a boostrap variable selection process for massive data sets.

The method should have comparable accuracy to the naive estimator, but hopefully be much faster computationally.

2.2 Bag of Little Bootstraps

The algorithm for the BLBVS procedure is as follows. Partition the set of observed data $(x_1, ..., x_n)$ into s randomly selected subsets (called bags) of size $b = \{n^\gamma : \gamma \in 0, 1\ \}$ chosen *without replacement*. Each subset produces r bootstrap resamples chosen *with replacement* of size n, the original sample size. Notice that each resample can only have at most b distinct data points.

2.2.1 Notation

2.2.2 Algorithm

Bag of Little Bootstraps Variable Selection (BLBVS)

```
Input: Original data, $X = (x_1, ..., x_n)$
$b$ : size of each subset
$s$ : number of subsets
$r$ : number of resamples
$\beta$ : parameter vector
$\xi$ : assessmend of estimator quality
Output: An estimate of $\xi$ and the selection result proportion p
for $i = 1$ to $s$ do
    Randomly select a subset $\mathcal{I} = \{i_1, ..., i_b\}$ of $b$ objects from $\{1, 2, ..., n\}$ without replacement
    Form the subset $\tilde{X}^{(i)}$ based on the original dataset $X$ and index set $\mathcal{I}$
    for $j = 1$ to $r$ do
        Generate $r$ bootstrap resamples: randomly draw a resample $X^*_{ij} = (\tilde{X}^{(i)}; n^{*ij})$ of size $n$ from subset $\tilde{X}_i$ with replacement, where each resample is $Multinomial(n, (1/b)\mathbf{1}_b)$ distributed with a weight vector of parameters $\beta$ and calculate $\mathbf{1}(\hat{\beta}(X^*_{ij}))$.
    end for
    Get $r$ estimated results of parameter vector $\beta$ and then calculate the standard error a parameter estimation $\xi^{*(i)}$
end for
Compute the final estimate of $\xi$ by $s^{-1} \sum_{i=1}^{s} \xi_i^{*(i)}$
return the proportion $p$
```

2.3 BLB Variable Selection

2.3.1 Voting

2.3.2 Selection

Let $\hat{\beta_{ij}}$ be a vetor of estimators for the jth resample of the ith subset, and $\hat{\xi}_i$ be the standard error of the parameter estimation in the ith subset. The overall standard error $\hat{\xi}$ is simply the average over the $\hat{\xi}_i$.

To perform variable selection, we consider indicator functions, $\mathbf{I}\left(\hat{\beta}_{ijk} \neq 0\right)$, to represent whether or not the parameter estimate k in the jth resample of the ith subset was set to zero. The selection procedure is based on model averaging using a vote criterion similar to Random Forests. Each bootstrap sample is given an equally weighted vote, for a total of $s \times r$ votes. A predictor is selected if it recieves a proportion of votes, p_k, greater than some pre-chosen cutoff c. Where p_k is defined as

$$p_k = \frac{\sum_{i=1}^{s} \sum_{j=1}^{r} \mathbf{I}(\hat{\beta}_k(X_{ij}) \neq 0)}{s \times r}.$$

2.4 BLBVS with Group Lasso Penalty for GLM

2.4.1 Extending to GLM

The variable selection method (BLBVS) works perfectly well for a traditional linear regression with continuous predictors. However, it does not handle generalized linear models with categorical predictors or different link functions properly.

For a categorical predictor with k categories, $k-1$ dummy variables must be added to the model. Under the original Lasso framework, the dummy variables are penalized individually. This may cause a partial choice of some of levels of the predictor. Ideally, all the dummy variables from a certain category should be chosen or eliminated together. An extension to the Lasso, called the Group Lasso, can overcome these issues by grouping sets of related dummy variables together and penalizing them together. The estimator is defined as

$$\hat{\beta}_\lambda = argmin(\|Y - X\beta\|_2^2 + \lambda \sum_{g=1}^{G} \|\beta I_g\|_2)$$

where I_g is the index set of the gth group of variables, $g = 1, ..., G$.

