Statistiques avancées — Régression : Erreur de prédiction du Lasso

24 Septembre 2021

L'objectif de ce TD est de traiter les capacités prédictives du Lasso (Least absolute shrinkage and selection operator); en particulier, nous calculons des bornes supérieures sur l'erreur de prédiction du Lasso, et montrons que celui-ci surpasse dans certains cas la régression linéaire classique.

Problème 1 (Bornes sur l'erreur de prédiction du Lasso)**.** Soit $(X_i, Y_i)_{1 \leq i \leq n}$ un n-échantillon i.i.d avec $X_i \in \mathbb{R}^p$ un vecteur de prédicteurs et $Y_i \in \mathbb{R}$ une réponse, pour tout $1 \leq i \leq n$. On note $X \in \mathbb{R}^{n \times p}$ la matrice de design, $Y \in \mathbb{R}^n$ le vecteur de réponses et $\varepsilon \in \mathbb{R}^n$ le vecteur de bruit. On considère le modèle linéaire suivant, avec $\beta^* \in \mathbb{R}^p$ inconnu et $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ pour tout $1 \leq i \leq n$:

$$Y_i = X_i^\top \beta^* + \varepsilon_i. \tag{1}$$

1) Dans cette question on suppose $n \geq p$ et la matrice de covariance empirique $X^\top X$ inversible.
   a) Rappeler la formule de l'estimateur des moindres carrés ordinaires $\hat{\beta}^{LS}$.
   b) Calculer l'erreur moyenne de prédiction $n^{-1}\mathbb{E}[\|X(\hat{\beta}^{LS} - \beta^*)\|_2^2]$.
   c) Préciser la valeur de $n^{-1}\mathbb{E}[\|X(\hat{\beta}^{LS} - \beta^*)\|_2^2]$ dans le cas d'un design orthogonal, où $X^\top X = I_p$.
   d) Que se passe-t-il si $p > n$?

Dans la suite du problème, on étudie un estimateur parcimonieux de β^*, reposant sur une pénalisation ℓ_1 des moindres carrés. L'estimateur du Lasso est défini comme suit :

$$\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \tag{2}$$

où $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$. Le paramètre $\lambda > 0$ est le paramètre de régularisation, contrôlant la parcimonie de l'estimateur $\hat{\beta}$. L'objectif du problème est de calculer une borne supérieure sur l'erreur de prédiction du Lasso $n^{-1}\|X(\hat{\beta} - \beta^*)\|_2^2$, et de la comparer à l'erreur de prédiction des moindres carrés ordinaires. Pour $\beta \in \mathbb{R}^p$, soit

$$\ell_n(\beta, \beta^*) = \frac{1}{n}\|X(\beta - \beta^*)\|_2^2.$$

On va prouver un résultat de la forme suivante :

$$\ell_n(\hat{\beta}, \beta^*) \leq R(\beta^*, \sigma^2, n, p, \delta) \text{ avec probabilité au moins } 1 - \delta,\ \delta \in (0, 1). \tag{3}$$

Dans l'équation (3), $R(\beta^*, \sigma^2, n, p, \delta)$ est une borne supérieure valide avec grande probabilité (par rapport à la distribution du bruit ε) qui dépend de la dimension du problème p, du nombre d'observations n, de la variance du bruit σ^2, et de la probabilité $1-\delta$ avec laquelle on veut contrôler l'erreur de prédiction $\ell_n(\hat{\beta}, \beta^*)$.

2) En utilisant la définition $\hat{\beta}$ comme un minimiseur de $\mathcal{F}(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$, montrer que, pour tout $\beta \in \mathbb{R}^p$:

$$\ell_n(\hat{\beta}, \beta^*) \leq \ell_n(\beta, \beta^*) + 4\lambda\|\beta\|_1 + \frac{2}{n}\varepsilon^\top X(\hat{\beta} - \beta) - 2\lambda\left(\|\beta\|_1 + \|\hat{\beta}\|_1\right). \quad (4)$$

