TD4 : Régression en grande dimension

27/03/2025

Le but de ce TD est d'aborder certaines techniques de régression en grande dimension. Lorsque le nombre de variables explicatives p est grand, les estimateurs MC des paramètres du modèle de RLM

$$Y = w_0 + w_1 X_1 + \cdots + w_p X_p + \varepsilon$$

possèdent généralement une grande variance. Pour réduire cette variance (quitte à augmenter un "peu" le biais), on impose une contrainte sur les paramètres $(w_1, \ldots, w_p)$ du type $\|(w_1, \ldots, w_p)\| \leq b$, où $b > 0$ est un seuil donné, et $\| \cdot \|$ est une norme sur $\mathbb{R}^p$.

L'estimateur MC "pénalisé" sera alors défini par

$$\widehat{\mathbf{w}}_{\text{pen}} := \underset{\{\mathbf{w} \in \mathbb{R}^{1+p} \text{ t.q. } \|(w_1,\ldots,w_p)\| \leq b\}}{\arg\min} \frac{1}{n} \|\mathbf{Y} - \mathbb{X}\mathbf{w}\|^2.$$

Plusieurs questions se posent alors : quelle norme utiliser pour la contrainte? existence, unicité et calcul de la solution $\widehat{\mathbf{w}}_{\text{pen}}$ du problème de minimisation précédent? le choix du seuil b?

Régression ridge

La régression ridge consiste à minimiser le critère des moindres carrés pénalisé par la norme L_2 du vecteur $(w_1, \ldots, w_p)$. L'estimateur ridge de $\mathbf{w} := (w_0, w_1, \ldots, w_p)^\top$ est défini alors comme suit

$$\widehat{\mathbf{w}}^R := \underset{\mathbf{w} \in \mathbb{R}^{1+p} \text{ t.q. } \sum_{j=1}^p w_j^2 \leq b}{\arg\min} \frac{1}{2n} \sum_{i=1}^n \left(Y_i - w_0 - \sum_{j=1}^p w_j X_{i,j} \right)^2 \tag{1}$$

ou de façon équivalente

$$\widehat{\mathbf{w}}^R := \underset{\mathbf{w} \in \mathbb{R}^{1+p}}{\arg\min} \left\{ \frac{1}{2n} \sum_{i=1}^n \left(Y_i - w_0 - \sum_{j=1}^p w_j X_{i,j} \right)^2 + \lambda \sum_{j=1}^p w_j^2 \right\}. \tag{2}$$

Les deux définitions précédentes sont équivalentes dans le sense où pour tout $b > 0$ il existe un unique $\lambda > 0$ tels que les deux solutions coïncident.

Remarques

(1) Le coefficient w_0 (de l'intercept) n'est pas pris en compte dans la pénalité;

(2) L'estimateur ridge dépend évidement du paramètre b (ou λ) : $\widehat{\mathbf{w}}^R := \widehat{\mathbf{w}}^R(b) := \widehat{\mathbf{w}}^R(\lambda)$;

(3) Les variables explicatives sont le plus souvent réduites pour éviter les problèmes d'échelle dans la pénalité;

(4) Sous R, le calcul des estimateurs ridge peut se faire à l'aide de la fonction `glmnet()` du package `glmnet` en spécifiant l'argument `alpha = 0`.

Propriétés de l'estimateur ridge

(1) Lorsque les variables explicatives sont centrées-réduites, l'estimateur ridge s'écrit

$$\widehat{\mathbf{w}}^R = \widehat{\mathbf{w}}^R(\lambda) = \left(\mathbb{X}^\top \mathbb{X} + \lambda\, \mathbb{I}_{1+p}\right)^{-1} \mathbb{X}^\top \mathbf{Y}.$$

(2) On en déduit

$$\text{biais}(\widehat{\mathbf{w}}^R) = -\lambda \left(\mathbb{X}^\top \mathbb{X} + \lambda\, \mathbb{I}_{1+p}\right)^{-1} \mathbf{w}$$

et

$$\text{Var}(\widehat{\mathbf{w}}^R) = \sigma^2 \left(\mathbb{X}^\top \mathbb{X} + \lambda\, \mathbb{I}_{1+p}\right)^{-1} \mathbb{X}^\top \mathbb{X} \ \left(\mathbb{X}^\top \mathbb{X} + \lambda\, \mathbb{I}_{1+p}\right)^{-1}.$$

(3) Si $\lambda = 0$, on retrouve le biais (null) et la variance de l'estimateur MC;

(4) Si λ augmente, le biais augmente et la variance diminue;

(5) Si λ diminue, le biais diminue et la variance augmente;

(6) Si $\lambda \approx 0$, alors $\widehat{\mathbf{w}}^R \approx \widehat{\mathbf{w}}$ l'estimateur MC. Si λ est grand, alors $\widehat{w}^R \approx \mathbf{0}$.

Choix du paramètre de pénalisation λ

Le choix de λ peut se faire, par validation croisée, de la manière suivante

(i) Estimation de l'erreur théorique (par validation croisée) pour toutes les valeurs de λ;

(ii) Choix de λ pour lequel l'erreur estimée est minimale.

