MAP553 Apprentissage Statistique
Corrigé de l'examen du 5 décembre 2011

A. Tsybakov

Exercice 1.

1.

$$\begin{aligned}
I\!E\left[\hat{f}_n(x)\right] &= \frac{1}{h} I\!E\left[YK\left(\frac{F(X)-F(x)}{h}\right)\right] \\
&= \frac{1}{h} I\!E\left[I\!E(Y|X)K\left(\frac{F(X)-F(x)}{h}\right)\right] \\
&= \frac{1}{h}\int f(w)K\left(\frac{F(w)-F(x)}{h}\right) dF(w) \\
&= \frac{1}{h}\int_0^1 f(F^{-1}(u))K\left(\frac{u-F(x)}{h}\right) du \\
&= \int_{-F(x)/h}^{(1-F(x))/h} f(F^{-1}(F(x)+hz))K(z)\,dz.
\end{aligned}$$

2. Notons que pour $h < b/2$ et z dans le support de K ($= [-1,1]$), on a: $F(x)+hz \in [b/2, 1-b/2]$. Or, la derivée de la fonction F^{-1} sur $[b/2, 1-b/2]$ est uniformément bornée par une constante $0 < L' < \infty$. En effet, $[F^{-1}(u)]' = 1/p(F^{-1}(u))$, et la densité $p = F'$ est strictement positive et continue sur $[F^{-1}(b/2), F^{-1}(1-b/2)]$. De plus, f appartient à $\Sigma(\beta, L)$, ce qui implique pour tout z dans le support de K :

$$|f(F^{-1}(F(x)+hz)) - f(F^{-1}(F(x)))| \le L|F^{-1}(F(x)+hz) - F^{-1}(F(x))|^\beta \le L|L'hz|^\beta.$$

Par ailleurs, pour $h < b/2$ on a aussi

$$\int_{-F(x)/h}^{(1-F(x))/h} f(F^{-1}(F(x)+hz))K(z)\,dz = \int_{\mathbf{R}} f(F^{-1}(F(x)+hz))K(z)\,dz$$

On obtient donc, pour tout $h < b/2$, le résultat désiré avec la constante

$$C = L(L')^\beta \int |z|^\beta K(z)dz.$$

3. On a : $I\!E(Z|X) = f(X)K\left(\frac{F(X)-F(x)}{h}\right)$ et

$$\mathrm{Var}(Z|X) = I\!E\left[(Y-f(X))^2 K^2\left(\frac{F(X)-F(x)}{h}\right)\Big|X\right] = \sigma^2(X)K^2\left(\frac{F(X)-F(x)}{h}\right).$$

Donc,

$$\begin{aligned}\mathrm{Var}[I\!E(Z|X)] \leq I\!E[(I\!E(Z|X))^2] &\leq \int f^2(w)K^2\left(\frac{F(w)-F(x)}{h}\right)dF(w)\\ &\leq h f_{\max}^2 \int K^2\,.\end{aligned}$$

L'inégalité

$$I\!E[\mathrm{Var}(Z|X)] \leq h\sigma_{\max}^2 \int K^2$$

se démontre de façon similaire. On en déduit:

$$\mathrm{Var}(Z) = \mathrm{Var}[I\!E(Z|X)] + I\!E[\mathrm{Var}(Z|X)] \leq h(f_{\max}^2 + \sigma_{\max}^2)\int K^2$$

et

$$\mathrm{Var}(\hat{f}_n(x)) = \frac{1}{nh^2}\mathrm{Var}(Z) \leq \frac{C'}{nh}\,,$$

où $C' = (f_{\max}^2 + \sigma_{\max}^2)\int K^2$.

4. On obtient de ce qui précède la borne pour le risque quadratique (MSE) de $\hat{f}_n(x)$:

$$MSE = b^2(x) + \mathrm{Var}(\hat{f}_n(x)) \leq C^2 h^{2\beta} + \frac{C'}{nh}\,.$$

Le minimum est atteint pour $h \sim n^{-1/(2\beta+1)}$, ce qui donne la vitesse classique : $MSE = O(n^{-2\beta/(2\beta+1)})$.

5. On suppose maintenant que F est inconnue. Il est alors naturel de remplacer F par son estimateur naturel F_n (fonction de répartition empirique). L'estimateur modifié prend alors la forme suivante :

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^n Y_i K\left(\frac{F_n(X_i)-F_n(x)}{h}\right) = \frac{1}{nh}\sum_{i=1}^n Y_{(i)} K\left(\frac{i/n - F_n(x)}{h}\right)$$

où $Y_{(1)}, \ldots, Y_{(n)}$ est la permutation de $Y_1, \ldots, Y_n$ telle que les X_i correspondants vérifient $X_{(1)} \leq \cdots \leq X_{(n)}$ et on a utilisé que $F_n(X_{(i)}) = i/n$.

Exercice 2.

