

# THÉORIE STATISTIQUE DE L'APPRENTISSAGE

OLIVIER CATONI

*DEA Probabilités et Applications*

*Examen du 7 juin 2002 – durée: trois heures.*

*Les candidats sont autorisés à consulter leurs notes de cours durant l'épreuve.*

Les différentes parties du sujet sont indépendantes. La notation  $\mathcal{M}_+^1(E)$  désigne l'ensemble des probabilités sur l'ensemble  $E$ .

## PREMIÈRE PARTIE

Sur l'ensemble à huit éléments  $E = \{1, 2, \dots, 8\}$ , considérons le vecteur de probabilités

$$(p_i)_{i=1}^8 = (0,1; 0,3; 0,2; 0,05; 0,07; 0,03; 0,12; 0,13).$$

**I.1.1)** Trouver un code binaire inversible  $c : E \rightarrow \{0;1\}^*$  (c'est-à-dire une application injective de  $E$  dans  $\bigcup_{n=1}^{+\infty} \{0;1\}^n$ ) qui minimise

$$\sum_{i=1}^8 p_i \ell[c(i)]$$

parmi tous les codes binaires inversibles, où  $\ell[c(i)]$  est la longueur, c'est-à-dire le nombre de bits, du code  $c(i)$  du  $i^e$  élément de  $E$ .

**I.1.2)** Calculer  $\min_c \sum_{i=1}^8 p_i \ell[c(i)]$ , où  $c$  varie dans l'ensemble des codes binaires inversibles.

**I.2.1)** Même question que **I.1.1)** pour un code binaire préfixe: trouver un code binaire préfixe de longueur moyenne minimum sous  $p$ .

**I.2.2)** Même question que **I.1.2)** pour un code binaire préfixe: calculer la longueur moyenne minimum d'un code préfixe sous  $p$ .

**I.3.1)** Même question que **I.1.1)** pour un code binaire préfixe n'utilisant que des mots de longueur paire: trouver un code binaire préfixe à valeurs dans  $\bigcup_{k=1}^{\infty} \{0,1\}^{2k}$  de longueur moyenne minimum sous  $p$  parmi tous les codes vérifiant ces propriétés.

**I.3.2)** Même question que **I.1.2)** pour un code binaire préfixe n'utilisant que des mots de longueur paire: calculer la longueur moyenne minimum sous  $p$  d'un tel code.

**I.4)** Expliquer comment construire un code arithmétique (code de Shannon-Fano-Elias) pour  $E$  muni de  $p$ . Donner le code de l'élément 4.

## DEUXIÈME PARTIE

Pour tout paramètre  $\theta \in [0; 1]$ , on considère la probabilité  $P_\theta$  sur l'ensemble  $\{1; 2; 3\}$  définie par

$$\begin{aligned} P_\theta(1) &= \frac{\theta}{2}; \\ P_\theta(2) &= \frac{1-\theta}{2}; \\ P_\theta(3) &= \frac{1}{2}. \end{aligned}$$

**II)** Pour toute valeur de  $\epsilon$  dans l'intervalle  $[0; 1[$ , calculer la loi de codage minimax sur l'ensemble  $\{P_\theta : \theta \in [\epsilon, 1]\}$ , c'est-à-dire la probabilité  $Q_\epsilon$  sur  $\{1; 2; 3\}$  vérifiant

$$\max_{\theta \in [\epsilon, 1]} \mathcal{R}(P_\theta, Q_\epsilon) = \min_{Q \in \mathcal{M}_+^1(\{1; 2; 3\})} \max_{\theta \in [\epsilon, 1]} \mathcal{R}(P_\theta, Q);$$

où  $\mathcal{R}$  désigne comme dans le cours la redondance.

## TROISIÈME PARTIE

Considérons une suite  $(E_N)_{N \in \mathbb{N}}$  d'ensembles finis et une famille

$$\{P_{\theta, N} \in \mathcal{M}_+^1(E_N) : \theta \in [0; 1]^d, N \in \mathbb{N}\}$$

de probabilités. Supposons qu'il existe pour tout  $N \in \mathbb{N}$  un estimateur

$$\hat{\theta}_N : E_N \rightarrow [0; 1]^d$$

de  $\theta$  vérifiant la condition suivante : pour tout  $\theta \in [0; 1]^d$

$$P_{\theta, N} \left( \|\hat{\theta}_N - \theta\| \geq \frac{c}{N^\beta} \right) \leq \alpha(c),$$

où  $\beta \in ]0; \frac{1}{2}]$  est un exposant fixé, où  $\lim_{c \rightarrow +\infty} \alpha(c) = 0$  et où  $\|\cdot\|$  désigne la norme Euclidienne sur  $\mathbb{R}^d$ .