Cela peut se faire sous R à l'aide de la fonction `cv.glmnet()` du package `glmnet`.

Exercice 1

(1) Appliquer la régression ridge pour expliquer/prédire la variable `Sales` en fonction des variables de la bd `Carseats`.

(2) Comparer les trois modèles suivants, en terme d'erreur théorique de prévision estimée par la méthode de validation croisée leave-one-out :

(i) le modèle de régression ridge optimal;

(ii) le modèle RLM complet;

(iii) le modèle de RLM optimal au sens du critère BIC.

Régression Lasso

La régression lasso consiste à minimiser le critère des moindres carrés pénalisé par la norme L_1 du vecteur $(w_1, \ldots, w_p)$.

L'estimateur lasso $\widehat{\mathbf{w}}^L$ est défini par

$$\widehat{\mathbf{w}}^L := \underset{\mathbf{w}\in\mathbb{R}^{1+p} \ \text{t.q.} \ \sum_{j=1}^p |w_j| \le b}{\arg\min} \ \frac{1}{2n} \sum_{i=1}^n \left(Y_i - w_0 - \sum_{j=1}^p w_j\, X_{i,j} \right)^2 \tag{3}$$

ou de façon équivalente

$$\widehat{\mathbf{w}}^L := \underset{\mathbf{w}\in\mathbb{R}^{1+p}}{\arg\min} \ \left\{ \frac{1}{2n} \sum_{i=1}^n \left(Y_i - w_0 - \sum_{j=1}^p w_j\, X_{i,j} \right)^2 + \lambda \sum_{j=1}^p |w_j| \right\}. \tag{4}$$

Remarques

(1) Comme pour la régression ridge, si λ augmente, alors le biais augmente et la variance diminue. Si λ diminue, le biais diminue et la variance augmente;

(2) Comme pour la régression ridge, les variables explicatives sont le plus souvent réduites pour éviter les problèmes d'échelle dans la pénalité;

(3) Le choix de λ peut se faire par validation croisée;

(4) Contrairement au ridge, si λ augmente, alors le nombre de coefficients nuls augmente. Donc le lasso permet de faire de la sélection de variables (contrairement au ridge);

(5) Sous R, on utilise la fonction `glmnet()` en spécifiant l'argument `alpha = 1`;

(6) Le choix de λ minimisant l'erreur théorique estimée (par validation croisée), peut se faire à l'aide de la fonction `cv.glmnet()` en spécifiant l'argument `alpha = 1`.

Exercice 2

(1) Appliquer la régression lasso pour expliquer/prédire la variable `Sales` en fonction des variables de la bd `Carseats`.

(2) Comparer les quatre modèles suivants, en terme d'erreur théorique de prévision estimée par la méthode de validation croisée leave-one-out :

(i) le modèle de régression ridge optimal;

(ii) le modèle de régression lasso optimal;

(iii) le modèle RLM complet;

(iv) le modèle de RLM optimal au sens du critère BIC.

Régression group-lasso

Dans certains cas, les variables explicatives appartiennent à des groupes de variables prédéfinis; c'est le cas par exemple des variables indicatrices des modalités d'une même variable qualitative (on veut les sélectionner toutes ou les exclure toutes). En présence de d variables réparties en L groupes $\mathbf{X}_1, \ldots, \mathbf{X}_L$ de cardinal $d_1, \ldots, d_L$, on note $\mathbf{w}_\ell := (w_{\ell,1}, \ldots, w_{\ell,d_\ell})^\top$, le vecteur des coefficients associé au groupe $\mathbf{X}_\ell$, $\ell = 1, \ldots, L$. Les estimateurs group-lasso s'obtiennent en minimisant, en $(w_0, \mathbf{w}_1, \ldots, \mathbf{w}_L)$, le critère suivant

$$(w_0, \mathbf{w}_1, \ldots, \mathbf{w}_L) \mapsto \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - w_0 - \sum_{\ell=1}^{L} \mathbf{X}_{i,\ell}\ \mathbf{w}_\ell \right)^2 + \lambda \sum_{\ell=1}^{L} \sqrt{d_\ell}\, \|\mathbf{w}_\ell\|_2.$$

Puisque $\|\mathbf{w}_\ell\|_2 = 0$ ssi $w_{\ell,1} = \cdots = w_{\ell,d_\ell} = 0$, la méthode group-lasso encourage la mise à zéro des coefficients d'un même groupe.

Pour calculer les estimateurs group-lasso, on peut utiliser la fonction `gglasso()` du package `gglasso`.

Exercice 3

(1) Construire le modèle de régression group-lasso en considérant que les variables indicatrices des modalités de la variables `Shelveloc` font partie d'un même groupe.

(2) Comparer le modèle group-lasso optimal obtenu avec tous les modèles de l'exercice 1 en terme de l'erreur théorique de prévision estimée par la méthode de validation croisée leave-one-out.