Exercices pour le cours d'Apprentissage et Grande Dimension en statistique

February 27, 2017

1 Pénalisation, parcimonie et classification linéaire

1.1 Méthode LASSO

On se place dans le modèle de régression

$$\boldsymbol{y} = X\boldsymbol{\beta} + \sigma\boldsymbol{\xi}, \quad \beta \in \mathbb{R}^M$$

où $\boldsymbol{y} \in \mathbb{R}^n$ et $\boldsymbol{\xi}$ est un vecteur gaussien centré de matrice de variance-covariance l'identité (avec les notations du cours).

$$\hat{\beta}_L = \text{argmin}_{\beta\in\mathbb{R}^M} \left(\frac{\|\boldsymbol{y} - X\boldsymbol{\beta}\|_n^2}{2} + \lambda\|\beta\|_1 \right)$$

1. On suppose l'hypothèse ORT sur la matrice de design, c'est-à-dire $n^{-1}X^TX = \text{Id}_M$. Montrer que $\hat{\beta}_L$ est aussi défini par

$$\hat{\beta}_L = \text{argmin}_{\beta\in\mathbb{R}^M} \left(\frac{\sum_{j=1}^M (z_j - \beta_j)^2}{2} + \lambda\|\beta\|_1 \right)$$

   pour un vecteur $\mathbf{z}$ dépendant de $\mathbf{y}$ et X que l'on déterminera

2. En déduire, dans le cas de la question 1 que pour tout $j \leq M$

$$\hat{\beta}_j^L = sgn(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+, \quad \text{où } (x)_+ = \max(x, 0)$$

3. Si le design est quelconque . Soit X_j la j ième colonne de X et supposond que $X_j^t X_j = n$ pour tout $j \leq M$. Soit $F_n(\beta) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|_n^2 + 2\lambda\|\beta\|_1$. Montrer que

$$\frac{\partial F_n(\beta)}{\partial \beta_j} = -\frac{X_j^t(\mathbf{y} - X\beta)}{n} + 2\lambda\frac{\beta_j}{|\beta_j|}$$

4. soit $\beta \in \mathbb{R}^M$ et $\beta^{(j)}$ défini par $\beta_k^{(j)} = \beta_k$ si $k \neq j$ et

$$\beta_j^{(j)} = R_j \left(1 - \frac{\tau}{|R_j|}\right)_+, \quad R_j = \frac{1}{n} X_j^t(\mathbf{y} - \sum_{k\neq j} \beta_k X_k)$$

   Montrer que $F_n(\beta^{(j)}) \leq F_n(\beta)$ et que si $\beta^{(j)} \neq \beta$ alors $F_n(\beta^{(j)}) < F_n(\beta)$. En déduire une possibilité d'algorithme pour minimiser F_n.

1.2 Pénalisation BIC et AIC

On se place dans le modèle de régression

$$\boldsymbol{y} = \boldsymbol{\theta} + \sigma\boldsymbol{\xi}, \quad \xi = (\xi_1, \cdots, \xi_n)^t, \quad \xi_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^t$.

1. Calculer l'estimateur de maximum de vraisemblance en $\boldsymbol{\theta}$ en supposant σ connu. Quel est l'estimateur de σ si σ est supposé inconnu? Pourquoi cela pose- t- il problème?

2. Considérons le modèle indéxé par $\boldsymbol{\gamma} \in \{0,1\}^n$ tel que $\theta_i = 0$ si $\gamma_i = 0$. Pour tout $\boldsymbol{\gamma} \in \{0,1\}^n$ calculer le maximum de vraisemblance dans ce modèle.

