MAP553 Apprentissage Statistique

PC6 : Seuillage, Lasso et Analyse Linéaire Discriminante

Les slides et codes `R` des PCs sont disponibles sur la page web :
`http ://www.cmap.polytechnique.fr/~giraud/MAP553/MAP553.html`

1 Seuillage dur et C_p de Mallows

Considérons le modèle linéaire

$$y = X\theta + \xi \tag{1}$$

avec $\theta \in \mathbb{R}^p$ un paramètre inconnu, ξ de loi $\mathcal{N}(0, \sigma^2 I_n)$, X une matrice $n \times p$ connue et I_n la matrice identité $n \times n$. Par exemple $X_{i,j} = \varphi_j(x_i)$ avec $x_i = i/n$ et φ_j la base trigonométrique. On supposera dans cette partie que X satisfait l'hypothèse :

Hypothèse (ORT) : $\frac{1}{n} X^T X = I_p$.

En posant $z = \frac{1}{n} X^T y \in \mathbb{R}^p$ on se ramène au modèle $z = \theta + \zeta$ avec ζ de loi $\mathcal{N}(0, \frac{\sigma^2}{n} I_p)$. Pour $m \subset \{1, \dots, p\}$, on note $\hat{\theta}_m$ le vecteur défini par $(\hat{\theta}_m)_j = z_j \mathbf{1}_{j \in m}$ pour $j = 1, \dots, p$. Les $\hat{\theta}_m$ sont des estimateurs linéaires de θ. On peut construire comme à la PC5 un estimateur "progressif" en posant $\hat{\theta}_{prog} = \hat{\theta}_{\{1,\dots,\hat{M}\}}$ avec $\hat{M}$ obtenu en minimisant le critère de Mallows

$$C_p\big(M\big) = |z - \hat{\theta}_{\{1,\dots,M\}}|_2^2 + \frac{2\sigma^2 M}{n}\,,$$

où $|\cdot|_2$ est la norme Euclidienne.

1. Pour $\tau > 0$, l'estimateur par seuillage dur est défini par $\hat{\theta}^H_j = z_j \mathbf{1}_{|z_j| > \tau}$ pour $j = 1, \dots, p$. Quel est l'avantage de $\hat{\theta}^H$ sur $\hat{\theta}_{prog}$?
2. Montrer que $\hat{\theta}^H$ est solution du problème de minimisation

$$\min_{\theta \in \mathbb{R}^p} \left\{ \sum_{j=1}^p (z_j - \theta_j)^2 + \tau^2 \sum_{j=1}^p \mathbf{1}_{\theta_j \neq 0} \right\}.$$

3. En déduire que $\hat{\theta}^H = \hat{\theta}_{\hat{m}}$ où $\hat{m}$ est solution de

$$\min_{m \subset \{1,\dots,p\}} \Big\{ |z - \hat{\theta}_m|_2^2 + \tau^2 \mathrm{card}(m) \Big\}.$$

4. Que reconnait-on lorsque $\tau^2 = 2\sigma^2/n$? Les conditions du théorème de Kneip sont-elles vérifées ?
   ⚠ le choix $\tau^2 = 2\sigma^2/n$ conduit à de très mauvais résultats en général ! (trop petit)
   Il faut prendre $\tau^2 = 2\sigma^2 \log(p)/n$ ou plus grand.

2 Seuillage doux et estimateur Lasso

On considère de nouveau le modèle (1). On s'intéresse désormais à l'estimateur Lasso $\hat{\theta}^L$ défini pour $\tau > 0$ comme solution du problème de minimisation

$$\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n} |y - X\theta|_2^2 + 2\tau \sum_{j=1}^{p} |\theta_j| \right\}. \tag{2}$$

2.1 Design orthogonal

Nous supposons dans cette partie que l'hypthèse (ORT) est vérifiée.

(a) En notant $z = \frac{1}{n} X^T y$, montrer que (2) est équivalent à

$$\min_{\theta \in \mathbb{R}^p} \left\{ \sum_{j=1}^{p} (z_j - \theta_j)^2 + 2\tau \sum_{j=1}^{p} |\theta_j| \right\}.$$

(b) En déduire que

$$\hat{\theta}_j^L = z_j \left(1 - \frac{\tau}{|z_j|} \right)_+, \quad \text{où } (x)_+ = \max(x, 0).$$

(c) Que reconnait-on ? Quels coefficients sont mis à 0 ?

