International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2015): 6.391

Study on the Properties of Penalized Logistic Regression with Adaptive Elastic Net

Qiang Hua[1], Shi Peng Zhao[2], Shaojing Lian[3]

College of Mathematics and Information Science, Hebei University, Baoding, China

Abstract: ***Logistic regression (LR) as an important data analysis method is widely used in various fields. In practical classification problem, the logistic regression always can receive a good effect. However, there are some obvious deficiencies in the traditional logistic regression model. The regularization method has been put forward. Nevertheless, some mainstream regularization logistic regression models are not "good" regularization methods for the lack of the Oracle property in theory. As a result, there is a certain degree of uncertainty when these methods are used. Based on this, the adaptive elastic net logistic regression (AEN-LR) is proposed. In this article we specially focus on the grouped selection property and the Oracle property of adaptive elastic net along with its model selection complexity. And the proofs of the properties are given. Through the detailed theoretical derivation, the reliable of the model can be guarantee essentially.***

Keywords: Logistic regression, Regularization, Adaptive elastic net, Oracle properties

1. Introduction

Formally, let $\mathbf{x} \subset R^{d \times n}$ denote the instance space and Y the set of class labels. The task of classification is to learn a function $f : \mathbf{x} \to \mathbf{y}$ from a given data set $\mathrm{D}=\left\{\left(\mathbf{x}_i, \mathbf{y}_i\right)\right\}_{i=1}^{n}$,where $\mathbf{x}_i \in \mathrm{R}^d$ is a training instance and $\mathbf{y}_i \in \{0,1\}$ is the associated binary class label. In order to solve this problem, lots of machine learning techniques have been proposed for classification problems, in which LR can be regarded as a popular one due to its well-established theoretical foundation and its good interpretability. The logistic regression model can be expressed as the following minimum optimization problem:

$$\min l(\boldsymbol{\beta})=-\sum_{i=1}^{n} \mathbf{y}_i \log\left(h_{\boldsymbol{\beta}}(\mathbf{x}_i)\right)+\left(1-\mathbf{y}_i\right)\log\left(1-h_{\boldsymbol{\beta}}(\mathbf{x}_i)\right) \quad (1)$$

Where $\boldsymbol{\beta}=[\beta_1, \beta_2, \ldots, \beta_{\mathrm{d}}]^T$ is regression coefficient vector, and $h_{\boldsymbol{\beta}}(\mathbf{x}_i)=\frac{1}{1+\exp(-\mathbf{x}_i{}^T\boldsymbol{\beta})}$ is the sigmoid function.

In practical classification problem, the logistic regression always can receive a good effect. However, there are some obvious deficiencies in the traditional logistic regression model. For this, the regularization method[1] has been put forward.. The ridge regression proposed by Hoerl and Kennard[2]is a continuous process that shrinks coefficients and hence makes the model stable, however, it does not set any coefficients to 0 and hence does not give an sparse model. In 1996, a promising technique named LASSO was proposed by Tibshirani[11]. It shrinks some coefficients and sets others to 0, and hence makes it retain the good features of the ridge regression. The LASSO has injected great vitality for the area of variable selection, especially when Least angle regression (LARS) algorithm was proposed by Efron[3], which has well solved the computational problems of LASSO. The LASSO method is not only applied to the simple linear regression model, but also other models, such as Generalized linear model (GLM)[10]. Park and Hastie[10]studied the coefficient estimate method with L1 planning in GLM, they adopted predict-correct algorithm to estimate coefficients meanwhile proceeding variable selection. As analyzed by Zou and Hastie [6], although LASSO shows good properties in many cases, it also has some limitations in the following cases:

a) When using LARS algorithm, for $n \times d$ design matrix, the LASSO can select at most $min(n, d)$ variables. So when $d > n$, LASSO can only select n variables, this is a limiting feature for a variable selection method.

b) If there is a group of variables among which the pair-wise correlations are very high, then the LASSO usually tends to select only one variable from the group and does not care which one is selected.

Zou and Hastie [6]proposed the elastic net penalty which is based on a combined penalty of LASSO and ridge regression penalty in order to overcome the drawbacks of using them on their own. Elastic net often performs better than LASSO in terms of prediction error when there is correlation among variables. Tutz and Ulbricht[5]proposed correlation-based penalty to deal with grouping effects. This penalty just makes variable shrinkage rather than variable selection. Elastic net penalty lacks consistent variable selection (oracle property) even though it outperforms LASSO. Zou and Zhang[8] proposed adaptive elastic net to handle grouping effects and enjoying oracle property simultaneously. El Anbari and Mkhadri [9] explained though experimental studies that elastic net seems to be slightly less reliable if the correlation between explanatory variables is not so extreme (i.e. $\rho > 0.95$).