3. En supposant σ connu, pour tout $r \in \{0, \cdots, n\}$ et tout $\boldsymbol{\gamma} \in \{0,1\}$ tel que $\sum_{i=1}^n \gamma_i = r$, soit
$$BIC(\gamma) = \ell_n(\hat{\theta}_\gamma, \gamma) - \frac{r}{2}\log n, \quad \text{où}$$
$\ell_n(\theta, \gamma)$ est la vraisemblance dans le modèle γ et $\hat{\theta}_\gamma$ le maximum de vraisemblance associé. Montrer que
$$BIC(\gamma) \leq -\frac{\sum_{i=1}^{n-r} y_{(i)}^2}{2\sigma^2} - \frac{r}{2}\log n := C(r), \quad y_{(1)}^2 \leq y_{(2)}^2 \leq \cdots \leq y_{(n)}^2$$

4. Supposons que le vrai paramètre de la loi de $\mathbf{y}$ soit $\boldsymbol{\theta} = (0, \cdots, 0)$. Montrer qu'avec une probabilité tendant vers 1,
$$C(0) \geq C(1)$$

5. Montrer que le résultat est faux si le BIC est remplacé par
$$AIC(\gamma) = \ell_n(\hat{\theta}_\gamma, \gamma) - 2r$$

1.3 Classification linéaire

Considérons un problème de classification à 2 classes $\{1,2\}$ basé sur des données $\{(y_i, x_i)_{i=1}^n$ On notera n_1 la taille de la population dans la classe 1 et n_2 celle de la classe 2. on considèrera une approche par analyses discrimnantes linéaires pour lesquelles $\hat{\pi}_1 = n_1/n$ et $\hat{\pi}_2 = n_2/n$.

1. Ecrire la résolution par l'analyse discriminante linéaire construite à partir des estimateurs des moments

2. Ecrire la résolution par l'analyse discriminante linéaire construite à partir du maximum de vraisemblance

3. Soit $\hat{\sigma}^2 = [\sum_{i:y_i=1}(x_i - \bar{x}_1)^2 + \sum_{i:y_2}(x_i - \bar{x}_2)^2]/(n-2)$ et $\hat{\sigma}_b^2 = n_1 n_2(\bar{x}_1 - \bar{x}_2)^2$ Montrer que l'estimateur $\hat{\beta}$ des moindre carrés associé à $\sum_i (y_i - \beta_0 - x_i\beta)^2$ vérifie
$$[(n-2)\hat{\sigma}^2 + n\hat{\sigma}_b^2]\hat{b} = n(\bar{x}_2 - \bar{x}_1)$$
et qu'il existe une constante c telle que
$$\hat{\beta} = c\hat{\sigma}^{-2}(\bar{x}_2 - \bar{x}_1)$$

1.4 Contrôle du risque

On suppose que la loi P du couple (X, Y) est donnée de la manière suivante: $Y \sim \mathcal{B}(p)$ pour un paramètre $p \in]0,1[$ puis $X \in \mathbb{R}^d$ est donné par sa loi conditionnelle sachant Y:
$$X|Y \sim \mathcal{N}(\mu_Y, \Sigma)$$
où Σ est une matrice définie positive et μ_0, μ_1 sont deux vecteurs distincts de $\mathbb{R}^d$.

1. Déterminer la fonction de régression

$$\eta(x) = E(Y|X = x).$$

2. En déduire la forme du classifieur de Bayes, g^*, et montrer que son risque de classification s'écrit

$$R^* = pP\big[Z > \delta/2 + \delta^{-1}\log(\frac{p}{1-p})\big] + (1-p)P\big[Z < -\delta/2 + \delta^{-1}\log(\frac{p}{1-p})\big]$$

où $\delta = \|\Sigma^{-1/2}(\mu_1 - \mu_0)\|$ et $Z \sim \mathcal{N}(0,1)$.

3. En pratique, on dispose d'un n-échantillon (X_i, Y_i), $1 \leq i \leq n$, de la loi P. On suppose que l'on est dans un cas où l'on connaît p et Σ et on se propose d'estimer μ_0 et μ_1 par maximum de vraisemblance. Donner la forme des estimateurs $\hat{\mu}_0$ et $\hat{\mu}_1$ ainsi obtenus.

