Partiel 2024/2025 – Durée 1h30

Vous serez évalués sur la démarche et la maîtrise des concepts utilisés. Par conséquent, peu d'importance sera accordée aux constantes numériques.

PRÉDICTION MONOTONE

On se donne $D_n = ((X_1, Y_1), \dots, (X_n, Y_n))$ un échantillon, où $X_i = i/n$ et Y_i est de la forme

$$Y_i = f^*(X_i) + \varepsilon_i,$$

où les ε_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. On cherche à construire $f : [0, 1] \to \mathbb{R}$ qui prédise au mieux $(Y_1, \dots, Y_n)$, au sens du risque

$$R(f) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2\right).$$

1. De quel type de problème (classification/régression) s'agit-il ? Trouver la forme d'un prédicteur de Bayes f^*.
2. Pour un prédicteur f on note μ_f le vecteur $(f(1/n), \dots, f(1))^T$. Donner l'expression de l'excès de risque $R(f) - R(f^*)$ en fonction de μ_f et μ_{f^*}.

À partir d'ici on oubliera totalement les f et f^* pour travailler exclusivement avec les μ et μ^*.

3. On note $\hat{\mu}_{LS}$ l'estimateur des moindres carrés associés à $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, c'est à dire

$$\hat{\mu}_{LS} \in \arg\min_{u \in \mathbb{R}^n} \|\mathbf{Y} - u\|^2.$$

   Que vaut $\mathbb{E}(\|\hat{\mu}_{LS} - \mu^*\|^2)$? On admettra que c'est la vitesse optimale au sens de $\inf_{\hat{\mu}} \sup_{\mu^* \in \mathbb{R}^*} \mathbb{E}(\|\hat{\mu} - \mu^*\|^2)$.

4. Soit V un sous-espace vectoriel de $\mathbb{R}^n$ de dimension k. On note $\hat{\mu}_V$ l'estimateur par moindre carrés associé ($\arg\min_{\mu \in V} \|\mathbf{Y} - \mu\|^2$), et ${\mu^*}_V = \pi_V(\mu^*)$ (projection orthogonale). Exprimer $E(\|\hat{\mu}_V - \mu^*\|^2)$ en fonction des quantités précédemment introduites.

On admettra pour la suite le résultat de concentration Gaussienne suivant : si $g = (g_1, \dots, g_n)$ est un vecteur Gaussien standard, et K est cône convexe fermé de $\mathbb{R}^n$, alors, pour tout $x > 0$,

$$\mathbb{P}\left(\|\pi_K(g)\| \geq \mathbb{E}(\|\pi_K(g)\|) + \sqrt{2x}\right) \leq e^{-x},$$

où π_K désigne la projection sur K.

5. Soit $\mathcal{C} = \{C_1, \dots, C_k\}$ une partition de $[\![1, n]\!]$. On suppose que μ^* (ou f^*) est constante sur chaque élément de la partition. Trouver un estimateur $\hat{\mu}_{\mathcal{C}}$ adapté et montrer que, pour tout $x > 0$,

$$\mathbb{P}\left(\|\hat{\mu}_{\mathcal{C}} - \mu^*\| \geq \sigma\sqrt{k} + \sigma\sqrt{2x}\right) \leq e^{-x}.$$

6. On suppose maintenant que μ^* (ou de manière équivalente f^*) est monotone. On note $V(\mu) = \mu_n - \mu_1$ (pour un $\mu \in \mathbb{R}^n$ quelconque). Montrer que, pour tout $k \leq n$, il existe μ^*_{kpm} constant sur k morceaux tel que

$$\|\mu^* - \mu^*_{kpm}\| \leq \sqrt{n}\frac{V(\mu^*)}{2k}.$$

On suppose à partir d'ici que μ^* est croissante. Pour $\mu \in \mathbb{R}^n$ on note $R(\mu) = \|\mu - \mu^*\|^2$, $R_n(\mu) = \|\mathbf{Y} - \mu^*\|^2$.

