Review: L_1 Regularization Path Algorithm for GLMs

Jieun Shin

2023-06-14

1. Introduction

이 논문은 볼록 최적화에서 predictor-corrector 방법을 사용하여 계산을 효율적으로 하는 path algorithm을 제안한다. 최적화는 1) λ에서 step-size를 결정하고 2) 현재시점의 변화를 예측하고 3) 이전의 예측에서 오차를 수정하는 과정으로 이루어진다.

전통적인 변수선택법인 forward or backward selection이 greedy한 방법인 것과는 달리, L_1정규화 방법은 보다 민주적이면서 smooth하다. 나아가 정규화 path는 변수가 enter 혹은 leave되는 순서를 알 수 있다. path 알고리즘으로 Efron (2004)는 LARS를 제안하였는데, 이는 forward stagewise와 least angle regression의 수정된 버전으로 exact piecewise linear한 계수 경로를 제공한다.

LARS 혹은 SVM path가 piece-wise linear인 것과 달리 GLM의 path는 piece-wise linear가 아니다. 따라서 계수가 정확히 계산되는 $lambda$를 선택해야 한다. 이 논문에서는 non-zero 계수의 변화의 정확한 추정을 제안한다.

Rosset (2004)는 any loss와 penalty에도 사용가능한 일반적인 경로알고리즘을 제안한다. 계수의 추정은 Newton iterative로 추정한다. Zhou and Yu (2004)는 boosted lasso를 제안하였다. 이는 any convex loss와 L_1정규화에서 사용 가능하며, 추정은 backward selection에 foward stagewise fitting을 결합한 방식으로 한다.

2. GLM path algorithm

p개의 factor와 response의 쌍이 n개 있다고 하자. 즉, $\{(\boldsymbol{x}_i, y_i) : \boldsymbol{x}_i \in \mathcal{R}^p, y_i \in \mathcal{R}, i = 1, \ldots, p\}$. Y는 평균 $\mu = E(Y)$, 분산 $V = Var(Y)$를 가지는 지수족 분포를 따른다고 하자. GLM은 random component Y와 systematic component η와 link function g가 다음의 관계를 이루고 있다:

$$\eta = g(\mu) = \beta_0 + \boldsymbol{x}'\beta$$

Y의 가능도함수는 다음과 같이 표현된다:

$$L(y; \theta, \phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y, \phi)\}$$

만약 산포모수 ϕ가 알려져 있다면 natural parameter θ에 대한 최대가능도 추정치를 찾을 수 있다. 고정된 λ에 대해 다음 식을 최소화하는 해를 찾는 문제와 동일하다:

$$\ell(\boldsymbol{\beta}, \lambda) = -\sum_{i=1}^{n}\{y_i\theta(\boldsymbol{\beta})_i - b(\theta(\boldsymbol{\beta})_i)\} + \lambda||\boldsymbol{\beta}||_1$$

$\boldsymbol{\beta}$에 zero값이 없다고 가정한다면, $\boldsymbol{\beta}$에 대한 1차 도함수는 다음과 같다:

$$H(\boldsymbol{\beta}, \lambda) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = -\boldsymbol{X}'\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{\mu})\frac{\partial \eta}{\partial \boldsymbol{\mu}} + \lambda \text{ sign}(0, \boldsymbol{\beta})$$

여기서 $\boldsymbol{X}$는 $n \times (p+1)$ 행렬이며, $\boldsymbol{W}$는 대각요소가 $V_i^{-1}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2$인 $n \times n$ 대각행렬이다. 그리고 $(\boldsymbol{y}-\boldsymbol{\mu})\frac{\partial \eta}{\partial \boldsymbol{\mu}}$는 각 요소가 $(y_i\mu_i)\left(\frac{\partial \eta}{\partial \mu}\right)_i$인 n차원 벡터이다.

