



Elastic-Net and algorithms for computing the regularization paths

Zou, Hastie, etc
Presenter: Zhiting Xu

May 6, 2010

Outline

- 1 Motivation
- 2 Elastic Net
 - Naive Elastic Net
 - Elastic Net
 - Adaptive Elastic Net
- 3 Computation



Limitations of Lasso

- In the $p > n$ case, the lasso selects at most n variables before it saturates
- If there is a groups of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected
- For usual $n > p$ situations, if there are high correlations between predictors, it has been empirically observed that the prediction performance of the lasso is dominated by the ridge regression



Naive Elastic Net

Definition

Suppose that the data set has n observations with p predictors. Let $\mathbf{y} = (y_1, \dots, y_n)^T$ be the response and $\mathbf{X} = (\mathbf{x}_1 | \dots | \mathbf{x}_p)$ be the model matrix, where $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})^T$, $j = 1 \dots, p$, are the predictors. Assume the response is centered and the predictors are standardized. For any fixed non-negative λ_1 and λ_2 , we define the naive elastic net criterion:

$$L(\lambda_1, \lambda_2, \beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1 \quad (1)$$

The naive elastic net estimator $\hat{\beta}$ is the minimizer of equation 1:

$$\hat{\beta} = \arg \min_{\beta} \{L(\lambda_1, \lambda_2, \beta)\} \quad (2)$$





Solution

Lemma

Given data set (\mathbf{y}, \mathbf{X}) and (λ_1, λ_2) , define an artificial data set $(\mathbf{y}^*, \mathbf{X}^*)$ by

$$\mathbf{X}_{(n+p) \times p}^* = (1 + \lambda_2)^{-1/2} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2} \mathbf{I} \end{pmatrix}, \mathbf{y}_{(n+p)}^* = \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix}. \quad (3)$$

Let $\gamma = \lambda_1 / \sqrt{1 + \lambda_2}$ and $\beta^* = \sqrt{1 + \lambda_2} \beta$. Then the naive elastic net criterion can be written as

$$L(\gamma, \beta) = L(\gamma, \beta^*) = \|\mathbf{y}^* - \mathbf{X}^* \beta^*\|^2 + \gamma \|\beta^*\|_1 \quad (4)$$

Let $\hat{\beta}^* = \arg \min_{\beta^*} L\{(\gamma, \beta^*)\}$, then $\hat{\beta} = \frac{1}{\sqrt{1 + \lambda_2}} \hat{\beta}^*$





The grouping effect

Theorem

Given data (\mathbf{y}, \mathbf{X}) and parameters (λ_1, λ_2) , the response \mathbf{y} is centered and the predictors \mathbf{X} are standardized. Let $\hat{\beta}(\lambda_1, \lambda_2)$ be the naive elastic net estimate. Suppose that $\hat{\beta}_i(\lambda_1, \lambda_2)\hat{\beta}_j(\lambda_1, \lambda_2) > 0$. Define

$$D_{\lambda_1, \lambda_2}(i, j) = \frac{1}{|\mathbf{y}|_1} |\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)|$$

then

$$D_{\lambda_1, \lambda_2}(i, j) \leq \frac{1}{\lambda_2} \sqrt{2(1 - \rho)}$$

where $\rho = \mathbf{x}_i^T \mathbf{x}_j$, the sample correlation.





The Elastic Net Estimate

- Given data (\mathbf{y}, \mathbf{X}) , penalty parameter (λ_1, λ_2) and augmented data $(\mathbf{y}^*, \mathbf{X}^*)$, the naive elastic net solves a lasso-type problem

$$\hat{\beta}^* = \arg \min_{\beta^*} |\mathbf{y}^* - \mathbf{X}^* \beta^*|^2 + \frac{\lambda_1}{\sqrt{1 + \lambda_2}} |\beta^*|_1$$

- The elastic net(corrected) estimates $\hat{\beta}$ are defined by

$$\hat{\beta}(\text{elastic net}) = \sqrt{1 + \lambda_2} \hat{\beta}^*$$

- Recall that $\hat{\beta}(\text{naive elastic net}) = 1 / \sqrt{1 + \lambda_2} \hat{\beta}^*$, thus

$$\hat{\beta}(\text{elastic net}) = (1 + \lambda_2) \hat{\beta}(\text{naive elastic net})$$



