
TD 2

Exercice 1. En 2007, Netflix a lancé un des premiers gros *challenges* à l'ensemble des statisticiens, informaticiens et autre *data scientists*. Il s'agissait de battre leur propre algorithme de prédiction de préférence des utilisateurs, **Cinematch**. Grossièrement, les données sont un sous-ensemble des entrées d'une matrice $M = (M_{i,j})_{1\leq i\leq m, 1\leq j\leq p}$: $M_{i,j} \in \{1,2,3,4,5\}$ est la note donnée par l'utilisateur i au film j. Il s'agit de reconstruire les entrées non observées. On propose la représentation suivante : on observe les entrées correspondant aux indices (supposés indépendants) $X_1 = (i_1, j_1), \dots, X_n = (i_n, j_n) \sim \Pi$. Ces entrées sont notées $Y_\ell = M_{i_\ell, j_\ell}$. Pour toute matrice $N \in \{1,\dots,5\}^{n\times p}$ on définit

$$R(N) = \mathbb{E}[(Y_\ell - N_{i_\ell,j_\ell})^2] = \sum_{i=1}^{m}\sum_{j=1}^{p}(N_{i,j} - M_{i,j})^2 \Pi((i,j))$$

et

$$r_n(N) = \frac{1}{n}\sum_{\ell=1}^{n}(Y_\ell - N_{i_\ell,j_\ell})^2.$$

Une hypothèse naturelle est qu'il existe un certain nombre, disons $K \ll m, p$, de comportements de base (amateurs de science-fiction, amateurs de cinéma russe...) et que les goûts de chaque utilisateur peuvent être approchés par une combinaison linéaire de ces goûts de base. On propose donc la loi π *a priori* suivante :

$$\underbrace{N}_{m\times p} = \underbrace{U}_{m,K}\underbrace{V}_{K,p}$$

où les entrées $U_{i,h}$ et $V_{h,j}$ sont i.i.d de loi $\mathcal{U}([-\sqrt{c/K}, \sqrt{c/K}])$ avec $c > 0$ constant. Noter qu'on a alors bien $N_{i,j} = \sum_{h=1}^{K} U_{i,h}V_{h,j}$, les $(U_{1,1},\dots,U_{m,1}),\dots,(U_{1,K},\dots,U_{m,K})$ sont interprétés comme les K comportements de base. Pour tout $\lambda > 0$ on pose

$$\hat{M}_\lambda = \int N \pi_{\exp(-\lambda r_n)}(\mathrm{d}N).$$

Pour simplifier les notations, pour toute matrice H on notera $\|H\| = \sup_{i,j}|H_{i,j}|$.

1) Expliquer pourquoi, pour tout $\lambda > 0$,

$$R(\hat{M}_\lambda) \leq \int R(N)\pi_{\exp(-\lambda r_n)}(\mathrm{d}N).$$

2) Pourquoi serait-il déraisonnable de prendre $c < 5$? On supposera donc que $c \geq 5$.
3) Vérifier que les variables aléatoires $|(Y_\ell - N_{i_\ell,j_\ell})^2 - R(N)|$ sont bornées par $4c^2$.
4) En déduire que l'on a, pour tout $\varepsilon > 0$, avec proba. au moins $1-\varepsilon$,

$$R(\hat{M}_\lambda) \leq \frac{2 \vee \left(1 + \frac{4c^2\lambda}{n} g\left(\frac{8c^2\lambda}{n}\right)\right)}{1 - \frac{4c^2\lambda}{n} g\left(\frac{8c^2\lambda}{n}\right)} \inf_\rho \left[\int R(N)\rho(\mathrm{d}N) + \frac{\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda}\right]$$

où g est la fonction de Bernstein.

On va tenter de majorer cet inf de façon explicite. Pour ceci, pour toutes matrices U^0 et V^0 et $\delta > 0$ on définit $\rho_{U^0,V^0,\delta}$ comme la loi π restreinte à l'ensemble $\{(U,V) : \|U - U^0\| \leq \delta, \|V - V^0\| \leq \delta\}$.

5) Démontrer que

$$\int R(N)\rho_{U^0,V^0,\delta}(\mathrm{d}N) \leq R(U^0V^0) + \delta^2 K(\sqrt{K}\delta + c)^2.$$

6) Démontrer que

$$\mathcal{K}(\rho,\pi) = (m+p)K\log\left(\frac{1}{\delta}\sqrt{\frac{c}{K}}\right).$$

7) En déduire qu'il existe une constante (pouvant dépendre du choix de c) $\mathcal{C}(c)$ telle que

$$R(\hat{M}_\lambda) \leq \mathcal{C}(c) \inf_{\|U^0\|,\|V^0\|\leq\sqrt{c/K}} \left[R(U^0V^0) + \frac{(m+p)K\log\left(\frac{cn}{(m+p)K}\right) + \log\left(\frac{2}{\varepsilon}\right)}{n} \right].$$

On pourra expliciter $\mathcal{C}(c)$.