**III)** Soit  $\{Q_N \in \mathcal{M}_+^1(E_N) : N \in \mathbb{N}\}$  une suite quelconque de probabilités. Montrer qu'il existe un ensemble borélien  $\Delta \subset [0; 1]^d$  de mesure de Lebesgue nulle tel que pour tout  $\theta \in [0; 1]^d \setminus \Delta$ ,

$$\limsup_{N \rightarrow +\infty} \frac{1}{\log_2(N)} \mathcal{R}(P_{\theta, N}, Q_N) \geq \beta d.$$

## QUATRIÈME PARTIE

**Compression d'une chaîne de Markov cachée.** Considérons deux ensembles finis  $\mathcal{X}$  et  $\mathcal{Y}$ . Soit  $\Theta$  l'ensemble des couples  $(p, q)$ , où  $p : \mathcal{X} \rightarrow \mathcal{M}_+^1(\mathcal{X})$  et  $q : \mathcal{X} \rightarrow \mathcal{M}_+^1(\mathcal{Y})$  sont des matrices de transition quelconques.

Pour toute valeur de  $(p, q) \in \Theta$  et tout entier  $N$ , définissons la loi conditionnelle  $Q_{p, q} : \mathcal{X} \rightarrow \mathcal{M}_+^1((\mathcal{X} \times \mathcal{Y})^N)$  par la formule

$$Q_{p, q} \left( (x_i)_{i=1}^N, (y_i)_{i=1}^N \mid x_0 \right) = \prod_{i=1}^N p(x_{i-1}, x_i) q(x_i, y_i).$$

Pour tout entier  $d \geq 2$  notons  $\psi_d : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  la plus petite fonction concave majorant la fonction

$$x \mapsto \begin{cases} 0 & \text{si } x = 0, \\ \frac{d-1}{2} \log(x) + \min \left\{ \log(d), -\frac{d-1}{2} \log \left( \frac{d-2}{2} \right) + \frac{(d-1)^2}{4x} + \frac{d}{2} \right\} & \text{si } x > 0. \end{cases}$$

**IV.1.1)** Trouver pour tout entier  $N$  une loi de codage  $Q : \mathcal{X} \rightarrow \mathcal{M}_+^1((\mathcal{X} \times \mathcal{Y})^N)$ , mélange des lois  $Q_{p,q}$ , telle que pour tout  $(x_i)_{i=0}^N \in \mathcal{X}^{N+1}$  et tout  $(y_i)_{i=1}^N \in \mathcal{Y}^N$ ,

$$-\log \left[ Q \left( (x_i)_{i=1}^N, (y_i)_{i=1}^N \mid x_0 \right) \right] \leq \inf_{\theta \in \Theta} -\log \left[ Q_\theta \left( (x_i)_{i=1}^N, (y_i)_{i=1}^N \mid x_0 \right) \right] + \gamma(N),$$

où

$$\gamma(N) = |\mathcal{X}| \left[ \psi_{|\mathcal{X}|} \left( \frac{N}{|\mathcal{X}|} \right) + \psi_{|\mathcal{Y}|} \left( \frac{N}{|\mathcal{X}|} \right) \right].$$

Montrer que

$$\lim_{N \rightarrow \infty} \frac{\gamma(N)}{\log(N)} = \left( |\mathcal{X}| + |\mathcal{Y}| - 2 \right) \frac{|\mathcal{X}|}{2}.$$

**IV.1.2)** En déduire que pour tout  $(x_i)_{i=0}^N \in \mathcal{X}^{N+1}$ , tout  $(y_i)_{i=1}^N \in \mathcal{Y}^N$ ,

$$-\log \left[ Q \left( (y_i)_{i=1}^N \mid x_0 \right) \right] \leq \inf_{\theta \in \Theta} -\log \left[ Q_\theta \left( (y_i)_{i=1}^N \mid x_0 \right) \right] + \gamma(N),$$