In this paper, we study the adaptive elastic net logistic regression model. And we focus on the Oracle property and the grouped selection property of the proposed model. The rest of the paper is organized as follows. In Section II, we propose the adaptive elastic net logistic regression. In Section III, we prove that it has the Oracle property and the grouping effect property. In Section IV, we give a discussion.

2. The Adaptive Elastic Net Logistic Regression

The adaptively weighted L1 penalty and the Elastic-Net penalty improve the lasso in two different directions. The adaptive lasso achieves the oracle property of the SCAD and the Elastic-Net handles the collinearity. However, following the arguments in [6,7], we can easily see that the adaptive lasso inherits the instability of the lasso for high-dimensional data, while the Elastic-Net is lack of the oracle property. Thus, it is natural to consider combining the ideas of the adaptively weighted L1 penalty and the Elastic-Net regularization to obtain a better method which can improve the lasso in both directions.

The adaptive Elastic-Net can be viewed as a combination of the L2-norm and the adaptive lasso. Suppose we first compute the Elastic-Net estimator $\hat{\beta}(\text{logistic enet})$ as defined in (6), and then we construct the adaptive weights by

$$w_j = \left|\hat{\beta}_j^*\right|^{-\gamma},\ j = 1,2,\cdots,d,$$

where γ is a positive constant[8]. Now we solve the following optimization problem to get the adaptive Elastic-Net estimates:

$$\hat{\beta}^{**} = \underset{\beta}{\operatorname{argmin}} \left\{ l(\boldsymbol{\beta}) + \lambda_1^* \sum_{j=1}^{d} w_j \left|\beta_j\right| + \lambda_2 \sum_{j=1}^{d} \beta_j^2 \right\} \quad (2)$$

From now on, we write $\hat{\beta} = \hat{\beta}^{**}$ for the sake of convenience.

If we force λ_2 to be zero in (2), then the adaptive Elastic-Net reduces to the adaptive lasso. The role of the L2 penalty in (2) is to further regularize the adaptive lasso fit whenever the collinearity may cause serious trouble.

We know the Elastic-Net naturally adopts a sparse representation. One can use $w_j = \left(\left|\hat{\beta}_j^*\right| + 1/n\right)^{-\gamma}$ to avoid dividing zeros (Zou & Zhang (2009)). Let active set $\mathrm{A} = \left\{ j : \hat{\beta}_j^* \neq 0 \right\}$ and denotes its complement set. Then we have $\hat{\beta}_{\mathrm{A}_C} = 0$ and

$$\hat{\beta}_{\mathrm{A}} = \underset{\beta}{\operatorname{argmin}} \left\{ l(\boldsymbol{\beta}) + \lambda_1^* \sum_{j\in \mathrm{A}}^{p} w_j \left|\beta_j\right| + \lambda_2 \sum_{j\in \mathrm{A}}^{p} \beta_j^2 \right\} \quad (3)$$

where $\boldsymbol{\beta}$ in (3) is a vector of length $|\mathrm{A}| = d_0$, namely, the size of A is p_0. Then $|\mathrm{A}_C| = d - d_0\ (p > p_0)$.

The L1 regularization parameters, λ_1^* and λ_1, are directly responsible for the sparsity of the estimates. Their values are allowed to be different. On the other hand, we use the same λ_2 for the L2 penalty component in the Elastic-Net and the adaptive Elastic-Net estimators, because the L2 penalty offers the same kind of contribution in both estimators.

3. The Property of Aen-Lr

Consider loss function based on n samples:

$$\min_{\boldsymbol{\beta}} \left\{ l(\boldsymbol{\beta}) + \lambda_1^* \sum_{j=1}^{d} w_j \left|\beta_j\right| + \lambda_2 \sum_{j=1}^{d} \beta_j^2 \right\} \quad (4)$$

Definition 1 For given data (X, Y) and penalty parameter (λ_1, λ_2), the logistic adaptive elastic net estimate in the logistic regression model is defined as

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin} \left\{ l(\boldsymbol{\beta}) + \lambda_1^* \sum_{j=1}^{d} w_j \left|\beta_j\right| + \lambda_2 \sum_{j=1}^{d} \beta_j^2 \right\} \quad (5)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)$ are regression coefficients, nonnegative λ_1^* and λ_2 are tuning parameters.

3.1 Oracle properties

In this section we briefly review that with the proper choice of regularization parameters (λ_1^*, λ_2), the adaptive elastic net enjoys oracle properties.