8) Discuter la borne obtenue. En particulier, on pourra supposer qu'il existe effectivement U^0 et V^0 avec $\|U^0\|, \|V^0\| \leq \sqrt{c/K}$ et $M = U^0V^0$, et se concentrer sur la vitesse $\frac{(m+p)K}{n}$. Est-il possible de reconstruire correctement une matrice à $m \times p$ entrées en n'observant que n entrées aléatoires, avec $n \ll m, p$?

9) Question subsidiaire : comment implémenteriez-vous le calcul de $\hat{M}_\lambda$ en pratique ?

10) Question encore plus subsidiaire : en pratique, on ne connaît pas K. Que proposeriez-vous ?

Exercice 2. Dans cet exercice, on va étendre des résultats du cours démontrés pour des variables bornées à des variables non bornées. Pour ceci, on commence par essayer d'étendre l'inégalité de Bernstein à des variables aléatoires non bornées.

1) Soit Z une variable aléatoire de loi exponentielle $\mathcal{E}(\theta)$. On pose $f(\lambda) = \log \mathbb{E}\mathrm{e}^{\lambda(Z-\mathbb{E}(Z))}$.
   (a) Rappeler une formule explicite pour $f(\lambda)$ et son domaine de définition.
   (b) Démontrer que pour tout $u \in]0,1[$, $-\log(1-u) - u \leq \frac{u^2}{2(1-u)}$ et en déduire que

   $$\log \mathbb{E}\mathrm{e}^{\lambda Z} \leq \frac{\mathrm{Var}(Z)\lambda^2}{2(1-c\lambda)}$$

   pour une constante c que l'on précisera.
   Par analogie, on dit que Z' est sous-exponentielle avec les paramètres (v,c) ssi pour tout $\lambda \in]-1/c, 1/c[$,

   $$\log \mathbb{E}\mathrm{e}^{\lambda Z'} \leq \frac{v\lambda^2}{2(1-c\lambda)}.$$

   (c) Soient Z et Z' deux variables sous-exponentielles de paramètres respectifs (v,c) et (v',c). Démontrer que $Z+Z'$ est sous-exponentielle de paramètres $(v+v',c)$. Démontrer que αZ, pour $\alpha \in \mathbb{R}$, est sous-exponentielle $(\alpha^2 v, c)$.
   (d) Soit Z une variable aléatoire $\mathcal{N}(0,\sigma^2)$. Démontrer que Z^2 est sous-exponentielle de paramètres $(2\sigma^2, 2)$.

2) Soient $Z_1,\dots,Z_n$ des variables aléatoires indépendantes, centrées, sous-exponentielles de paramètre (v,c). Démontrer que pour tout $\lambda \in]-n/c, n/c[$ on a

$$\mathbb{E}\mathrm{e}^{\frac{\lambda}{n}\sum_{i=1}^n Z_i} \leq \mathrm{e}^{\frac{v\lambda^2}{2n(1-c\frac{\lambda}{n})}}.$$

3) On suppose que l'on observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d de loi P. On suppose que l'on a $f_1, \ldots, f_M$ des fonctions $\mathcal{X} \to \mathbb{R}$, un ensemble $\Theta \subset \mathbb{R}^M$ et une fonction de perte ℓ (pas forcément bornée). On définit pour $\theta \in \Theta$ la fonction $f_\theta(\cdot) = \sum_{i=1}^M \theta_i f_i(\cdot)$ et on suppose que pour tout $\theta \in \Theta$, $\ell(Y_i, f_\theta(X_i))$ est une variable sous-exponentielle de paramètres (v, c).

(a) En suivant la preuve du cours, démontrer que, pour tout $\varepsilon \in]0,1[$ on a

$$\mathbb{P}\left\{ \int R(f_\theta)\pi_{\exp(-\lambda r_n)}(\mathrm{d}\theta) \leq \inf_\rho \left[\int R(f_\theta)\rho(\mathrm{d}\theta) + \frac{v\lambda}{n\left(1 - c\frac{\lambda}{n}\right)} + 2\frac{\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right] \right\}.$$

(b) Comparer avec le résultat du cours pour les variables bornées (par exemple, en appliquant ce théorème au cas de la MS-agrégation).