1. D'après le cours (cf. page 75),

$$\eta(x) = \frac{p_1(x)/2}{p_1(x)/2 + p_{-1}(x)/2}$$

où p_k désigne la densité d'un vecteur gaussien d'espérance m_k et de matrice de covariance Σ :

$$p_k(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-m_k)^T\Sigma^{-1}(x-m_k)\right) .$$

Posons $z = \Sigma^{-1/2}x$, $\mu_k = \Sigma^{-1/2}m_k$, $u_k = \mu_k^T z - \|\mu_k\|^2/2$. On a alors :

$$\begin{aligned} \eta(x) &= \frac{\exp(u_1)}{\exp(u_1)+\exp(u_{-1})} \\ &= \frac{\exp(\alpha+\beta^T x)}{\exp(\alpha+\beta^T x)+1} \end{aligned}$$

car $u_1 - u_{-1} = \alpha + \beta^T x$ avec

$$\begin{aligned} \alpha &= \frac{1}{2}(m_{-1}^T\Sigma^{-1}m_{-1} - m_1^T\Sigma^{-1}m_1), \\ \beta &= \Sigma^{-1}(m_1 - m_{-1}). \end{aligned}$$

On en déduit facilement (1). La loi P_X est un mélange de deux gaussiennes : la densité de P_X vaut $(p_1 + p_{-1})/2$.

2. D'après le cours (cf. (3.26)), le classifieur de Bayes est

$$h^*(x) = \text{sign}(\eta(x) - 1/2) = \text{sign}(\alpha + \beta^T x),$$

où $\alpha \in \mathbf{R}$ et $\beta \in \mathbf{R}^d$ sont les "vrais" paramètres du modèle logistique figurant dans (1). Comme $\alpha \in \mathbf{R}$ et $\beta \in \mathbf{R}^d$ sont inconnus, on introuduit l'ensemble de classifieurs linéaires

$$\mathcal{H}^* = \{h \, : h(x) = \text{sign}(\bar{\alpha} + \bar{\beta}^T x),\, \bar{\alpha} \in \mathbf{R},\, \bar{\beta} \in \mathbf{R}^d\},$$

et on cherche un estimateur de $\alpha \in \mathbf{R}$ et $\beta \in \mathbf{R}^d$ en minimisant une fonction de risque sur cet ensemble. D'après le cours (cf. page 86-87), la dimension de Vapnik-Chervonenkis de $\mathcal{H}^*$ est égale à celle de la collection d'ensembles correspondante

$$\mathcal{A}^* = \{A \subset \mathbf{R}^d \, : A = \{x \, : \bar{\alpha} + \bar{\beta}^T x > 0\},\, \bar{\alpha} \in \mathbf{R},\, \bar{\beta} \in \mathbf{R}^d\},$$

donc $V = d + 1$.

3. Par définition, $\hat{h}_n^{\text{erm}} = \text{argmin}_{h\in\mathcal{H}^*} R_n(h)$. D'après le cours (cf. (3.22)),

$$I\!E[R(\hat{h}_n^{\text{erm}})] \leq \min_{h\in\mathcal{H}^*} R(h) + 4\sqrt{\frac{2(V\log(n+1)+\log 2)}{n}} ,$$

où $V = d + 1$ vu la Question 2. Puisque pour le modèle de régression logistique le classifieur de Bayes appartient à la famille $\mathcal{H}^*$ (cf. Question 2), on a :

$$\min_{h\in\mathcal{H}^*} R(h) = \min_{\forall\, h} R(h) = R^*.$$

Par conséquent, l'excès de risque $I\!E[R(\hat{h}^{\mathrm{erm}})] - R^*$ admet la majoration suivante (ne dépendant que de n et d et se comportant comme $O\Big(\sqrt{\frac{\log n}{n}}\Big)$ quand $n \to \infty$) :

$$I\!E[R(\hat{h}_n^{\mathrm{erm}})] - R^* \leq 4\sqrt{\frac{2((d+1)\log(n+1) + \log 2)}{n}}\,.$$

4. Le φ-risque s'écrit sous la forme :

$$\begin{aligned} R_\varphi(f) = I\!E\varphi(-Yf(X)) &= I\!E[I\!E(\varphi(-Yf(X))|X)] \\ &= I\!E[\eta(X)\varphi(-f(X)) + (1-\eta(X))\varphi(f(X))]. \end{aligned}$$

Montrons qu'il existe $f = f_\varphi^*$ qui minimise $\eta(x)\varphi(-f(x)) + (1-\eta(x))\varphi(f(x))$ pour tout x fixé. Par conséquent, f_φ^* minimise $R_\varphi(f)$. D'après la définition de φ, pour tout $u \in \mathbf{R}$,

$$\eta\varphi(-u) + (1-\eta)\varphi(u) = \eta\log(1+e^{-u}) + (1-\eta)\log(1+e^u) = Q(u).$$