In the absence of generalized loss, the true parameter $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02}, \cdots, \beta_{0d})^T = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T$, where $\boldsymbol{\beta}_1 \neq 0$ is the first d_0 components of $\boldsymbol{\beta}$, $\boldsymbol{\beta}_2 = 0$ is the remaining $d - d_0$ components of $\boldsymbol{\beta}$. Accordingly, $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)$. For the Fisher information matrix $\mathrm{I}(\boldsymbol{\beta})$, $\mathrm{I}_1(\boldsymbol{\beta}_1) = \mathrm{I}_1(\boldsymbol{\beta}_1, 0)$, where $\mathrm{I}_1(\boldsymbol{\beta}_1)$ is a $d_0 \times d_0$ matrix, and is the subset of $\mathrm{I}(\boldsymbol{\beta})$, while $\boldsymbol{\beta}_2 = 0$.

We assume that the following conditions are established:

(A1)The Fisher information matrix of logistic regression $\mathrm{I}(\boldsymbol{\beta}) = E\left[\dfrac{\exp(\mathbf{x}^T\boldsymbol{\beta})}{\left(1+\exp(\mathbf{x}^T\boldsymbol{\beta})\right)^2} \mathbf{x}\mathbf{x}^T \right]$ is a finite and positive definite matrix.

(A2) There exists an open subset Ω that contains the true parameter point $\boldsymbol{\beta}_0$. For all $\boldsymbol{\beta} \in \Omega$, here exist functions M such that

$$\left| \frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta})\left(1-\exp(\mathbf{x}_i^T\boldsymbol{\beta})\right)}{\left(1+\exp(\mathbf{x}_i^T\boldsymbol{\beta})\right)^3} \right| \leq M(\mathbf{x}_i^T) < \infty$$

And for any p dimensional vector $\mathbf{u}$, it exists $E\left[M(x_i^T) \left| x_i^T u \right|^3 \right] < \infty (1 \leq i \leq n)$.

(A3) There is a sequence $\{a_n\}$,so that $a_n \to \infty$, $a_n\left(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0\right) = O_p(1)$,and $\lambda_1 = o\left(\sqrt{n}\right)$, $a_n^{\gamma} \dfrac{\lambda_1^*}{\sqrt{n}} \to \infty$.

(A4) $\dfrac{\lambda_2}{\sqrt{n}} \sqrt{\sum_{j\in \mathrm{A}} \beta_j^2} \to 0$

Theorem 1 (Oracle properties) Suppose that $\lambda_1^* / \sqrt{n} \to 0$, $\lambda_1^* n^{(\gamma-1)/2} \to \infty$ and $\lambda_2 \to 0$. Under conditions (A1)-(A3), the adaptive elastic net has the oracle property, that is, the estimate $\hat{\boldsymbol{\beta}}$ must satisfy the following:

1. Asymptotic normality: $\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{\mathrm{d}} N\left(\mathbf{0}, I_1^{-1}(\beta_1)\right)$.
2. Sparsity: $\lim_{n\to\infty} P(\hat{\boldsymbol{\beta}}_2 = \mathbf{0}) = 1$.

Proof: Let $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \frac{\mathbf{u}}{\sqrt{n}}$.

Define

$$\Psi^{(n)}(\mathbf{u}) = l(\boldsymbol{\beta}_0 + \frac{\mathbf{u}}{\sqrt{n}}) + \lambda_1^{(n)*}\sum_{j=1}^{p} w_j \left|\boldsymbol{\beta}_{0j} + \frac{u_j}{\sqrt{n}}\right| + \lambda_2^{(n)}\sum_{j=1}^{p}\left(\boldsymbol{\beta}_{0j} + \frac{u_j}{\sqrt{n}}\right)^2$$

Let $\hat{\mathbf{u}} = \arg\min_{\mathbf{u}} \Psi^{(n)}(\mathbf{u})$, then $\hat{\mathbf{u}} = \sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$.