7. Pour une partition $\mathcal{C}$ à k éléments, montrer que, avec probabilité plus grande que $1 - e^{-x}$, pour tout $\mu_{\mathcal{C}}$ constant par morceaux sur $\mathcal{C}$,

$$|(R(\mu_{\mathcal{C}}) - R(\mu^*)) - (R_n(\mu_{\mathcal{C}}) - R_n(\mu^*))| \leq \frac{1}{2}(R(\mu_{\mathcal{C}}) - R(\mu^*)) + 2\sigma^2\left(\sqrt{k+1} + \sqrt{2x}\right)^2.$$

8. En déduire que, avec probabilité plus grande que $1 - e^{-x}$, **pour toute partition $\mathcal{C}$ à k éléments**, pour tout $\mu_{\mathcal{C}}$ constant par morceaux sur $\mathcal{C}$,

$$|(R(\mu_{\mathcal{C}}) - R(\mu^*)) - (R_n(\mu_{\mathcal{C}}) - R_n(\mu^*))| \leq \frac{1}{2}(R(\mu_{\mathcal{C}}) - R(\mu^*)) + \sigma^2 p_{k,n}(x),$$

avec $p_{k,n}(x) = 32\left(\sqrt{(k+1)\log(n)} + \sqrt{x}\right)^2$.

9. Soit $k \leq n$. Construire $\hat{\mu}_k$ vérifiant

$$\mathbb{P}\left(\|\hat{\mu}_k - \mu^*\|^2 \geq 3\inf_{\mathcal{C}\in\mathcal{P}_k^n}\|\mu^*_{\mathcal{C}} - \mu^*\|^2 + 6\sigma^2 p_{k,n}\right) \leq e^{-x},$$

où $\mathcal{P}_k^n$ est l'ensemble des partitions à k éléments de $[\![1; n]\!]$, et $\mu^*_{\mathcal{C}}$ est le meilleur approximant de μ^* constant par morceaux sur $\mathcal{C}$ (au sens de $\mu^*_{\mathcal{C}} \in \arg\min_{\mu \text{ cpm } \mathcal{C}} \|\mu - \mu^*\|^2$).

10. On suppose σ^2 et $V(\mu^*)$ connus. Construire un prédicteur $\hat{\mu}$ vérifiant

$$\mathbb{P}\left(\|\hat{\mu} - \mu^*\|^2 \geq c_1 n^{1/3}\log(n)^{2/3}V^{2/3}(\mu^*)\sigma^{4/3} + c_2\sigma^2 x\right) \leq e^{-x},$$

où c_1 et c_2 sont des constantes. Au vu de ce résultat, quelle "dimension" peut-on donner à l'ensemble des fonctions monotones sur n points, du point de vue de ce problème d'apprentissage ?

SOLUTION

1. On cherche à prédire $(f^*(1/n), \dots, f^*(1))^T \in \mathbb{R}^n$, c'est donc un problème de régression (multivariée). Pour un prédicteur f, on a

$$R(f) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}(f^*(X_i) + \varepsilon_i - f(X_i))^2\right) = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - f^*(X_i))^2.$$

   Un prédicteur de Bayes est donc caractérisé par $f(i/n) = f^*(i/n)$, pour tout $i \in [\![1, n]\!]$ (en dehors de ces points peu importe).

2. D'après ce qui précède, on a

$$R(f) - R(f^*) = \sigma^2 + \frac{1}{n}\|\mu_f - \mu_{f^*}\|^2 - \sigma^2 = \frac{1}{n}\|\mu_f - \mu_{f^*}\|^2.$$

   On se ramène donc à un problème de stats inférentielles 'classique'.

3. C'est un piège grossier : on a évidemment $\hat{\mu}_{LS} = \mathbf{Y}$, et

$$\mathbb{E}(\|\hat{\mu}_{LS} - \mu^*\|^2) = \mathbb{E}(\|\varepsilon\|^2) = n\sigma^2,$$

   en notant $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$.