이 논문은 $(p+2)$차원 공간에서 $(\boldsymbol{\beta} \in \mathcal{R}^{p+1}, \lambda \in \mathcal{R}_+)$ $H(\boldsymbol{\beta}, \lambda)$의 유일하게 결정되는 경로를 그리는 것에 있다. $\ell(\boldsymbol{\beta}, \lambda)$은 $\boldsymbol{\beta}$에 대해 convex이므로 각 λ에 대해 유일한 최솟값을 얻을 수 있는 $\boldsymbol{\beta}(\lambda)$가 존재한다. 사실 변수의 어떤 active set을 산출하는 λ의 범위 내에 $H(\boldsymbol{\beta}, \lambda) = 0$가 있는 연속이고 미분가능한 유일한 함수 $\boldsymbol{\beta}$가 존재한다.

2-1. Predictor-Corrector algorithm

predictor-corrector 알고리즘은 수치적 연속을 구현하기 위한 기본 전략 중 하나이다. 수치 연속은 1차원 모수를 통해 추적되는 비선형 방정식에 대한 솔루션 세트를 식별하기 위해 사용된다. predictor-corrector방법은 초기 조건(모수의 한 극단값에서의 해)을 사용하여 해를 명시적으로 찾고 현재 해를 기반으로 인접한 해를 계속해서 찾는 방법이다.

[Lemma 1: 초기값] *lambd*가 어떤 threshold를 초과할 때 intercept는 non-zero 계수 $\hat{\beta}_0 = g(\bar{y})$이며 다음이 성립한다:

$$H\{(\hat{\beta}_0, 0, \ldots, 0)^T, \lambda\} = 0 \text{ for } \lambda > \max_{j \in \{1,\ldots,p\}} |x_i^T \hat{W}(y - \bar{y}1)g'(\bar{y})|$$

증명 'log likelihood+penalty'를 다음과 같이 재표현할 수 있다:

$$-\sum_{i=1}^{n}\{y_i\theta(\beta)_i - b\{\theta(\beta)_i\}\} + \sum_{j=1}^{p}\{\lambda(\beta_j^+ + \beta_j^-) - \lambda_j^+\beta_j^+ - \lambda_j^-\beta_j^-\}$$

여기서 $\beta_j^+, \beta_j^- \geq 0$와 $\lambda_j^+, \lambda_j^- \geq 0$이며, $\beta_j^+, \beta_j^- = 0$와 $\beta = \beta^+ + \beta^-$을 만족한다. KKT 조건으로부터

$$\begin{aligned}
-x_j^T\hat{W}(y-\hat{\mu})\frac{\partial \eta}{\partial \mu} + \lambda - \lambda_j^+ &= 0, \\
-x_j^T\hat{W}(y-\hat{\mu})\frac{\partial \eta}{\partial \mu} + \lambda - \lambda_j^- &= 0, \\
\lambda_j^+\hat{\beta}_j^+ &= 0, \\
\lambda_j^-\hat{\beta}_j^- &= 0,
\end{aligned}$$

첫 번째와 두 번째 줄은 stationary part로 각 λ_j^+와 λ_j^-로 미분한 식이다. 세 번째와 네 번째 줄은 complementary slackness part로 부등제약식과 관련된 조건이다. 만약 λ_j^+혹은 λ_j^-가 0이거나 β_j^+ 혹은 β_j^-가 0이 되어야 한다. 이로부터

$$\left|x_j^T\hat{W}(y-\hat{\mu})\frac{\partial \eta}{\partial \mu}\right| < \lambda \Rightarrow \hat{\beta}_j = 0$$

을 암시한다. 그리고 이는 다시 $1^T\hat{W}(y-\hat{\mu})\frac{\partial \eta}{\partial \mu} = 0$에 대해 $\hat{\mu} = \bar{y}1 = g^{-1}(\hat{\beta}_0)1$을 유도한다. ■

(1) predictor step

k번째 predictor step에서 $\beta(\lambda_{k+1})$은 다음과 같이 근사된다:

$$\begin{aligned}
\hat{\beta}^{k+} &= \hat{\beta}^k + (\lambda_{k+1} - \lambda_k)\frac{\partial \beta}{\partial \lambda} \\
&= \hat{\beta}^k + (\lambda_{k+1} - \lambda_k)(X_A^T W_k X_A)^{-1}\mathrm{Sgn}(0, \beta^k)^T
\end{aligned}$$

여기서 W_k는 현재시점의 가중행렬이고 X_A는 현재 active set에 속하는 X의 열들의 모임을 의미한다. 위 방정식에서 β는 현재 non-zero인 계수로만 구성되어 있다. 이러한 선형화는 log-likelihood의 이차근사와 같으며, 가중 lasso step을 취함으로써 현재시점의 해 $\hat{\beta}^k$로 확장된다.