Using the Taylor expansion ,we have

$$\begin{aligned}\Psi^{(n)}(\mathbf{u}) - \Psi^{(n)}(\mathbf{0}) &= l(\boldsymbol{\beta}_0 + \frac{\mathbf{u}}{\sqrt{n}}) - l(\boldsymbol{\beta}_0) \\ &+ \lambda_1^{(n)*}\sum_{j=1}^{p} w_j\left(\left|\boldsymbol{\beta}_{0j} + \frac{u_j}{\sqrt{n}}\right| - \left|\boldsymbol{\beta}_{0j}\right|\right) \\ &+ \lambda_2^{(n)}\sum_{j=1}^{p}\left[\left(\boldsymbol{\beta}_{0j} + \frac{u_j}{\sqrt{n}}\right)^2 - \left(\boldsymbol{\beta}_{0j}\right)^2\right] \\ &\equiv (\mathbf{I}) + (\mathbf{II}) + (\mathbf{III})\end{aligned}$$

where

$$\begin{aligned}(I) &= \frac{1}{\sqrt{n}}\nabla^T l(\boldsymbol{\beta}_0)\mathbf{u} + \frac{1}{2}\mathbf{u}^T\nabla^2 l(\boldsymbol{\beta}_0)\frac{\mathbf{u}}{n} \\ &+ \frac{1}{6}\nabla^T\left[\mathbf{u}^T\nabla^T l(\tilde{\boldsymbol{\beta}}_0)\mathbf{u}\right]\frac{\mathbf{u}}{n^{3/2}} \\ &= T_1 + T_2 + T_3\end{aligned}$$

$$(II) = \lambda_1^{(n)*}\sum_{j=1}^{p} w_j\left(\left|\beta_{0j} + \frac{u_j}{\sqrt{n}}\right| - \left|\beta_{0j}\right|\right) = \frac{\lambda_1^{(n)*}}{\sqrt{n}}\sum_{j=1}^{p} w_j\sqrt{n}\left(\left|\beta_{0j} + \frac{u_j}{\sqrt{n}}\right| - \left|\beta_{0j}\right|\right)$$

$$(III) = \lambda_2^{(n)}\sum_{j=1}^{p}\left[\left(\beta_{0j} + \frac{u_j}{\sqrt{n}}\right)^2 - \left(\beta_{0j}\right)^2\right] = \sum_{j=1}^{p} 2\left[\frac{\lambda_2^{(n)}}{\sqrt{n}}\beta_{0j}u_j + \frac{\lambda_2^{(n)}}{n}u_j^2 + o_p(1)\right]$$

with

$$T_1 = \sum_{i=1}^{n}\left[-y_i + \frac{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}{1+\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}\right]\frac{\mathbf{x}_i^T\mathbf{u}}{\sqrt{n}}$$

$$T_2 = \frac{1}{2}\sum_{i=1}^{n}\frac{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}{\left(1+\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)\right)^2}\mathbf{u}^T\frac{\left(\mathbf{x}_i\mathbf{x}_i^T\right)}{n}\mathbf{u}$$

$$T_3 = \frac{1}{6}\sum_{i=1}^{n}\frac{\exp\left(\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_0\right)\left(1-\exp\left(\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_0\right)\right)}{\left(1+\exp\left(\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_0\right)\right)^3}\frac{\left(\mathbf{x}_i^T\mathbf{u}\right)^3}{n^{3/2}},$$

$$\left(\boldsymbol{\beta}_0 < \tilde{\boldsymbol{\beta}}_0 < \boldsymbol{\beta}_0 + \frac{\mathbf{u}}{\sqrt{n}}\right)$$

We analyze the asymptotic limit of each part.
By the properties of the exponential family, then

$$\begin{aligned}&E_{y_i,\mathbf{x}_i}\left(\left[y_i - \frac{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}{1+\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}\right]\mathbf{x}_i^T\mathbf{u}\right) \\ &= E_{y_i,\mathbf{x}_i}\left(\left[y_i - \frac{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}{1+\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}\right]\right)\mathbf{x}_i^T\mathbf{u} \\ &= 0\end{aligned}$$

$$\begin{aligned}&Var_{y_i,\mathbf{x}_i}\left(\left[-y_i + \frac{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}{1+\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}\right]\mathbf{x}_i^T\mathbf{u}\right) \\ &= E_{\mathbf{x}_i}\left(\left[\frac{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}{\left(1+\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)\right)^2}\right]\left(\mathbf{x}_i^T\mathbf{u}\right)^2\right) \\ &= \mathbf{u}^T E_{\mathbf{x}_i}\left(\left[\frac{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}{\left(1+\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)\right)^2}\right]\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{u} \\ &= \mathbf{u}^T\mathbf{I}(\boldsymbol{\beta}_0)\mathbf{u}\end{aligned}$$

According to the central limit theorem, we have that

$$T_1 \xrightarrow{d} -\mathbf{u}^T N\left(0, \mathbf{I}(\boldsymbol{\beta}_0)\right) \quad (6)$$

For the second part T_2, we can observe than

$$\frac{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}{\left(1+\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)\right)^2}\frac{\mathbf{x}_i\mathbf{x}_i^T}{n} \xrightarrow{p} \mathbf{I}(\boldsymbol{\beta}_0)$$