4. De manière immédiate encore on a $\hat{\mu}_V = \pi_V(\mathbf{Y})$, dont il s'ensuit

$$\begin{aligned}\mathbb{E}(\|\hat{\mu}_V - \mu^*\|^2) &= \mathbb{E}\left(\|\hat{\mu}_V - \mu_V^*\|^2 + \|\mu_V^* - \mu^*\|^2\right) \qquad \text{(Pythagore)}\\ &= \|\mu_V^* - \mu^*\|^2 + \mathbb{E}\left(\|\pi_V(\mathbf{Y} - \mu^*)\|^2\right)\\ &= \|\mu_V^* - \mu^*\|^2 + \mathbb{E}(\|\pi_V(\varepsilon)\|^2).\end{aligned}$$

   Comme $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, $\pi_V(\varepsilon) \sim \mathcal{N}(0, \sigma^2\pi_V)$, et $\mathbb{E}(\|\pi_V(\varepsilon)\|^2) = \sigma^2\mathrm{Tr}(\pi_V) = k\sigma^2$. On conclut alors

$$\mathbb{E}(\|\hat{\mu}_V - \mu^*\|^2) = \|\mu_V^* - \mu^*\|^2 + k\sigma^2.$$

5. Pour $j \in [\![1, k]\!]$, on note v_j le vecteur de $\mathbb{R}^n$ de coordonnées $(v_j)_i = \mathbb{1}_{i \in C_j}$, pour $1 \le i \le n$, et V le sous-espace vectoriel engendré par les v_j. V est de dimension k, et un vecteur μ est constant par morceaux sur $\mathcal{C}$ si et seulement si il appartient à V. Par hypothèse $\mu^* \in V$. On choisit alors naturellement comme estimateur $\hat{\mu}_{\mathcal{C}} = \pi_V(\mathbf{Y})$. Avec le même calcul que précédemment, on a

$$\|\hat{\mu}_{\mathcal{C}} - \mu^*\| = \|\pi_V(\varepsilon)\|.$$

   En utilisant le résultat de concentration Gaussienne, on a que, avec probabilité plus grande que $1 - e^{-x}$,

$$\|\pi_V(\varepsilon)\| \le \mathbb{E}(\|\pi_V(\varepsilon)\|) + \sigma\sqrt{2x},$$

   et donc

$$\|\hat{\mu}_{\mathcal{C}} - \mu^*\| \le \mathbb{E}(\|\pi_V(\varepsilon)\|) + \sigma\sqrt{2x}.$$

   On conclut en remarquant que $\mathbb{E}(\|\pi_V(\varepsilon)\|) \le \sqrt{\mathbb{E}(\|\pi_V(\varepsilon)\|^2)} = \sqrt{k\sigma^2}$ (Jensen).

6. On suppose que μ^* est monotone. Pour $k \leq n$, on divise $[\mu_1^*, \mu_n^*]$ en k intervalles $I_j = [\mu_1^* + (j-1)V(\mu^*)/k; \mu_1^* + jV(\mu^*)/k[$ (en prenant le dernier intervalle fermé). On peut alors partitionner $[\![1; n]\!]$ en $C_1, \ldots, C_k$, où $C_j = \{i \mid \mu_i^* \in I_j\}$. On prend alors μ_{kpm}^* constant sur C_j, de valeur $c_j = \mu_1^* + (j-1/2)V(\mu^*)/k$, de telle sorte que si $x \in I_j$, $|x - c_j| \leq V(\mu^*)/2k$. On peut alors écrire

$$\begin{aligned}\|\mu^* - \mu_{kpm}^*\| &= \sqrt{\sum_{j=1}^{k} \sum_{i \in C_j} (\mu_i^* - c_j)^2} \\ &\leq \sqrt{\sum_{j=1}^{k} |C_j| V(\mu^*)^2/4k^2} \\ &\leq \frac{V(\mu^*)}{2k}\sqrt{n}.\end{aligned}$$