La fonction $Q : \mathbf{R} \to \mathbf{R}$ est strictement convexe et admet un seul minimum : sa dérivée ne s'annule qu'au point $u = \log(\eta/(1-\eta))$. Le minimiseur du φ-risque est donc donné par :

$$f_\varphi^*(x) = \log\left(\frac{\eta(x)}{1-\eta(x)}\right).$$

La fonction $\mathrm{sign}(f_\varphi^*)$ coïncide avec le classifieur de Bayes, car $\log\left(\frac{\eta(x)}{1-\eta(x)}\right) > 0$ ssi $\eta(x) > 1/2$. De plus, comme on suppose le modèle de régression logistique, f_φ^* minimise $R_\varphi(f)$ parmi toutes les fonctions f dans la classe

$$\mathcal{F} = \Big\{f(x) = \bar{\alpha} + \bar{\beta}^T x :\ \bar{\alpha} \in \mathbf{R},\ \bar{\beta} \in \mathbf{R}^d\Big\}.$$

En effet, sous le modèle de régression logistique, le classifieur de Bayes est de la forme $\mathrm{sign}(f)$ pour une fonction $f \in \mathcal{F}$ (cf. Question 2).

5. Considérons un RKHS avec le noyau reproduisant $K(\cdot,\cdot)$. Le *kernel trick* consiste à passer des classifiers linéaires dans l'espace de départ (dans notre cas, $\mathbf{R}^d$) aux classifiers linéaires par rapport au dictionnaire de fonctions $\phi_1, \ldots, \phi_n$ données par $\phi_i(x) = K(X_i, x)$. La *kernelisation* de la procédure $\min_{f\in\mathcal{F}} R_{n,\varphi}(f)$ est donc donnée par

$$\min_{f\in\mathcal{F}'} R_{n,\varphi}(f),$$

où $\mathcal{F}' = \{f\ : f(x) = \sum_{i=1}^n \theta_i K(X_i, x),\ \theta \in \mathbf{R}^n\}$. L'intérêt d'une telle extension : on peut approximer une variété plus riche de fonctions f^* donnant la frontière entre les classes, notamment des fonctions dans l'ensemble non-paramétrique associé à l'RKHS donné.

6. a) En faisant le changement de variable $f(x) = \log \frac{\bar{\eta}(x)}{1-\bar{\eta}(x)}$ et en utilisant la Question 4, on écrit

$$\begin{aligned} R_\varphi(f) &= I\!E[\eta(X)\log(1+e^{-f(X)}) + (1-\eta(X))\log(1+e^{f(X)})] \\ &= -I\!E[\eta(X)\log(\bar{\eta}(X)) + (1-\eta(X))\log(1-\bar{\eta}(X))], \end{aligned}$$

ce qui implique le résultat. D'après l'inégalité de Jensen, $\eta \log \frac{\eta}{\bar{\eta}} + (1-\eta)\log\frac{1-\eta}{1-\bar{\eta}} \geq 0$ avec l'égalité ssi $\eta = \bar{\eta}$. Donc, $D(\bar{\eta}) \geq 0$ avec l'égalité ssi $\eta(X) = \bar{\eta}(X)$ pour P_X-presque tout X.

b)

$$D(\bar{\eta}) = I\!E\left\{\eta(X)\log\frac{\eta(X)}{\bar{\eta}(X)} + (1-\eta(X))\log\frac{1-\eta(X)}{1-\bar{\eta}(X)}\right\}.$$

Effectuons un développement de Taylor au second ordre autour de q de la fonction définie par

$$K(p) = p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right),$$

où $p, q \in]0,1[$. On a:

$$K'(p) = \log\left(\frac{p}{q}\right) - \log\left(\frac{1-p}{1-q}\right), \quad K''(p) = \frac{1}{p(1-p)}.$$

Comme $K(q) = 0$, $K'(q) = 0$,

$$K(p) \geq \frac{(p-q)^2}{2} \min_{0\leq p\leq 1} \frac{1}{p(1-p)} = 2(p-q)^2,$$

d'où le résultat.

c) Comme prouvé dans le cours (la proposition montrant que "la classification est plus facile que l'estimation de la régression") :

$$R(\bar{h}) - R^* \leq 2I\!E\{|\bar{\eta}(X) - \eta(X)|\} \leq 2\sqrt{I\!E\{|\bar{\eta}(X) - \eta(X)|^2\}}.$$

Avec les inégalités des Questions 6a) et 6b), ceci implique le résultat. En particulier, pour le minimiseur du φ-risque empirique $\hat{f}_{n,\varphi}$,

$$R(\mathrm{sign}(\hat{f}_{n,\varphi})) - R^* \leq \sqrt{2(R_\varphi(\hat{f}_{n,\varphi}) - R_\varphi(f_\varphi^*))}.$$

Interprétation : si l'on peut garantir la convergence du "substitut convexe" $R_\varphi(\hat{f}_{n,\varphi})$ vers sa valeur oraculaire $R_\varphi(f_\varphi^*)$, alors on garantit automatiquement la convergence du vrai risque de classification $R(\mathrm{sign}(\hat{f}_{n,\varphi}))$ vers le risque de Bayes.