Then ,

$$T_2 \xrightarrow{p} \frac{1}{2}\mathbf{u}^T\mathbf{I}(\boldsymbol{\beta}_0)\mathbf{u} \quad (7)$$

The condition (A2) says that

$$\begin{aligned}6\sqrt{n}\mathbf{I}_3 &= \sum_{i=1}^{n}\frac{\exp\left(\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_0\right)\left(1-\exp\left(\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_0\right)\right)}{\left(1+\exp\left(\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_0\right)\right)^3}\frac{\left(\mathbf{x}_i^T\mathbf{u}\right)^3}{n} \\ &\le \sum_{i=1}^{n}\frac{1}{n}M\left(\mathbf{x}_i^T\right)\left|\mathbf{x}_i^T\mathbf{u}\right|^3 \xrightarrow{P} E\left[M\left(\mathbf{x}_i^T\right)\left|\mathbf{x}_i^T\mathbf{u}\right|^3\right] \quad (8) \\ &< \infty\end{aligned}$$

Now consider the limiting behavior of the (II).

If $\beta_{0j} \neq 0$, then $w_j \xrightarrow{p} \left|\beta_{0j}\right|^{-\gamma}$ and $\sqrt{n}\left(\left|\beta_{0j} + \frac{u_j}{\sqrt{n}}\right| - \left|\beta_{0j}\right|\right) \to u_j\,\mathrm{sgn}\left(\beta_{0j}\right)$ .By the Slutsky theorem, we have $(II) \xrightarrow{p} 0$.

If $\beta_{0j} = 0$, then $\sqrt{n}\left(\left|\beta_{0j} + \frac{u_j}{\sqrt{n}}\right| - \left|\beta_{0j}\right|\right) \to \left|u_j\right|$,and $a_n\left(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0\right) = O_p(1)$, $\lambda_1^{*(n)} = o\left(\sqrt{n}\right)$, $a_n^{\gamma}\frac{\lambda_1}{\sqrt{n}} \to \infty$,then $\frac{\lambda_1^{(n)*}}{\sqrt{n}}w_j = \frac{\lambda_1^{(n)*}}{\sqrt{n}}a_n^{\gamma}\frac{\lambda_1^{(n)*}}{\left|a_n^{\gamma}\hat{\beta}_j^*\right|^{\gamma}} \to \infty$,hence $(II) \xrightarrow{p} \infty$

We summarize the results as follow:

$$(II) \xrightarrow{p} \begin{cases} 0, & if\ \beta_{0j} \neq 0 \\ 0, & if\ \beta_{0j} = 0, u_j = 0 \\ \infty, & if\ \beta_{0j} = 0, u_j \neq 0 \end{cases}.$$

Consider the third term. By the condition(A4),we have

$$(III) \xrightarrow{p} 0.$$

Thus, by the Slutsky's theorem, for every $\boldsymbol{u}$,we have

$$\Psi^{(n)}(\mathbf{u}) - \Psi^{(n)}(\mathbf{0}) \xrightarrow{d} \Psi(\mathbf{u})$$

$$= \begin{cases} \frac{1}{2}\mathbf{u}_{\mathrm{A}}^T \mathbf{I}(\boldsymbol{\beta}_0)\mathbf{u}_{\mathrm{A}} - \mathbf{u}_{\mathrm{A}}^T W_{\mathrm{A}}, & u_j = 0 \forall j \notin \mathrm{A} \\ \infty, & \text{otherwise} \end{cases}$$

Where $W_{\mathrm{A}} = N\left(0, \mathbf{I}_1(\boldsymbol{\beta}_{10})\right)$. $\Psi^{(n)}(\mathbf{u}) - \Psi^{(n)}(\mathbf{0})$ is convex and the unique minimum of $\Psi^{(n)}(\mathbf{u}) - \Psi^{(n)}(\mathbf{0})$ is $\mathbf{u}_{\mathrm{A}} = \mathbf{I}_1^{-1}(\boldsymbol{\beta}_{10}) W_{\mathrm{A}}$ .Then we have

$$\hat{\mathbf{u}}_{\mathrm{A}} \xrightarrow{d} \mathbf{I}_1^{-1}(\boldsymbol{\beta}_{10}) W_{\mathrm{A}}, \hat{\mathbf{u}}_{\mathrm{A}^C} \xrightarrow{d} 0$$

The asymptotic normality part is proven.

Now we show the second part of this theorem.

If $\lim_{n\to\infty} P(\hat{\boldsymbol{\beta}}_2 = \boldsymbol{0}) = 1$,then it satisfied that $\beta_{0k} = 0$, $P(\hat{\beta}_k \neq 0) \to 0$.