현재 시점의 active set을 산출하는 도메인에서 $f(\lambda) = H(\beta(\lambda), \lambda)$를 정의하자. $f(\lambda)$는 모든 λ에 대해 zero이다. f를 λ에 대해 미분하면 다음의 식을 얻는다:

$$f'(\lambda) = \frac{\partial H}{\partial \lambda} + \frac{\partial H}{\partial \beta}\frac{\partial \beta}{\partial \lambda} = 0$$

정리1을 통해 predictor step에서 수행하는 근사가 $\lambda_k - \lambda_{k+1}$ 차이를 작게 한다면 실제 해에 가깝게 됨을 보여준다

[정리 1] $h_k = \lambda_k - \lambda_{k+1}$라 하자. 그리고 h_k이 $lambda = \lambda_k$와 $\lambda = \lambda_{k+1}$에서의 active set이 같을 정도로 충분히 작다고 하자. 그러면 근사된 해 $\hat{\beta}^{k+}$은 실제 해 $\hat{\beta}^{k+1}$와의 차이가 $O(h_k^2)$만큼 차이가 난다.

증명 $\partial\beta/\partial\lambda = -(X_A^T W_k X_A)^{-1}\mathrm{Sgn}(0, \beta^k)^T$가 $\lambda \in (\lambda_{k+1}, \lambda_k]$에 대해 연속적으로 미분가능하므로

$$\hat{\beta}^{k+1} = \hat{\beta}^k - h_k \frac{\partial \beta}{\partial \lambda}\bigg|_{\lambda=\lambda_k} + O(h_k^2) = \hat{\beta}^{k+} + O(h_k^2)$$

이다. 테일러 전개식에 의해 이차항부터는 remained term $O(h_k^2) = h_k^2 \frac{\partial^2 \beta}{\partial \lambda^2}\Big|_{\lambda=\lambda_k} + \cdots$으로 둔다. ■

(2) corrector step

corrector step에서는 $\hat{\beta}^{k+}$를 $\ell(\beta, \lambda_{k+1})$를 최소화하는 β를 찾기 위한 초기값이라고 하자 (즉, $H(\beta, \lambda k + 1) = 0$을 만족시키는 β). 이를 위해 선형 제약이 있는 미분가능한 목적함수를 최소화하는 데 적용할 최적화 알고리즘이 실행되어야 한다. predictor step은 warm start를 수행하며, $\hat{\beta}^{k+}$이 항상 정확한 해 $\hat{\beta}^{k+1}$에 가깝도록 하기 때문에 계산비용이 작다. corrector step은 λ가 주어질 때 정확한 해를 찾을 뿐만 아니라 β의 방향을 산출한다.

[정리 2] λ_k와 $\lambda_{k+1} = \lambda_k - h_k$에서의 해 $\hat{\beta}^k$와 $\hat{\beta}^{k+1}$가 $\alpha \in [0, 1]$에 대해 $\lambda = \lambda_k - \alpha h_k$에서 다음과 같이 연결되어 있다고 하자:

$$\hat{\beta}(\lambda - \alpha h_k) = \hat{\beta}^k + \alpha(\hat{\beta}^{k+1} - \hat{\beta}^k)$$

그러면 $\hat{\beta}(\lambda - \alpha h_k)$은 실제 해 $\beta(\lambda - \alpha h_k)$와 $O(h_k^2)$만큼 떨어져 있다.