3) Pour un vecteur $x \in \mathbb{R}^p$, on définit la norme ℓ_∞ de x : $\|x\|_\infty = \max_{1 \leq j \leq p} |x_j|$. Montrer que, pour tout $\beta \in \mathbb{R}^p$:

$$\frac{1}{n}\varepsilon^\top X(\hat{\beta} - \beta) - \lambda\left(\|\beta\|_1 + \|\hat{\beta}\|_1\right) \leq \left(\frac{1}{n}\|\varepsilon^\top X\|_\infty - \lambda\right)\left(\|\hat{\beta}\|_1 + \|\beta\|_1\right). \quad (5)$$

Indice : utiliser l'inégalité suivante pour $x, y \in \mathbb{R}^p$, découlant de la dualité des normes ℓ_1 et ℓ_∞ : $x^\top y \leq \|x\|_\infty \|y\|_1$.

On cherche à présent à montrer que, pour une valeur de λ bien choisie, le membre de droite de l'inégalité (5) est négatif ou nul avec grande probabilité

4) Pour $1 \leq j \leq p$, notons X^j la j-ième colonne de la matrice de design X. Soit $\delta \in (0, 1)$ fixé. On suppose $\|X^j\|_2^2 \leq n$ pour tout $1 \leq j \leq p$, et

$$\lambda = \sigma\sqrt{\frac{2}{n}\ln(p/\delta)}.$$

   a) On définit la variable aléatoire $\zeta_j = \frac{\varepsilon^\top X^j}{\sigma\|X^j\|_2}$. Quelle loi suit ζ_j ?

   b) Montrer que, pour tout $1 \leq j \leq p$,

$$\mathbb{P}\left(|\varepsilon^\top X^j| > \sigma\sqrt{2n\ln(p/\delta)}\right) \leq \frac{\delta}{p}.$$

   Indice : utiliser l'inégalité de concentration Gaussienne $\mathbb{P}(\xi > x) \leq \frac{1}{2}\exp(-x^2/2)$ pour $\xi \sim \mathcal{N}(0, 1)$.

   c) En déduire que $\mathbb{P}\left(n^{-1}\|\varepsilon^\top X\|_\infty > \lambda\right) \leq \delta$.

5) En utilisant les réponses aux questions précédentes, conclure que, pour $\delta \in (0, 1)$ fixé, si $\lambda = \sigma\sqrt{\frac{2}{n}\ln(p/\delta)}$ alors, avec probabilité au moins $1 - \delta$,

$$\ell_n(\hat{\beta}, \beta^*) \leq \inf_{\beta \in \mathbb{R}^p}\left\{\ell_n(\beta, \beta^*) + 4\sqrt{2}\sigma\sqrt{\frac{\ln(p/\delta)}{n}}\|\beta\|_1\right\}. \quad (6)$$

L'inégalité (6) est appelée une "inégalité oracle". En effet, elle compare l'erreur de prédiction de l'estimateur du Lasso, $\ell_n(\hat{\beta}, \beta^*)$, à l'erreur de prédiction du meilleur estimateur parcimonieux de β^*. Ce meilleur estimateur, noté $\bar{\beta}$, satisfait :

$$\bar{\beta} \in \mathrm{argmin}_{\beta \in \mathbb{R}^p}\left\{\ell_n(\beta, \beta^*) + 4\sqrt{2}\sigma\sqrt{\frac{\ln(p/\delta)}{n}}\|\beta\|_1\right\}.$$

En pratique, on ne connait pas l'estimateur oracle $\bar{\beta}$, car on ne peut pas calculer la perte $\ell_n(\beta, \beta^*)$ qui dépend de β^*, lui-même inconnu. Cependant, sous certaines conditions sur β^*, on peut obtenir un résultat plus précis.

7) On suppose $\|\beta^*\|_0 = \sum_{j=1}^p 1_{\{|\beta_j|>0\}} \leq s$, et $\|\beta^*\|_\infty \leq a$. Montrer que

$$\ell_n(\hat{\beta}, \beta^*) \leq C\sigma\sqrt{\frac{\ln(p/\delta)}{n}}as \tag{7}$$

avec probabilité au moins $1-\delta$, et C une constant numérique que l'on précisera.

8) Comparer la borne supérieure sur l'erreur de prédiction du Lasso obtenue en (7) à l'erreur de prédiction de l'estimateur des moindres carrés calculé à la question 1) et conclure.