7. Se référer au cours pour une version plus générale et détaillée. Ici, pour n'importe quel $\mu_{\mathcal{C}}$ constant par morceaux sur $\mathcal{C}$, on a

$$\begin{aligned}R(\mu_{\mathcal{C}}) - R(\mu^*)) - (R_n(\mu_{\mathcal{C}}) - R_n(\mu^*)) &= \|\mu_{\mathcal{C}} - \mu^*\|^2 - \|\mu_{\mathcal{C}} - \mu^* - \varepsilon\|^2 + \|\varepsilon\|^2 \\ &= 2\langle \varepsilon, \mu_{\mathcal{C}} - \mu^* \rangle .\end{aligned}$$

On en déduit

$$|R(\mu_{\mathcal{C}}) - R(\mu^*)) - (R_n(\mu_{\mathcal{C}}) - R_n(\mu^*))| \leq 2\,|\langle \varepsilon, \mu_{\mathcal{C}} - \mu^* \rangle| .$$

On note V' l'espace vectoriel engendré par les μ constant par morceaux sur $\mathcal{C}$ et μ^*. On a alors $\dim(V') = k+1$, et

$$\begin{aligned}2\,|\langle \varepsilon, \mu_{\mathcal{C}} - \mu^* \rangle| &= 2\,|\langle \pi_{V'}(\varepsilon), \mu_{\mathcal{C}} - \mu^* \rangle| \\ &\leq 2\|\pi_{V'}(\varepsilon)\|\|\mu_{\mathcal{C}} - \mu^*\| \\ &\leq 2\|\pi_{V'}(\varepsilon)\|^2 + \frac{1}{2}\|\mu_{\mathcal{C}} - \mu^*\|^2,\end{aligned}$$

en utilisant $2ab \leq (1/2)a^2 + 2b^2$. On utilise maintenant le résultat de concentration Gaussienne : avec probabilité $1 - e^{-x}$ (portant sur ε donc ne dépendant pas de $\mu_{\mathcal{C}}$),

$$\begin{aligned}\|\pi_{V'}(\varepsilon)\| &\leq \mathbb{E}\left(\|\pi_{V'}(\varepsilon)\|\right) + \sigma\sqrt{2x} \\ &\leq \sqrt{\mathbb{E}(\|\pi_{V'}(\varepsilon)\|^2)} + \sigma\sqrt{2x} \\ &\leq \sigma(\sqrt{(k+1)} + \sqrt{2x}).\end{aligned}$$

On conclut en remarquant que $\|\mu_{\mathcal{C}} - \mu^*\|^2 = R(\mu_{\mathcal{C}}) - R(\mu^*)$.

8. La différence avec la question précédente est l'uniformité en le choix de la partition. Il s'agit alors de dénombrer ce choix, puis de prendre une borne d'union. Choisir

une partition de $[\![1;n]\!]$ à k éléments revient à choisir $(k-1)$ frontières consécutives, on a alors immédiatement

$$|\mathcal{P}_k^n| \leq \binom{n}{k-1} \leq n^{k-1}.$$

Pour un α quelconque, on a alors, au vu de la question précédente

$$\mathbb{P}\left(\bigcup_{C\in\mathcal{P}_k^n}\bigcup_{\mu_C}\{|(R(\mu_C)-R(\mu^*))-(R_n(\mu_C)-R_n(\mu^*))| > \frac{1}{2}(R(\mu_C)-R(\mu^*))\right.$$
$$\left.+2\sigma^2\left(\sqrt{k+1}+\sqrt{2(x+\alpha)}\right)^2\}\right) \leq \left(\sum_{C\in\mathcal{P}_k^n} e^{-\alpha}\right)e^{-x} \leq n^{k-1}e^{-\alpha}e^{-x}.$$

En choisissant $\alpha = (k-1)\log(n)$, on a, avec probabilité plus grande que $1-e^{-x}$, uniformément en $\mathcal{C}$ et $\mu_{\mathcal{C}}$,

$$\begin{aligned}
&|(R(\mu_{\mathcal{C}})-R(\mu^*))-(R_n(\mu_{\mathcal{C}})-R_n(\mu^*))| \\
&\quad\leq \frac{1}{2}(R(\mu_{\mathcal{C}})-R(\mu^*)) + 2\sigma^2\left(\sqrt{k+1}+\sqrt{2(x+(k-1)\log(n))}\right)^2 \\
&\leq \frac{1}{2}(R(\mu_{\mathcal{C}})-R(\mu^*)) + 2\sigma^2 4^2\left(\sqrt{(k+1)\log(n)}+\sqrt{x}\right)^2.
\end{aligned}$$

On remarque au passage que la dimension associée à chaque $\mathcal{C}$ étant la même, la pénalité correspondante est uniforme en $\mathcal{C}$ (on peut faire plus fin bien sûr).