증명 $\frac{\partial\beta}{\partial\lambda}$이 $\lambda \in (\lambda_{k+1}, \lambda_k]$에서 연속적으로 미분 가능하므로 다음의 방정식을 만족한다:

$$\begin{aligned}\hat{\beta}(\lambda - \alpha h_k) &= \hat{\beta}^k - \alpha h_k \frac{\hat{\beta}^{k+1} - \hat{\beta}^k}{-h_k} \\ &= \hat{\beta}^k - \alpha h_k \frac{\partial \beta}{\partial \lambda}\bigg|_{\lambda=\lambda_k} + O(h_k^2)\end{aligned}$$

그리고 비슷하게 $\lambda = \lambda_k - alphah_k$에서의 true solusion은

$$\begin{aligned}\beta(\lambda - \alpha h_k) &= \hat{\beta}^k - \alpha h_k \frac{\partial \beta}{\partial \lambda}\bigg|_{\lambda=\lambda_k} + O(h_k^2) \\ &= \hat{\beta}(\lambda - \alpha h_k) + O(h_k^2)\end{aligned}$$

이다. ■

(3) active set

active set $\mathcal{A}$은 Lemma 1의 intercept에서부터 시작한다. 각 corrector step후에 $\mathcal{A}$의 크기가 커지는지 확인한다. $\mathcal{A}$의 확인을 위한 다음의 절차는 Rosset & Zhu (2003)와 Rosset (2004)에 의해 정당화한다:

$$\left| x_j^T W(y - \mu)\frac{\partial \beta}{\partial \lambda} \right| > \lambda \quad \text{for any } j \in \mathcal{A}^c \Rightarrow \mathcal{A} \leftarrow \mathcal{A} \cup \{j\}$$

수정된 active set을 가지고 corrector step을 active set이 더 이상 커지지 않을 때까지 반복한다. 그리고 active set으로부터 zero인 계수를 가진 변수를 제거한다:

$$|\hat{\beta}_j| = 0 \text{ for any } j \in \mathcal{A} \Rightarrow \mathcal{A} \leftarrow \mathcal{A} \setminus \{j\}$$

(4) step length

step length를 $\Delta_k = \lambda_k - \lambda_{k+1}$에 대한 두 개의 자연스러운 선택은 다음과 같다:

(a) 모든 k 에 대해 $\Delta_k = \Delta$

(b) 고정된 변화 L는 길이이며, $\Delta_k = L/||\partial\beta/\partial\lambda||_1$로 정함으로써 얻어진다.

step size를 줄임으로써 정확한 해는 λ의 finer grid에서 계산된다. 그리고 계수 경로는 더 정확해진다. 이 논문에서는 더 효율적이고 유용한 전략을 제안한다:

(c) active set을 변화시키는 가장 작은 Δ_k을 선택한다.

이것을 어떻게 얻을지 직관적으로 LARS 알고리즘처럼 설명해보자. k번째 반복의 끝에서, corrector step은 $-X_A^T W_k(y-\mu)\frac{\partial\beta}{\partial\lambda} + \text{Sgn}(0,\beta^k)^T$를 만족하는 weighted lasso의 해를 찾는 꼴이다. 이 weighted lasso는 다음 시점의 predictor step을 위한 방향을 찾는다. 만약 weight W_k가 고정되어 있자면, weighted Lars는 다음 시점의 active-set의 변화점에 대한 정확한 step length를 계산할 수 있다. 이 논문에서는 weight의 변화가 경로의 과정을 변화시킨다고 하더라도 이 step length를 사용한다.

> **[Lemma 2]** $\hat{\mu}$를 corrector step에서의 y의 추정치라 하고, 해당 weighted correlation을 다음과 같이 정의하자:
>
> $$\hat{c} = X^T\hat{W}(y-\hat{\mu})\frac{\partial\eta}{\partial\mu}$$
>
> $\mathcal{A}$의 요소들의 (intercept는 제외) 절대 상관은 λ이며, $\mathcal{A}^c$의 요소들의 절대 상관은 λ보다 작다.
>
> *증명* KKT 조건에 의해 $\hat{\beta}_j \neq 0$인 추정치에 대해 $|X^T\hat{W}(y-\hat{\mu})\frac{\partial\eta}{\partial\mu}| = \lambda$이다. ■

