Apprentissage statistique et grande dimension

TD 2 : régression en grande dimension

Dans tout le TD, nous considérons $Y = (Y_1, ..., Y_n)^t$, $X = (x_{i,j})_{1 \leq i \leq n, 1 \leq j \leq d}$ non aléatoire vérifiant

$$Y = X\beta^* + \varepsilon,$$

avec $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ et $\beta^* = (\beta_1^*, ..., \beta_d^*)^t \in \mathbb{R}^d$.

Exercice 1 (Moindres carrés et ridge et Lasso en dimension 1)

Nous supposons ici que $d = 1$, le modèle considéré peut-être réecrit

$$Y_i = \beta^* x_i + \varepsilon_i, \quad i = 1, ..., n,$$

avec $\beta \in \mathbb{R}$. L'objectif de cet exercice est de comparer dans ce cadre-là l'estimateur des moindres carrés

$$\widehat{\beta}^{(MC)} \in \arg\min_{\beta \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - \beta x_i)^2,$$

l'estimateur *ridge* $\widehat{\beta}_\lambda^{(R)}$ et l'estimateur Lasso $\widehat{\beta}_\lambda^{(L)}$.

1. Donner l'écriture de l'estimateur $\widehat{\beta}^{(MC)}$ en fonction de $\{(Y_i, x_i), i = 1, ..., n\}$. Calculer le biais et la variance de cet estimateur.
2. Écrire le problème de minimisation que doit vérifier l'estimateur *ridge* dans ce cadre-là et donner son écriture.
3. Calculer son biais, sa variance et son risque quadratique.
4. Nous considérons maintenant l'estimateur Lasso, c'est-à-dire la solution du critère de minimisation

$$\widehat{\beta}_\lambda^{(L)} \in \arg\min_{\beta \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - \beta x_i)^2 + \lambda |\beta|.$$

Calculer la solution du problème de minimisation.

Exercice 2 (Propriétés de l'estimateur Ridge)

Nous considérons, pour $\lambda > 0$, l'estimateur Ridge

$$\widehat{\beta}_\lambda^{(R)} = \left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^t\mathbf{Y}$$

L'objectif de cet exercice est de prouver les propriétés de l'estimateurs Ridge vues en cours.

1. Montrer que le problème de minimisation

$$\min_{\beta \in \mathbb{R}^d} \left\{ \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \|\beta\|^2 \right\} \tag{1}$$

admet $\widehat{\beta}_\lambda^{(R)}$ comme unique solution.

2. Montrer que toute solution du problème d'optimisation sous contrainte suivant

$$\min_{\beta \in \mathbb{R}^d, \|\beta\| \leq M_\lambda} \left\{ \sum_{i=1}^{n} (Y_i - X_i\beta)^2 \right\}, \tag{2}$$

avec $M_\lambda = \left\| \left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^t\mathbf{Y} \right\|$ est aussi solution du problème (1). En déduire que $\widehat{\beta}_\lambda^{(R)}$ est aussi l'unique solution du problème (2).

3. Exprimer la norme au carré du biais de $\widehat{\beta}_\lambda^{(R)}$ en fonction des valeurs propres $\lambda_1, \ldots, \lambda_d$ (comptées avec multiplicité) de $\mathbf{X}^t\mathbf{X}$.
$$B_\lambda^{(R)} := \left\| \mathbb{E}\left[\widehat{\beta}_\lambda^{(R)}\right] - \beta^* \right\|^2$$
4. Exprimer le variance
$$V_\lambda^{(R)} = \mathbb{E}\left[\left\| \widehat{\beta}_\lambda^{(R)} - \mathbb{E}\left[\widehat{\beta}_\lambda^{(R)}\right] \right\|^2 \right]$$
de $\widehat{\beta}_\lambda^{(R)}$ en fonction de la variance du bruit σ^2 et des valeurs propres $\lambda_1, \ldots, \lambda_d$.