2.2 Design quelconque

On ne suppose plus que l'hypothèse (ORT) est vérifiée. Dans ce cas on ne dispose plus d'une formule explicite pour $\hat{\theta}^L$. Nous allons exhiber dans la suite un algorithme (efficace) pour minimiser la fonction convexe

$$F(\theta) = \frac{1}{n} |y - X\theta|_2^2 + 2\tau \sum_{j=1}^{p} |\theta_j|.$$

On notera X_j la j-ième colonne de X et on supposera pour simplifier que $\frac{1}{n} X_j^T X_j = 1$ pour $j = 1, \dots, p$.

(a) Montrer que

$$\frac{\partial}{\partial \theta_j} F(\theta) = -\frac{2}{n} X_j^T (y - X\theta) + 2\tau \frac{\theta_j}{|\theta_j|} \quad \text{pour } \theta_j \neq 0.$$

(b) Soit $\theta \in \mathbb{R}^p$ et $\theta^{(j)}$ défini par $\theta_k^{(j)} = \theta_k$ si $k \neq j$ et

$$\theta_j^{(j)} = R_j \left(1 - \frac{\tau}{|R_j|} \right)_+ \quad \text{avec} \quad R_j = \frac{1}{n} X_j^T \left(y - \sum_{k \neq j} \theta_k X_k \right).$$

Montrer que $F(\theta^{(j)}) \leq F(\theta)$ avec inégalité stricte si $\theta^{(j)} \neq \theta$.

(c) En déduire un algorithme de minimisation numérique de F.

3 Analyse Linéaire Discriminante

Considérons un couple (X, Y) de variables aléatoires à valeurs dans $\mathbb{R}^p \times \{0, 1\}$, de loi donnée par

$$\mathbb{P}(Y = k) = \pi_k > 0 \quad \text{et} \quad \mathbb{P}(X \in dx | Y = k) = g_k(x)\, dx, \quad k \in \{0, 1\},\ x \in \mathbb{R}^p, \tag{3}$$

où $\pi_0 + \pi_1 = 1$ et g_0, g_1 sont deux densités de probabilité sur $\mathbb{R}^p$.

On note $h_* : \mathbb{R}^p \to \{0, 1\}$ le classifieur

$$h_*(x) = \mathbf{1}_{\{\pi_1 g_1(x) > \pi_0 g_0(x)\}},\ x \in \mathbb{R}^p.$$

1. Quelle est la loi de X ?
2. Montrer que le classifieur h_* vérifie

$$\mathbb{P}(h_*(X) \neq Y) = \min_h \mathbb{P}(h(X) \neq Y).$$

3. Supposons dans la suite que

$$g_k(x) = (2\pi)^{-p/2} \sqrt{\det(\Sigma_k^{-1})} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right), \qquad k = 0, 1,$$

pour deux matrices Σ_0, Σ_1 inversibles et $\mu_0, \mu_1 \in \mathbb{R}^p$, $\mu_0 \neq \mu_1$. Montrer que si $\Sigma_0 = \Sigma_1 = \Sigma$, la condition $\pi_1 g_1(x) > \pi_0 g_0(x)$ est équivalente à

$$(\mu_1 - \mu_0)^T \Sigma^{-1} \left(x - \frac{\mu_1 + \mu_0}{2} \right) > \log(\pi_0/\pi_1).$$

Commenter.

4. Supposons maintenant que π_k, μ_k, Σ sont inconnus, mais que l'on dispose d'un échantillon $(X_i, Y_i)_{i=1,\dots,n}$ i.i.d. de loi donnée par (3). Dans le cas $n > p$, proposer un classifieur $\hat{h} : \mathbb{R}^p \to \{0, 1\}$.
5. Revenons au cas où π_k, μ_k, Σ sont connus. Si $\pi_1 = \pi_0$, montrer que

$$\mathbb{P}(h_*(X) = 1 | Y = 0) = \Phi(-d(\mu_1, \mu_0)/2)$$

où Φ est la fonction de répartition d'une loi gaussienne standard et $d(\mu_1, \mu_0)$ est la distance de Mahalanobis donnée par $d(\mu_1, \mu_0)^2 = (\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0)$.

6. Lorsque $\Sigma_1 \neq \Sigma_0$, quelle est la nature de la frontière séparant $\{h_* = 1\}$ et $\{h_* = 0\}$?