다음 시점의 predictor step은 $\hat{\beta}$를 확장시킨다. 따라서 현재의 correlation은 변한다. λ에서 한 단위 감소할때의 correlation에서 변화량을 다음과 같이 정의하자:

$$\begin{aligned} c(h) &= \hat{c} - ha \\ &= \hat{c} - hX^T\hat{W}X_A(X_A^T\hat{W}X_A)^{-1}\text{Sgn}(0,\hat{\beta})^T \end{aligned}$$

여기서 $h > 0$은 λ에서의 감소량이다. $\mathcal{A}^c$에 있는 어떤 요소도 $\mathcal{A}$에 있는 것과 동일한 절대 correlation을 가지게 하는 h를 찾기 위해 다음의 방정식을 풀게된다:

$$|c_j(h)| = |\hat{c}_j - ha_j| = \lambda - h \quad \text{for any } j \in \mathcal{A}^c$$

이 방정식은 λ에서 step length 추정을 다음과 같이 제안한다:

$$h = \min_{j\in\mathcal{A}}{}^{+}\left\{\frac{\lambda-\hat{c}_j}{1-a_j}, \frac{\lambda+\hat{c}_j}{1+a_j}\right\}$$

λ가 h에 의해 감소하기 전에 active set에서 어떤 변수가 0에 도달하는지 확인하기위해 다음의 방정식을 푼다:

$$\beta_j(\tilde{h}) = \hat{\beta}_j + \tilde{h}(X_A^T\hat{W}X_A)^{-1}\text{Sgn}(0,\hat{\beta})^T = 0 \quad \text{for any } j \in \mathcal{A}$$

어떤 $j \in \mathcal{A}$에 대해 $0 < \tilde{h} < h$이면, 다른 변수가 active set에 들어오기 전에 해당 변수가 active set으로부터 제거될것이다. 그러므로 h보다는 $\tilde{h}$가 다음시점의 step length로 사용된다.

이 step length의 근사에 덧붙이자면, λ의 작은 감소에 따라 active set이 변하므로 predictor step의 역할은 중요하지 않다. 그러나 λ의 크게 감소하는 경우에도 predictor step을 여전히 포함한다. predictor step 방향이 자동으로 쉽게 계산되므로 나머지의 계산은 사소하다.

active set이 변하는 지점에 knot을 두어 계수 경로를 piecewise linear로 만드는 것은 문제를 단순화시킨다. 만약 active set을 수정하는 가장 작은 step length가 우리가 추정한 값보다 크다면, 올바른 보정단계 이후에도 active set은 변하지 않는다. 또는 실제 step length가 예상보다 작아서 새로운 active variable의 진입점을 놓쳤다면 λ를 증가시켜 보정 단계를 반복한다. 따라서 경로 알고리즘은 active set이 변경되는 λ를 거의 정확하게 감지하는데, 이는 절댓값이 작은 고정값보다 커지기전에 적어도 한 번은 정확한 계수를 계산하기 때문이다.

ldentity link를 가진 가우시안 분포의 경우를 보면 piecewise linear path가 정확한 것을 볼 수 있다. $i = 1, \ldots, n$에 대해 $\hat{\mu} = X\hat{\beta}$와 $V_i = \text{var}(y_i)$가 constant이기 때문에 $H(\beta, \lambda) = -X^T(y - \mu) + \lambda \text{Sgn}(0, \beta)^T$로 단순해질 수 있다. step length는 오차없이 계산된다. predictor step가 정확한 계수값을 제공하기 때문에 corrector step은 필요하지 않다.

2-2. Degrees of freedom

자유도로 active set의 size를 사용한다. active set의 size는 경로를 따라 반드시 monotonically하게 변하지는 않는다. 즉,

$$\text{df}(\lambda) = |\mathcal{A}(\lambda)|$$

으로 표현된다. 여기서 $|\mathcal{A}(\lambda)|$는 λ에 따른 active set의 크기를 의미한다. $\text{df}(\lambda) = E|\mathcal{A}(\lambda)|$에 기반하며, lasso의 경우에 충족하는 식이다. 위 식은 Efron (2004)에서 시작하여 Zou and Hastie (2004)가 발전시켰다. 축소 (shrink)의 효과는 모형에 포함할 최적의 변수를 탐색하는 과정에서 분산에 지불되는 대가를 상쇄시킨다. Zou and Hastie (2004)의 결과에 기초하여 GLM의 일반적인 경우에 위 식을 사용하는 것을 휴리스틱하게 정당화한다.