Exercice 3 (Propriétés de l'estimateur Lasso)

L'objectif est de montrer que les problèmes d'optimisation

$$\min_{\beta \in \mathbb{R}^d} \left\{ \sum_{i=1}^n (Y_i - X_i\beta)^2 + \lambda \|\beta\|_1 \right\} \tag{3}$$

et

$$\min_{\beta \in \mathbb{R}^d, \|\beta\|_1 \leq M_\lambda} \left\{ \sum_{i=1}^n (Y_i - X_i\beta)^2 \right\}, \tag{4}$$

sont équivalents lorsque M_λ est bien choisi dans le sens où l'ensemble des solutions des deux problèmes est identique (on rappelle que pour le LASSO on n'a pas unicité de la solution).

1. Montrer que, pour toutes solutions $\widehat{\beta}_\lambda^{(L,1)}$ et $\widehat{\beta}_\lambda^{(L,2)}$ du problème pénalisé (3),
$$\left\| \widehat{\beta}_\lambda^{(L,1)} \right\|_1 = \left\| \widehat{\beta}_\lambda^{(L,2)} \right\|_1 .$$
Nous noterons par la suite cette valeur commune N_λ.
2. Montrer que toute solution de (3) est aussi solution de (4) lorsque $M_\lambda = N_\lambda$.
3. Montrer que toute solution de (4) est aussi solution de (3) lorsque $M_\lambda = N_\lambda$.

Exercice 4 (Elastic net)

Nous considérons le problème de minimisation suivant :

$$\widehat{\beta}_{\lambda,\mu}^{(EN)} \in \arg\min_{\beta \in \mathbb{R}^d} \mathcal{L}_{EN}(\beta),$$

avec

$$\mathcal{L}_{EN}(\beta) = \sum_{i=1}^n (Y_i - x_i^t\beta)^2 + \lambda\|\beta\|^2 + \mu\|\beta\|_1.$$

1. Que se passe-t'il dans le cas $\lambda = 0$? Dans le cas $\mu = 0$?
2. Supposons que la condition ORT est vérifiée. Calculer la valeur minimale de $\mathcal{L}_{EN}$.
3. Pour tout $j = 1, ..., d$, calculer la dérivée partielle de $\mathcal{L}_{EN}(\beta)$ par rapport à $\beta_j \neq 0$.
4. En déduire un algorithme de descente de gradient coordonnées par coordonnées pour approcher $\widehat{\beta}_{\lambda,\mu}^{(EN)}$ lorsque, pour tous $j = 1, ..., d$, $n^{-1}\sum_{i=1}^n x_{i,j}^2 = 1$.

Exercice 5 (Group-Lasso)

Nous supposons maintenant que les covariables se répartissent en deux groupes c'est-à-dire que nous séparons l'ensemble $\{1, ..., d\}$ en deux sous-ensembles $I_1 := \{1, ..., d_1\}$ et $I_2 := \{d_1 + 1, ..., d\}$ avec $2 \leq d_1 \leq d - 1$. Nous souhaitons soit garder tous les coefficients β_j pour $j \in I_1$ non nuls, soit les annuler tous en même temps, de même pour les coefficients β_j pour $j \in I_2$. Nous considérons que l'ordonnée à l'origine est nulle ($\beta_0 = 0$ et $X = (x_{i,j})_{1\leq i\leq n, 1\leq j\leq d}$). Nous considérons le problème de minimisation suivant

$$\widehat{\beta}_{\lambda_1,\lambda_2}^{(GL)} \in \arg\min_{\beta \in \mathbb{R}^d} \mathcal{L}_G(\beta), \tag{5}$$

avec

$$\mathcal{L}_G(\beta) = \sum_{i=1}^{n}(Y_i - x_i^t\beta)^2 + \lambda_1\|\beta\|_{I_1} + \lambda_2\|\beta\|_{I_2}$$

et

$$\|\beta\|_{I_1} = \sqrt{\sum_{j=1}^{d_1}\beta_j^2} \text{ et } \|\beta\|_{I_2} = \sqrt{\sum_{j=d_1+1}^{d}\beta_j^2}.$$

1. Montrer que $\mathcal{L}_G$ est de classe $\mathcal{C}^1$ sur l'ensemble

$$D_G = \{\beta \in \mathbb{R}^d, \|\beta\|_{I_1} \neq 0 \text{ et } \|\beta\|_{I_2} \neq 0\}.$$

   et calculer son gradient de $\nabla\mathcal{L}_G(\beta)$ en tout point $\beta \in D_G$.