To extend the method to a GLM framework, the cost function must be extended to incorporate the link function. The first term of the Lasso cost function then becomes $\|Y - g^{-1}(X\beta)\|_2^2$ insetad of simply $\|Y - X\beta\|_2^2$. Consider the following setup of logistic regression for more details. Let Y_i be a binary response variable, and $\mathbf{X}$ be the matrix of p independently and idendically distributed predictor variables. The predictor variables are divided into G groups. Each continuous predictor is its own group with $df_g = 1$. Each set of dummy variables pertaining to a single categorical variable (with k levels) is collected as one group with $df_g = k - 1$. $\mathbf{X}$ can now be partitioned groupwise by columns so that the linear model becomes

$$g(\mathbf{Y}) = log\left(\frac{p_\beta(\mathbf{X})}{1 - p_\beta(\mathbf{X})}\right) = \beta_0 + \sum_{g=1}^{G} X_g^T \beta_g$$

where $p_\beta(x) = P_\beta(Y = 1|x)$ and $\beta_g \in \mathcal{R}^{df_g}$. This problem is equivalent to minimizing the convex function:

$$S_\lambda(\beta) = -l(\beta) + \lambda \sum_{g=1}^{G} s(df_g) \|\beta_g\|_1$$

where $l(\cdot)$ is the log-likelihood and $s(\cdot)$ is chosen to rescale the regularization term with respect to the degrees of freedom in the group. Typically, the value of λ is chosen via cross-validation.

2.5 Simulations

2.5.1 Setup

2.5.2 Results

2.5.3 Evaluation

For the simulation analysis of the BLBVS method, they considered logistic regression with continuous independent predictors as the true model and Group Lasso as the penalty funciton. In this paper, they performed simulations considering $b = n^\gamma$ with $\gamma \in \{0.6, 0.7, 0.8\}$ and let $r = 100$ in each subset.

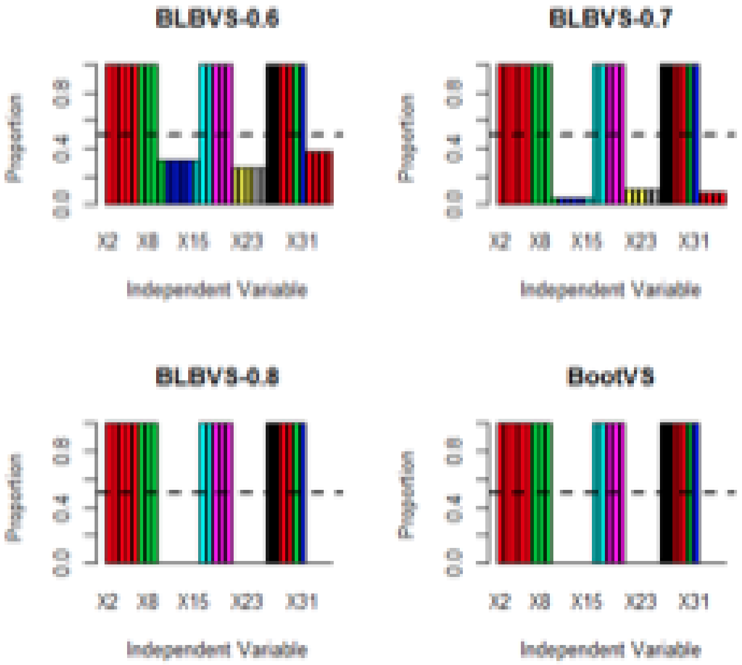
The simulated dataset was set to have $n = 20,000$ and 35 continuous predictors goruped into 8 groups:

Index Var	Index Variables
$\mathcal{I}_1$	$\{1, ..., 5\}$
$\mathcal{I}_2$	$\{6, ..., 9\}$
$\mathcal{I}_3$	$\{10, ..., 15\}$
$\mathcal{I}_4$	$\{16, ..., 20\}$
$\mathcal{I}_5$	$\{21, ..., 25\}$
$\mathcal{I}_6$	$\{26, ..., 28\}$
$\mathcal{I}_7$	$\{29, ..., 31\}$
$\mathcal{I}_8$	$\{32, ..., 35\}$