corrector step의 끝에서 β의 추정은 weighted lasso 문제를 푼다:

$$\min_{\beta}\{(z - X\beta)^T W(z - X\beta)\} + \lambda||\beta||_1$$

여기서 $z = \eta + (y - \mu)\frac{\partial \eta}{\partial \mu}$는 working response를 의미한다. 위 식의 해는 선형식 $z = X\beta + \epsilon, \epsilon \sim (0, W^{-1})$에 근사 적합된다. 이 공분산 W^{-1}은 η와 μ의 실제 값에서 올바르며, 적절한 가정 하에서 만들어진다면 점근적으로 방어된다. 실제로 $\lambda = 0, \epsilon \sim N$일 때, working response는 지수족 분포에서 mle에 대한 가우시안성과 점근적 formula로 이어진다.

이러한 휴리스틱에 따라 우리는 변환된 반응변수 $W^{1/2}z$에 stein보조정리를 적용하여 오차가 등분산성을 따르게 한다. 자세한 내용은 (Zou and Hastie 참조). Thm2에서 lasso의 경우 $\text{df}(\lambda) = E|\mathcal{A}(\lambda)|$임을 보여준다. 우리도 비슷한 방식으로 L_1정규화된 GLM의 자유도를 보여준다. 시뮬레이션 결과 $\text{df}(\lambda)$ formula가 자유도를 상당히 정확하게 근사한다는 것을 보여준다.

2-3. Adding a quadratic penalty

X의 열이 강한 상관이 있을 때, 계수 추정이 매우 불안정하며 솔루션이 고유하지 않을 수 있다. Osborneet (2000)은 고유한 솔루션이 존재하기 위한 필욥충분조건을 제공했다. 이러한 상황을 극복하기 위해 이 논문에서는 2차 패널티항을 추가할 것을 제안한다:

$$\hat{\beta}(\lambda_1) = \text{argmin}_{\beta}\left\{\log L(\beta; y) + \lambda_1||\beta||_1 + \frac{\lambda_2}{2}||\beta||_2^2\right\}, \quad \lambda_1 \in (0, \infty), \lambda_2 \text{ is fixed, small, positive constant}$$

결과적으로 변수 간 상관성은 적합의 안정성에 영향을 미치지 않는다. 상관이 강하지 않을 경우 작은 λ_2의 2차 패널티효과는 무시할 수 있다. β의 모든 원소가 non-zero라고 가정하자. X가 full rank가 아니면 $\frac{\partial H}{\partial \beta} = X^T W X$는 singular이지만 여기에 2차 패널티를 추가함으로써

$$\tilde{H}(\beta, \lambda_1, \lambda_2) = -X^T W(y-\mu)\frac{\partial \eta}{\partial \mu} + \lambda_1 \text{sign}(0, \beta)^T + \lambda_2(0, \beta)^T$$

가 되고, 이를 미분하면 $\frac{\partial \tilde{H}}{\partial \beta}$는 non-singular가 된다:

$$\frac{\partial \tilde{H}}{\partial \beta} = X^T W X + \lambda_2 \begin{pmatrix} 0 & ,\mathbf{0}^T \\ \mathbf{0} & ,I \end{pmatrix}$$

즉 λ_2는 constant, λ_1는 open set에서 움직일 때, 현재 active set은 고유하고 연속적이며 미분가능한 함수 β는 $\tilde{H}(\beta, \lambda_1, \lambda_2) = 0$을 충족한다. non-singular와 *고유하고 연속적이며 미분가능한 계수경로*의 연결은 *암시적 함수이론 (Munkres, 1991)*을 기반으로 한다. λ_1과 λ_2의 조정으로 변수선택과 grouping effect를 조절하는데 strong correlation의 input을 다루기 위해 λ_2는 매우작은 값으로, λ_1은 다른 값으로 조정하면서 정한다.