Theorem

Given data (\mathbf{y}, \mathbf{X}) and (λ_1, λ_2) , then the elastic net estimates $\hat{\beta}$ are given by

$$\hat{\beta} = \arg \min_{\beta} \beta^T \left(\frac{\mathbf{X}^T \mathbf{X} + \lambda_2 \mathbf{I}}{1 + \lambda_2} \right) \beta - 2\mathbf{y}^T \mathbf{X} \beta + \lambda_1 |\beta|_1 \quad (5)$$

It is easy to see that

$$\hat{\beta}(\text{lasso}) = \arg \min_{\beta} \beta^T (\mathbf{X}^T \mathbf{X}) \beta - 2\mathbf{y}^T \mathbf{X} \beta + \lambda_1 |\beta|_1$$

The theorem interprets the elastic net as a stabilized version of the lasso.



Adaptive Elastic Net

Definition

Suppose we first compute the elastic-net estimator $\hat{\beta}(\text{enet})$, and then we construct the adaptive weights by

$$\hat{w}_j = (|\hat{\beta}_j(\text{enet})|)^{-\gamma}, j = 1, 2, \dots, p$$

where γ is a positive constant. We solve the following optimization problem to get the adaptive elastic-net estimates

$$\hat{\beta}(\text{AdaEnet}) = \left(1 + \frac{\lambda_2}{n}\right) \left\{ \arg \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\| + \lambda_2 \|\beta\|_2^2 + \lambda_1^* \sum_{j=1}^p \hat{w}_j |\beta_j| \right\} \quad (6)$$



Let $\hat{\mathcal{A}}_{\text{enet}} = \{j : \hat{\beta}_j(\text{enet}) \neq 0\}$ and $\hat{\mathcal{A}}_{\text{enet}}^c$ denotes its complement set. Then we have $\hat{\beta}_{\hat{\mathcal{A}}_{\text{enet}}^c} = 0$, and

$$\hat{\beta}_{\hat{\mathcal{A}}_{\text{enet}}} = \left(1 + \frac{\lambda_2}{n}\right) \left\{ \arg \min_{\beta} \|\mathbf{y} - \mathbf{X}_{\hat{\mathcal{A}}_{\text{enet}}} \beta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1^* \sum_{j \in \hat{\mathcal{A}}_{\text{enet}}} \hat{w}_j |\beta_j| \right\} \quad (7)$$

where β is a vector of length $|\hat{\mathcal{A}}_{\text{enet}}|$



Regularity conditions

- A1** We use $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ to denote the minimum and maximum eigenvalues of a positive definite matrix \mathbf{M} . Then we assume

$$b \leq \lambda_{\min}\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right) \leq \lambda_{\max}\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right) \leq B$$

where b and B are two positive constants

- A2** $\lim_{n \rightarrow \infty} \frac{\max_{i=1,2,\dots,n} \sum_{j=1}^p x_{ij}^2}{n} = 0$
- A3** $E[|\epsilon|^{2+\delta}] < \infty$ for some $\delta > 0$



A4 $\lim_{n \rightarrow \infty} \frac{\log p}{\log n} = \nu$ for some $0 \leq \nu < 1$. To construct the adaptive weights \hat{w} , we take a fixed γ such that $\gamma > \frac{2\nu}{1-\nu}$.

A5

$$\lim_{n \rightarrow \infty} \frac{\lambda_2}{n} = 0, \lim_{n \rightarrow \infty} \frac{\lambda_1}{\sqrt{n}} = 0$$

and

$$\lim_{n \rightarrow \infty} \frac{\lambda_1^*}{\sqrt{n}} = 0, \lim_{n \rightarrow \infty} \frac{\lambda_1^*}{\sqrt{n}} n^{((1-\nu)(1+\gamma)-1)/2} = \infty$$

A6

$$\lim_{n \rightarrow \infty} \frac{\lambda_2}{\sqrt{n}} \sqrt{\sum_{j \in \mathcal{A}} \beta_j^{*2}} = 0$$

$$\lim_{n \rightarrow \infty} \min\left(\frac{n}{\lambda_1 \sqrt{p}}, \left(\frac{\sqrt{n}}{\sqrt{p} \lambda_1^*}\right)^{1/\gamma} (\min_{j \in \mathcal{A}} |\beta_j^*|)\right) \rightarrow \infty$$



Theorem

Given the data (\mathbf{y}, \mathbf{X}) , let $\hat{\mathbf{w}} = (\hat{w}_1, \dots, \hat{w}_p)$ be a vector whose components are all nonnegative and can depend on (\mathbf{y}, \mathbf{X}) . Define

$$\hat{\beta}_{\hat{\mathbf{w}}} = \arg \min_{\beta} \{ \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \sum_{j=1}^p \hat{w}_j |\beta_j| \}$$

for nonnegative parameter λ_2 and λ_1 .

