Machine Learning Under a Modern Optimization Lens	**(PUC-Rio - IND2076, 2023.1)**

Lecture 2: Sparce Regression

24/03/2023

Lecturer: André Ramos *Scribe: Mateus Waga*

1 Introdução

Nesta aula foi apresentado o capítulo 3 do livro [1], baseado nos artigos [2, 3, 4].

Dado um conjunto de dados com n observações e p variáveis preditoras, denotadas por $\mathbf{X} = [x_1, x_2, ..., x_p]$ e uma variável de destino $\mathbf{Y} = [y_1, y_2, ..., y_n]$, o modelo de regressão linear com regularização é definido como

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{1}{2\gamma}\|\boldsymbol{\beta}\|_2^2 \\ &\text{s.t. } \|\boldsymbol{\beta}\|_0 \leq k \end{aligned}$$

onde $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_p]$ é um vetor de coeficientes de regressão e γ é um parâmetro de regularização que controla a força da penalidade de regularização e $\|\boldsymbol{\beta}\|_0$ é o número de componentes não zero.

Se $n < p$ O problema sem a restrição de norma 0 possui infinitas , logo o problema não está bem definido. Se $n \geq p$ o problema é bem definido mas a restrição adicional torna o problema mais simples e mais interpretável.

A seguir, descreveremos alguns métodos para solucionar o problema.

2 Método Primal

Trata-se de uma reformualação do problema inicial como um MIP quadrático. É introduzida a variável binária s_j que é igual a 1 se a variável x_j é selecionada e 0 caso contrário.

$$\begin{aligned} &\min \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{1}{2\gamma}\|\boldsymbol{\beta}\|_2^2 \\ &\text{s.t. } \boldsymbol{\beta} \in \mathbb{R}^p, \mathbf{s} \in \{0, 1\} \\ &\quad \mathbf{1}'\mathbf{s} \leq k \\ &\quad M^- s_j \leq \beta_j \leq M^+ s_j, \quad \forall j \in [p] \end{aligned}$$

A escolha do M afeta o desempenho dos métodos de solução. Assim para casos onde $n \geq p$ podemos utilizar a seguinte formulação para encontrar valores de M mais "apertados":

$$
\begin{aligned}
M^{-} = \min_{\boldsymbol{\beta}} & \beta_j \\
& \text{s.t. } \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{1}{2\gamma}\|\boldsymbol{\beta}\|_2^2 \leq \text{UB}
\end{aligned}
$$

$$
\begin{aligned}
M^{+} = \max_{\boldsymbol{\beta}} & \beta_j \\
& \text{s.t. } \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{1}{2\gamma}\|\boldsymbol{\beta}\|_2^2 \leq \text{UB}
\end{aligned}
$$

Onde UB é uma solução viável do problema original (escolher a priori k variáveis não zero).

Uma alternativa proposta no livro, para não utilizar o M é considerar:

$$g(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{1}{2\gamma}\|\boldsymbol{\beta}\|_2^2,$$

com

$$\boldsymbol{\nabla} g(\boldsymbol{\beta}) = -\mathbf{X}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \frac{1}{\gamma}\boldsymbol{\beta}.$$

uma vez que $g(\boldsymbol{\beta}) \geq 0$ e possui um gradiente Lipschitz contínuo, temos:

$$\|\nabla g(\boldsymbol{\beta}) - \nabla g(\widetilde{\boldsymbol{\beta}})\| \leq \ell\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|.$$

Assim, considere:

$$\widehat{\boldsymbol{\beta}} \in \underset{\|\boldsymbol{\beta}\|_0 \leq k}{\arg\min}\|\boldsymbol{\beta} - \mathbf{c}\|_2^2$$

e a função $H_k(c)$

$$H_k(c) = \widehat{\beta}_i = \begin{cases} c_i, & \text{if } i \in \{(1), \ldots, (k)\} \\ 0, & \text{otherwise.} \end{cases}$$

Abaixo o algoritmo proposto no livro, ele considera como input: $g(\boldsymbol{\beta}), L, \epsilon$ e uma solução inicial $\beta_1 \in \mathbb{R}^p$ tal que $\|\beta_1\|_0 \leq k$. Output: uma solução estacionária de primeira ordem β^*.

1: $m \leftarrow 1$
2: repeat
3: $\quad m \leftarrow m+1$
4: $\quad \boldsymbol{\eta}_m \leftarrow \mathbf{H}_k\left(\beta_{m-1} - \frac{1}{L}\nabla g\left(\beta_{m-1}\right)\right)$
5: $\quad \lambda_m \leftarrow \arg\min_\lambda g\left(\lambda\eta_m + (1-\lambda)\boldsymbol{\beta}_{m-1}\right)$
6: $\quad \boldsymbol{\beta}_m \leftarrow \lambda_m\boldsymbol{\eta}_m + (1-\lambda_m)\,\boldsymbol{\beta}_{m-1}$
7: until $g\left(\boldsymbol{\beta}_m\right) - g\left(\beta_{m-1}\right) \leq \epsilon$
8: $\text{return } \boldsymbol{\beta}_m$

3 Método Dual

A ideia deste método é utilizar uma solução dual que encontra o vetor **s** e "plugar" estes valores no modelo original. O modelo utilizado para encontrar o **s** é:

$$\begin{aligned} \min \ & \frac{1}{2}\mathbf{y}^T \left(\mathbf{I}_n + \gamma \sum_{j \in [p]} s_j \mathbf{K}_j \right)^{-1} \mathbf{y} \\ \text{s.t.} \quad & \mathrm{s} \in \mathrm{S}_k^p, \end{aligned}$$

onde $\mathbf{K}_j$ em S_+^n é definido como:

$$\mathbf{K}_j = \mathbf{X}_j \mathbf{X}_j^T$$

References

[1] D. Bertsimas and J. Dunn. *Machine learning under a modern optimization lens*. Dynamic Ideas LLC Charlestown, MA, 2019.

[2] D. Bertsimas and A. King. Or forum—an algorithmic approach to linear regression. *Operations Research*, 64(1):2–16, 2016.

[3] D. Bertsimas, A. King, and R. Mazumder. Best subset selection via a modern optimization lens. 2016.

[4] D. BERTSIMAS and B. VAN PARYS. Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. *The Annals of Statistics*, 48(1):300–323, 2020.