9. Deux points de vue qui coïncident ici : on peut faire de la sélection de modèle en pénalisant avec la pénalité au-dessus, ou prendre directement la partition qui colle le mieux. C'est à dire on choisit

$$\hat{\mathcal{C}} \in \arg\min_{\mathcal{C}} R_n(\hat{\mu}_{\mathcal{C}}) + pen(\mathcal{C}).$$

La pénalité donnée par la borne d'union étant uniforme sur $\mathcal{P}_k^n$, cela revient à choisir

$$\hat{\mathcal{C}} \in \arg\min_{\mathcal{C}} R_n(\hat{\mu}_{\mathcal{C}}),$$

et à prendre comme prédicteur $\hat{\mu}_k = \hat{\mu}_{\hat{\mathcal{C}}}$. Pour ce choix de $\hat{\mu}_k$, uniformément en le choix de $\mathcal{C}$ on a

$$R(\hat{\mu}_k)-R(\mu^*) \leq R_n(\hat{\mu}_k)-R_n(\mu^*)+\frac{1}{2}(R(\hat{\mu}_k)-R(\mu^*))+\sigma^2 p_{k,n}(x),$$

et donc

$$\begin{aligned}
R(\hat{\mu}_k)-R(\mu^*) &\leq 2\left(R_n(\hat{\mu}_k)-R_n(\mu^*)\right)+2\sigma^2 p_{k,n}(x) \\
&\leq 2\left(R_n(\hat{\mu}_{\hat{\mathcal{C}}})-R_n(\mu^*)\right)+2\sigma^2 p_{k,n}(x) \\
&\leq 2\left(R_n(\mu^*_{\mathcal{C}})-R_n(\mu^*)\right)+2\sigma^2 p_{k,n}(x) \\
&\leq 2\left(R(\mu^*_{\mathcal{C}})-R(\mu^*)+(1/2)\left(R(\mu^*_{\mathcal{C}})-R(\mu^*)\right)+2\sigma^2 p_{k,n}(x)\right)+2\sigma^2 p_{k,n}(x) \\
&\leq 3\left(R(\mu^*_{\mathcal{C}})-R(\mu^*)\right)+6\sigma^2 p_{k,n}(x).
\end{aligned}$$

10. D'après le résultat de la question 6-, on a alors, pour $k \leq n$, avec probabilité plus grande que $1 - e^{-x}$,

$$\begin{aligned}\|\hat{\mu}_k - \mu^*\|^2 &\leq 3\left(\sqrt{n}\frac{V(\mu^*)}{2k}\right)^2 + 6\sigma^2 p_{k,n}(x) \\ &\leq \frac{3nV^2(\mu^*)}{4k^2} + \sigma^2 32 \times 6\left(\sqrt{(k+1)\log(n)} + \sqrt{x}\right)^2 \\ &\leq c_1\frac{nV^2(\mu^*)}{k^2} + c_2 k\sigma^2\log(n) + c_3\sigma^2 x.\end{aligned}$$

Il choisit maintenant de choisir k de manière 'optimale' (possible quand on connaît $V(\mu^*)$ et σ^2. Pour $k = c\frac{n^{1/3}V^{2/3}(\mu^*)}{\sigma^{2/3}\log(n)^{2/3}}$ on obtient le résultat voulu. Au facteur $\log(n)$ près (dû ici à la non-connaissance de la bonne partition), la dimension de l'ensemble des fonctions monotones du point de vue de ce problème d'apprentissage est $n^{1/3}$, et on peut remarquer au passage que c'est la même vitesse/dimension que pour l'ensemble des fonctions Lipschitz.