If we assume the condition A1, then

$$E(\|\hat{\beta}_{\hat{\mathbf{w}}}(\lambda_2, \lambda_1) - \beta^*\|_2^2) \leq 4 \frac{\lambda_2^2 \|\beta^*\|_2^2 + Bpn\sigma^2 + \lambda_1^2 E(\sum_{j=1}^p \hat{w}_j^2)}{(bn + \lambda_2)^2}$$



Theorem

Let us write $\beta^* = (\beta_{\mathcal{A}}^*, 0)$, and define

$$\tilde{\beta}_{\mathcal{A}}^* = \arg \min_{\beta} \{ \|\mathbf{y} - \mathbf{X}_{\mathcal{A}}\beta\|_2^2 + \lambda_2 \sum_{j \in \mathcal{A}} \beta_j^2 + \lambda_1^* \sum_{j \in \mathcal{A}} \hat{w}_j |\beta_j| \}$$

Then with probability tending to 1, $((1 + \frac{\lambda_2}{n} \tilde{\beta}_{\mathcal{A}}^*), 0)$ is solution to (6).

The definition of $\tilde{\beta}_{\mathcal{A}}^*$ borrows the concept of "oracle". If there was an oracle informing us the true subset model, then we would use this oracle information and the adaptive elastic-net criterion would become that in 7.



Theorem

Under conditions (A1)-(A6), the adaptive elastic-net has the oracle property; that is, the estimator $\hat{\beta}(\text{AdaEnet})$ must satisfy:

- *Consistency in selection: $Pr(\{j : \hat{\beta}(\text{AdaEnet})_j \neq 0\} = \mathcal{A}) \rightarrow 1$*
- *Asymptotic normality:*

$$\alpha^T \frac{\mathbf{I} + \lambda_2 \Sigma_{\mathcal{A}}^{-1}}{1 + \lambda_2 / \bar{n}} \Sigma_{\mathcal{A}}^{1/2} (\hat{\beta}(\text{AdaEnet})_{\mathcal{A}} - \beta_{\mathcal{A}}^*) \rightarrow_d N(0, \sigma^2), \text{ where}$$

$$\Sigma_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}} \text{ and } \alpha \text{ is a vector of norm } 1.$$



Least Angle Regression

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}, \beta_1, \dots, \beta_p = 0$
2. Find the predictor \mathbf{x}_j most correlated with \mathbf{r}
3. Move β_j from 0 towards its least-squares coefficient $\langle \mathbf{x}_j, \mathbf{r} \rangle$, until some other competitor \mathbf{x}_k has as much correlation with the current residual as does \mathbf{x}_j
4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor \mathbf{x}_l has as much correlation with the current residual
5. Continue in this way until all p predictors have been entered. After $\min(N - 1, p)$ steps, we arrive at the full least-squares solution.



LAR: Lasso Modification




Let $\hat{\beta}$ be a Lasso solution, with $\hat{\mu} = \mathbf{X}\hat{\beta}$. Then it is easy to show that the sign of any non-zero coordinate $\hat{\beta}_j$ must agree with the sign s_j of the current correlation $\hat{c}_j = \mathbf{x}'_j(\mathbf{y} - \hat{\mu})$

Lasso Modification

4.a If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint least squares direction.



Reference

-  Efron, Bradley and Hastie, Trevor and Johnstone, Lain and Tibshirani, Robert. *Least angle regression*, Annals of Statistics, 2004.
-  Hui Zou and Trevor Hastie. *Regularization and variable selection via the elastic net*, Journal Of The Royal Statistical Society Series B, 2005.
-  Hui Zou and Hao Helen Zhang. *On the adaptive elastic-net with a diverging number of parameters*, Annals of Statistics, 2009.