Consider $\beta_{0k} \neq 0$ .by the KKT conditions, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[x_{ik} \left(-y_i + \frac{\exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_0\right)}{1 + \exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_0\right)} \right) + 2\lambda_2^{(n)} \hat{\beta}_k \right] = \frac{\lambda_1^{*(n)}}{\sqrt{n}} w_k$$

$P(\hat{\beta}_k \neq 0)$

Then $\leq P\left(\begin{array}{l} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[x_{ik} \left(-y_i + \frac{\exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_0\right)}{1 + \exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_0\right)} \right) + 2\lambda_2^{(n)} \hat{\beta}_k \right] \\ = \frac{\lambda_1^{*(n)}}{\sqrt{n}} w_k \end{array} \right)$

$$= P\left(\begin{array}{l} \frac{\sqrt{n}}{\left(\lambda_1^{*(n)} w_k\right)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[x_{ik} \left(\frac{\exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_0\right)}{1 + \exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_0\right)} - y_i \right) + 2\lambda_2^{(n)} \hat{\beta}_k \right] \\ = 1 \end{array} \right)$$

Note that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[x_{ik} \left(-y_i + \frac{\exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_0\right)}{1 + \exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_0\right)} \right) + 2\lambda_2^{(n)} \hat{\beta}_k \right],$$

$$= K_1 + K_2 + K_3 + K_4$$

with

$$K_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ik} \left(-y_i + \frac{\exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_0\right)}{1 + \exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_0\right)} \right)$$

$$K_2 = \left(\frac{1}{n} \sum_{i=1}^{n} x_{ik} \frac{\exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_0\right)}{\left(1 + \exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_0\right)\right)^2} \mathbf{x}_i^T \right) \sqrt{n}\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_0\right)$$

$$K_3 = \frac{1}{n^{3/2}} \sum_{i=1}^{n} x_{ik} \frac{\exp\left(\mathbf{x}_i^T \tilde{\boldsymbol{\beta}}_{0*}\right)\left(1 - \exp\left(\mathbf{x}_i^T \tilde{\boldsymbol{\beta}}_0\right)\right)}{\left(1 + \exp\left(\mathbf{x}_i^T \tilde{\boldsymbol{\beta}}_0\right)\right)^3} \left(\mathbf{x}_i^T \left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_0\right)\right)^2,$$

$$\left(\hat{\boldsymbol{\beta}}_0 < \tilde{\boldsymbol{\beta}}_0 < \boldsymbol{\beta}_0\right)$$

and

$$K_4 = 2\frac{\lambda_2^{(n)} \hat{\beta}_k}{\sqrt{n}}.$$

By the previous arguments, we have that

$$K_1 \xrightarrow{d} N\left(0, \mathbf{I}(\boldsymbol{\beta}_0)\right).$$

Note that $\frac{1}{n} \sum_{i=1}^{n} x_{ik} \frac{\exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_0\right)}{\left(1 + \exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_0\right)\right)^2} \mathbf{x}_i^T \xrightarrow{p} \mathbf{I}_k$,where $\boldsymbol{I}_k$ is the kth row of $\boldsymbol{I}$.

Thus (7)implies that K_2 converges to some normal random variable.

It follows the condition(A2), (A4)and (8) that $K_3 \to 0$, $K_4 \to 0$.

Meanwhile, we have $\frac{1}{\left(\lambda_1^{*(n)} w_k\right) / \sqrt{n}} \xrightarrow{p} 0$.

Thus, $\lim_{n\to\infty} P(\hat{\boldsymbol{\beta}}_2 = \boldsymbol{0}) = 1$ .The theorem is proved.

3.2 The grouping effect

The logistic adaptive elastic net has the grouping effect which means that the strongly correlated predictors tend to be in or out of the model together, it is shown by the following theorem.

Theorem 2 (the grouping effect) For logistic regression model, given data (X, Y) and parameter (λ_1, λ_2) , design matrix X has been standardized. Let $\hat{\beta}(\lambda_1, \lambda_2)$ present logistic adaptive elastic net estimate. Assume $\hat{\beta}_k(\lambda_1, \lambda_2)\hat{\beta}_l(\lambda_1, \lambda_2) > 0$. Define $D_{\lambda_1, \lambda_2}(k, l) = |\hat{\beta}_k(\lambda_1, \lambda_2) - \hat{\beta}_l(\lambda_1, \lambda_2)|$, Then

$$D_{\lambda_1, \lambda_2}(k, l) \leq \frac{1}{\lambda_2} \left[\left| \sum_{i=1}^{n} (x_{il} - x_{ik}) \hat{r}_i \right| + \left| \frac{\lambda_1}{2} (w_l - w_k) \right| \right]$$

If we assume that $w_m = \left|\hat{\beta}_m^o\right|^{-\gamma}$ is a consistent estimator of β_m , $\gamma > 0$.