4. En déduire un estimateur de la fonction de régression, $\hat{\eta}$ et une règle de classification, $\hat{g}$.

5. Montrer que $R(\hat{g}) \to R^*$ en probabilité, quand $n \to \infty$.

6. Montrer que $\hat{g}$ n'est pas universellement consistante. Il suffira de fabriquer une autre loi P' telle que si (X_i, Y_i) et (X, Y) sont iid de loi P', alors $R(\hat{g})$ ne converge pas vers R^*.

2 Minimisation du risque empirique

2.1 Dictionnaire fini et inégalité de Hoeffding

Soit $\mathcal{G} = \{f_1, ..., f_p\}$ une famille finie de classifieurs. Soit, avec la notation habituelle, $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ un famille de couples aléatoires i.i.d. de loi P. Soit $\widehat{R}_n$ le risque empirique. Soit $\hat{g}$ la règle de minimisation du risque empirique.

1. Ecrire le risque théorique $R(f_j)$ pour chaque f_j ainsi que le risque empirique $\widehat{R}_n(f_j)$.
2. En utilisant l'inégalité de Markov, montrer que pour tout $t > 0$, pour tout $j \in \{1, ..., p\}$,
$$P\left[R(f_j) - \widehat{R}_n(f_j) > t\right] \leq \frac{R(f_j)(1 - R(f_j))}{nt^2}.$$
3. En déduire que
$$P\left[R(\hat{g}) - \inf_{1\leq i\leq p} R(f_i) \leq \sqrt{\frac{p}{n\varepsilon}}\ \right] \geq 1 - \varepsilon$$
et
$$E\left[R(\hat{g}) - \inf_{1\leq i\leq p} R(f_i)\right] \leq 2\sqrt{p/n}.$$
4. Comparer avec le résultat obtenu en cours où l'on utilisait l'inégalité de Hoeffding.
5. Montrer que s'il existe un j tel que $R(f_j) = 0$, on a
$$P\left[R(\hat{g}) - \inf_{1\leq i\leq p} R(f_i) \leq \frac{1}{n}\log(p/\epsilon)\right] \geq 1 - \varepsilon.$$
Donner la majoration qui en découle pour $E\left[R(\hat{g}) - \inf_{1\leq i\leq p} R(f_i)\right]$.

2.2 Dictionnaire dénombrable et inégalité de Hoeffding

Soit $\mathcal{G} = \{f_j : j \in \mathbb{N}\}$ une famille dénombrable de classifieurs. Soit $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ un famille de couples aléatoires i.i.d. de loi P. Soit $\widehat{R}_n$ le risque empirique. Enfin, soient $p_j > 0$ des réels tels que
$$\sum_{j=0}^{\infty} p_j = 1.$$

1. En utilisant l'inégalité de Hoeffding, montrer que:
$$P\left[\forall j, R(f_j) - \widehat{R}_n(f_j) \leq \sqrt{\frac{\log\left(\frac{2}{p_j\varepsilon}\right)}{2n}}\ \right] \geq 1 - \frac{\varepsilon}{2}.$$

2. On suppose qu'il existe $\tilde{j} \in \mathbb{N}^*$ tel que

$$\inf_{j \geq 0} R(f_j) = R(f_{\tilde{j}}).$$

Montrer que la règle de classification

$$\hat{g} = f_{\hat{j}}(\cdot) \text{ où } \hat{j} \in \arg\min_{j \in \mathbb{N}} \left(\widehat{R}_n(f_j) + \sqrt{\frac{\log\left(\frac{2}{p_j \varepsilon}\right)}{2n}} \right)$$

satisfait pour tout $\epsilon > 0$:

$$P\left[R(\hat{g}) - R(f_{\tilde{j}}) \leq \sqrt{\frac{\log\left(\frac{2}{p_{\tilde{j}} \varepsilon}\right)}{2n}} + \sqrt{\frac{\log\left(\frac{2}{\varepsilon}\right)}{2n}} \right] \geq 1 - \epsilon.$$