Coefficients were chosen groupwise to be included in the true model, T, and if $\mathcal{I}_g \in \mathcal{I}_T$, then the corresponding parameter was given a value of 10. Variables not chosen for the true model, were given random values chosen from a $N(0, 1)$ distribution. These are intended to be regularized to zero. Simulated Y_i were drawn from independent Bernoulli distributions where,

$$p_\beta(x) = \left[1 + \exp -\tilde{\mathbf{X}}_{i,G}^T \beta_G\right]^{-1}.$$

For the simulations, $\mathcal{I}_T = \mathcal{I}_{1,2,4,6,7}$. The figure below shows the selection results for BLBVS with the various values of γ as well as the results for the BootVS without bagging.



The selection results perform equally well under both procedures, even when we set a simple cutoff at being chosen in 50 percent of the replicates. The results show that the proportion estimate becomes more accurate as γ, the proportion of the original dataset included in each bagged subset, increases. Logically, the performance of BootVS is superior to BLBVS when γ is small since it resamples from all distinct data points instead of subsets. It has been estimated that each resample in BootVS contains approximately $0.632n$ different data points in comparison to BootVS's n^{γ}. Thus, it makes sense that the superiority of BootVS decreases as γ increases, and the figure shows that they are effectively equivalent when $\gamma = 0.8$. Notice however, that BootVS and BLBVS select the correct model.

2.6 TODO Convergence Properties

2.6.1 Relative Deviation

2.7 TODO Scalability and Computational Analysis

2.7.1 Performance of BLBVS

2.7.2 Paralellization of Bags

2.8 Real Data Analysis

2.8.1 Data

Credit card records from Taiwanese bank

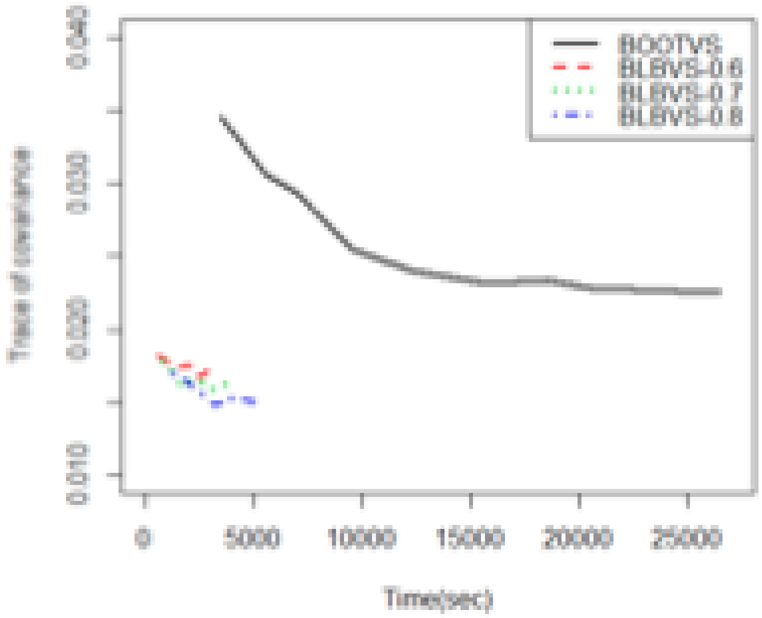
- Goal: Select predictors that important for predicting credit risk
- Greater than 800 million observations (11 GB)
- 25 Categorical predictors
- Binary response (1 - Risky, 0 - Not risky)

2.8.2 Selected Variables

Consistent across methods

- Force to stop credit card
- Living Area
- Education Background
- Occupation
- Housing Situation
- Avg. Monthly Income
- Family Economic Level

2.8.3 Computational Performance



3 Conclusions

-