Then

$$D_{\lambda_1, \lambda_2}(k, l) \leq \frac{1}{\lambda_2} \left[\left| \sum_{i=1}^{n} (x_{il} - x_{ik}) \hat{r}_i \right| + \frac{\gamma \lambda_1}{2 \min\left(\left|\hat{\beta}_l^o\right|, \left|\hat{\beta}_k^o\right|\right)^{\gamma+1}} \left| \hat{\beta}_l^o - \hat{\beta}_k^o \right| \right]$$

where $\hat{r}_i = y_i - \frac{\exp\left(x_i^T \hat{\beta}\right)}{1 + \exp\left(x_i^T \hat{\beta}\right)}$ is prediction residual.

Proof: If $\hat{\beta}_k(\lambda_1, \lambda_2)\hat{\beta}_l(\lambda_1, \lambda_2) > 0$,then $\hat{\beta}_k(\lambda_1, \lambda_2) \neq 0$, $\hat{\beta}_l(\lambda_1, \lambda_2) \neq 0$, and $\operatorname{sgn}\left\{\hat{\beta}_k(\lambda_1, \lambda_2)\right\} = \operatorname{sgn}\left\{\hat{\beta}_l(\lambda_1, \lambda_2)\right\}$.

If $\hat{\beta}_m(\lambda_1, \lambda_2) \neq 0$, $\hat{\beta}(\lambda_1, \lambda_2)$, it must satisfy

$$\frac{\partial L(\lambda_1, \lambda_2, \beta)}{\partial \beta_m}\Big|_{\beta = \hat{\beta}(\lambda_1, \lambda_2)} = 0.$$

Then we can have

$$\sum_{i=1}^{n} \left[-y_i x_{ik} + \frac{\exp\left(x_i^T \hat{\beta}\right)}{1 + \exp\left(x_i^T \hat{\beta}\right)} x_{ik} \right] + \lambda_1 w_l \operatorname{sgn}\left\{\hat{\beta}_k(\lambda_1, \lambda_2)\right\} + 2\lambda_2 \hat{\beta}_k(\lambda_1, \lambda_2) = 0 \quad (9)$$

$$\sum_{i=1}^{n}\left[-y_i x_{il}+\frac{\exp\left(x_i^T\hat{\beta}\right)}{1+\exp\left(x_i^T\hat{\beta}\right)}x_{il}\right] \quad (10)$$
$$+\lambda_1 w_k \operatorname{sgn}\left\{\hat{\beta}_l(\lambda_1,\lambda_2)\right\}+2\lambda_2\hat{\beta}_l(\lambda_1,\lambda_2)=0$$

According to Eqs.(9) and (10), we get

$$2\lambda_2\left[\hat{\beta}_k(\lambda_1,\lambda_2)-\hat{\beta}_l(\lambda_1,\lambda_2)\right]$$
$$=\sum_{i=1}^{n}\left[y_i\left(x_{il}-x_{ik}\right)-\frac{\exp\left(x_i^T\hat{\beta}\right)}{1+\exp\left(x_i^T\hat{\beta}\right)}\left(x_{il}-x_{ik}\right)\right]$$
$$-\lambda_1\left(w_l-w_k\right)\operatorname{sgn}\left\{\hat{\beta}_k(\lambda_1,\lambda_2)\right\}$$

which implies

$$\lambda_2\left|\hat{\beta}_k(\lambda_1,\lambda_2)-\hat{\beta}_l(\lambda_1,\lambda_2)\right|$$
$$=\left|\sum_{i=1}^{n}\left(x_{il}-x_{ik}\right)\hat{r}_i-\frac{\lambda_1}{2}\left(w_l-w_k\right)\operatorname{sgn}\left\{\hat{\beta}_k(\lambda_1,\lambda_2)\right\}\right|$$

where $\hat{r}_i=y_i-\frac{\exp\left(x_i^T\hat{\beta}\right)}{1+\exp\left(x_i^T\hat{\beta}\right)}$ is prediction residual.