2. Montrer que le problème de minimisation (5) admet une solution et que toute solution $\widehat{\beta}$ vérifie

$$\begin{cases} \widehat{\beta}_{I_1} = \left(1 - \frac{\lambda_1}{2\|X_{I_1}^t R_2\|}\right)_+ \frac{1}{n} X_{I_1}^t R_2 \\ \widehat{\beta}_{I_2} = \left(1 - \frac{\lambda_2}{2\|X_{I_2}^t R_1\|}\right)_+ \frac{1}{n} X_{I_2}^t R_1, \end{cases}$$

   avec, pour $k = 1, 2$, $\beta_{I_k} = (\beta_i, i \in I_k)$, $X_{I_k} = (X_{i,j})_{i=1,...,n, j\in I_k}$, et $R_j = Y - X_{I_j}\beta_{I_j}$.

3. Proposer un algorithme de descente de gradient coordonnées par coordonnées pour approcher $\widehat{\beta}^{(G)}_{\lambda_1,\lambda_2}$.

Exercice 6 (Modèle logit)

Nous nous plaçons dans un premier temps dans un cadre de classification binaire où $Y_i \in \{0, 1\}$ dépend de $X_i = (X_{i,1}, ..., X_{i,d})^t \in \mathbb{R}^d$ non aléatoire. Nous supposons que

$$g(\mathbb{E}[Y_i]) = X_i^t\beta^*,$$

avec $g(x) = \log(x/(1-x))$ la fonction de lien *logit*, $\beta^* \in \mathbb{R}^d$ inconnu. Nous souhaitons estimer β^* à partir de l'observation de $\{(X_i, Y_i), i = 1, ..., n\}$. Pour cela, nous considérons deux critères basés sur la log-vraisemblance des données, que nous noterons par la suite $\ell_n(\beta)$. Un critère de type *ridge*

$$\widehat{\beta}^{(R)}_\lambda = \arg\min_{\beta\in\mathbb{R}^d}\left\{-\ell_n(\beta) + \lambda\|\beta\|^2\right\},$$

et un critère de type Lasso,

$$\widehat{\beta}^{(L)}_\lambda = \arg\min_{\beta\in\mathbb{R}^d}\left\{-\ell_n(\beta) + \lambda\|\beta\|_1\right\},$$

dépendant tous deux d'un paramètre $\lambda \geq 0$.

1. Nous commençons par étudier la log-vraisemblance $\ell_n(\beta)$.
   (a) Soit Y une variable aléatoire de loi binomiale de paramètre $\mu \in]0, 1[$, calculer la densité de la loi de Y par rapport à la mesure de comptage sur $\{0, 1\}$.
   (b) En déduire la vraisemblance de $Y_1, ..., Y_n$ en fonction de $(\mu_1, ..., \mu_n)$ telle que, pour tout i, $\mu_i = \mathbb{E}[Y_i]$.
   (c) Montrer que

$$\ell_n(\beta) = \sum_{i=1}^{n}\left(Y_i X_i^t\beta - \log\left(1 + e^{X_i^t\beta}\right)\right).$$

   (d) La fonction ℓ_n est-elle convexe ? concave ? dérivable ?
2. Peux-t'on calculer facilement l'estimateur $\widehat{\beta}^{(R)}_\lambda$? Que vaut-il pour $\lambda = 0$?
3. Supposons que nos données sont des données génomiques, c'est-à-dire que pour n femmes ayant ou non un cancer du sein, nous avons prélevé et analysé l'ARN présent dans un échantillon de leurs cellules et nous disposons des données suivantes :
   — $Y_i = 1$ si le i-ème individu est atteint d'un cancer du sein ($Y_i = 0$ sinon),
   — $X_{i,j}$ le nombre de fois où de l'ARN correspondant au j-ème gène étudié pour l'individu i a été retrouvé.
   Nous souhaitons déterminer les gènes ayant de l'influence sur l'apparition d'un cancer du sein. Devons-nous calculer l'estimateur *ridge* ou l'estimateur Lasso ?