**IV.2)** Comment adapter la loi de codage  $Q$  construite à la question **IV.1.1)** au cas où le premier élément  $x_0$  n'est pas connu? Plus précisément, construire une loi de codage  $Q' \in \mathcal{M}_+^1(\mathcal{Y}^N)$  telle que pour tout  $(y_i)_{i=1}^N \in \mathcal{Y}^N$ ,

$$-\log Q' \left( (y_i)_{i=1}^N \right) \leq \inf_{\theta \in \Theta} \inf_{x_0 \in \mathcal{X}} -\log Q_\theta \left( (y_i)_{i=1}^N \mid x_0 \right) + \gamma'(N),$$

où

$$\lim_{N \rightarrow \infty} \frac{\gamma'(N)}{\log(N)} = \left( |\mathcal{X}| + |\mathcal{Y}| - 2 \right) \frac{|\mathcal{X}|}{2}.$$

## CINQUIÈME PARTIE

**Modèle de rupture:** On considère une variable à deux états  $Y \in \{0; 1\}$  dépendant d'un paramètre continu  $X \in [0; 1]$ . On observe un échantillon i.i.d.  $(X_i, Y_i)_{i=1}^N$  de loi  $P^{\otimes N}$  inconnue, où  $P \in \mathcal{M}_+^1([0; 1] \times \{0; 1\})$  ( $[0; 1]$  est muni de la tribu des boréliens et  $[0; 1] \times \{0; 1\}$  de son produit avec la tribu des parties de  $\{0; 1\}$ ). On souhaite estimer la loi conditionnelle  $P(dY \mid X)$ . On se place dans un modèle de rupture où la loi de  $Y$  ne dépend que de la position de  $X$  par rapport à un seuil inconnu. Plus précisément, pour tout réel  $\tau \in [0; 1]$  et tout couple de lois  $(p_0, p_1) \in \left( \mathcal{M}_+^1(\{0; 1\}) \right)^2$ , on pose

$$Q_{\tau, p_0, p_1}(y \mid x) = p_{\mathbf{1}(x \geq \tau)}(y).$$

On notera  $\Theta = [0; 1] \times \left( \mathcal{M}_+^1(\{0; 1\}) \right)^2$  l'espace des paramètres correspondant à ce modèle.

On construit un estimateur  $\hat{Q}$  de la façon suivante. On choisit un entier  $K$  compris entre 1 et  $N$  et un exposant  $\beta \in ]0, \frac{1}{2}[$ . On pose pour tout  $\tau \in [0; 1]$ , tout  $\sigma \in \{0; 1\}$ , tout  $x \in [0; 1]$ , tout  $y \in \{0; 1\}$ ,

$$\hat{p}_\sigma^\tau(y) = \frac{1 + \sum_{i=1}^K \mathbb{1}[\mathbb{1}(X_i \geq \tau) = \sigma] \mathbb{1}(Y_i = y)}{2 + \sum_{i=1}^K \mathbb{1}[\mathbb{1}(X_i \geq \tau) = \sigma]},$$

$$\hat{Q}(y|x) = \frac{\sum_{j=K+1}^{N+2} \left( \prod_{i=K+1}^N \hat{p}_{\mathbb{1}(X_i \geq X_j)}^{X_j}(Y_i) \right)^\beta \hat{p}_{\mathbb{1}(x \geq X_j)}^{X_j}(y)}{\sum_{j=K+1}^{N+2} \left( \prod_{i=K+1}^N \hat{p}_{\mathbb{1}(X_i \geq X_j)}^{X_j}(Y_i) \right)^\beta}$$

où par convention on a posé  $X_{N+1} = x$  et  $X_{N+2} = 1$ .

V) Montrer que pour toute probabilité  $P \in \mathcal{M}_+^1([0; 1] \times \{0; 1\})$ , pour tout paramètre  $\beta$  vérifiant

$$\inf_{\alpha \in ]\beta, 1]} \frac{\beta^2}{(\alpha - \beta)(2 - \alpha - \beta)} (K + 2)^\alpha \leq 1,$$

$$\mathbb{E} \left( \int_{x \in [0; 1]} \mathcal{K}[P(dy|x), \hat{Q}(dy|x)] P(dx) \right)$$

$$\leq \inf_{(\tau, p_0, p_1) \in \Theta} \int_{x \in [0; 1]} \mathcal{K}[P(dy|x), Q_{\tau, p_0, p_1}(dy|x)] P(dx)$$

$$+ \frac{\log(N - K + 2)}{\beta(N - K + 1)} + \frac{2}{K + 1},$$

où  $\mathcal{K}$  désigne la divergence de Kullback et  $\mathbb{E}$  l'espérance par rapport à l'échantillon  $(X_i, Y_i)_{i=1}^N$  de loi  $P^{\otimes N}$ .