Note the identity $|A\pm B|\le|A|+|B|$ holds true for all A and B .Applying this to the right hand side of the above equation

$$D_{\lambda_1,\lambda_2}(k,\ l)=|\hat{\beta}_k(\lambda_1,\lambda_2)-\hat{\beta}_l(\lambda_1,\lambda_2)|$$
$$\le\frac{1}{\lambda_2}\left[\left|\sum_{i=1}^{n}\left(x_{il}-x_{ik}\right)\hat{r}_i\right|+\left|\frac{\lambda_1}{2}\left(w_l-w_k\right)\right|\right]$$

To prove the second part lets assume $w_m=\left|\hat{\beta}_m^o\right|^{-\gamma}$ for some $\gamma>0$.

By applying mean value theorem to the function $f(x)=x^{-\gamma}$, we have $\left|f(x)-f(y)\right|=\left|f'(c)\right||x-y|$ for some $c\in[x,y]$. Hence,

$$\left|w_l-w_k\right|=\left|\left|\hat{\beta}_l^o\right|^{-\gamma}-\left|\hat{\beta}_k^o\right|^{-\gamma}\right|=\frac{\gamma}{c^{\gamma+1}}\left|\left|\hat{\beta}_l^o\right|-\left|\hat{\beta}_k^o\right|\right|$$
$$\le\frac{\gamma}{\min\left(\left|\hat{\beta}_l^o\right|,\left|\hat{\beta}_k^o\right|\right)^{\gamma+1}}\left|\hat{\beta}_l^o-\hat{\beta}_k^o\right|$$

$$D_{\lambda_1,\lambda_2}(k,\ l)\le$$

Then $$\frac{1}{\lambda_2}\left[\left|\sum_{i=1}^{n}\left(x_{il}-x_{ik}\right)\hat{r}_i\right|+\frac{\gamma\lambda_1}{2\min\left(\left|\hat{\beta}_l^o\right|,\left|\hat{\beta}_k^o\right|\right)^{\gamma+1}}\left|\hat{\beta}_l^o-\hat{\beta}_k^o\right|\right]$$

$D_{\lambda_1,\lambda_2}(k,\ l)$ describes the difference of coefficient path between the variables k and l . if x_k and x_l are highly correlated, Theorem 1 shows that the difference of coefficient path between the variables k and l is almost equal to 0, which demonstrates when there exists strong relationship between variables, the logistic adaptive elastic net can select all of them.

Another important observation from the above bound is that the grouping effect has contributions not only form L2 penalty but also from L1 type adaptive penalty. However, if $\lambda_2\to 0$, the above bound becomes trivial. Hence it is not possible to capture any grouping effect by only adaptive lasso type penalty.

4. Conclusion

In this article we proposed a kind of adaptive elastic net logistic regression which carry with oracle property along with other desirable properties of lasso and elastic net together. The adaptive elastic net logistic regression produces a sparse solution and enjoys the grouped selection property of elastic net. However, this comes at an increasing cost of model complexity.

5. Acknowledgements

This paper is supported by the ResearchFundof Hebei University and the Foundation of Research and Practice of Teaching Reform in Higher Education of Hebei province(2016GJJG012)and the Foundation of Hebei Educational Committee (QN2017019)

References

[1] Tikhonov. Solution of incorrectly formulated problems and the regularization method[C],Soviet Math. Dokl., 5: 1035-1038, (1963).
[2] A.E. Hoerl and R.W. Kennard, Ridge regression: Biased estimation for nonorthogonal problem[J], Technometrics, (12):55–67, (1970).
[3] Efron, T. Hastie, I. Johnstone and R. Tibshirani, Least angle regression[J], The Annals of Statistics, 32(2):407–499, (2004).
[4] X. Zhou. On grouping effect of elastic net[J]. Statistics & Probability Letters,83(9):2108-2112,(2013).
[5] G. Tutz and Ulbricht.Penalized regression with correlation-based penalty [J].Statistics and Computing, 19(3):239-253,(2009).
[6] H. Zou and T. Hastie, Regularization and variable selection via the elastic net[J], the Royal Statistical Society, Series B,67:(2),301–320, (2005).
[7] H. Zou. The adaptive lasso and its oracle properties[J], Journal of the American Statistical Association 101, 1418–1429,(2006).
[8] H. Zou and Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters. Annals of Statistics, 37(4):1733-1751.
[9] M. El Anbari, and A. Mkhadri, Penalized regression combining the l1 norm and a correlation based penalty[J]. Sankhya B, 76(1):82-102. (2014).
[10] M.Y. Park and T. Hastie, L1-regularization-path algorithm for generalized linear models[J], Journal of the Royal Statistical Society, Series B, 69(4):659–677, (2007).
[11] R. Tibshirani, Regression shrinkage and selection via the LASSO, Journal of the Royal Statistical Society, Series B,58(1):267–288, (1996)