INTRODUCTION AU MACHINE LEARNING
TD 1 - Corrigé

25/03/2020

Exercice 1

On observe $D_n = ((X_1, Y_1), ..., (X_n, Y_n))$ i.i.d de loi P avec $X_i \in 1, ..., k$ et $Y_i \in 0, 1$. On pose, pour $x \in \{1, ..., k\}$, $N_x = card(\{i : X_i = x\})$. On considère la définition de la fonction

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

et on définit son estimateur:

$$\hat{\eta}(x) = \begin{cases} \frac{1}{N_x} \sum_{i:X_i=x} Y_i & \text{si } N_x > 0 \\ 0 & \text{sinon} \end{cases}$$

et on définit la règle de classification $\hat{f}(D_n, x) = \mathbb{1}_{\hat{\eta}(x) \geq 1/2}$.

1. Démontrer que, pour $x \in \{1, ..., k\}$ fixé:

$$E(|\hat{\eta}(x) - \eta(x)||X_1, ..., X_n) \leq \frac{1}{\sqrt{N_x}} \mathbb{1}_{N_x>0} + \mathbb{1}_{N_x=0}$$

En déduire que

$$E[R(\hat{f}) - R^*] \leq 2E[\frac{1}{\sqrt{N_x}} \mathbb{1}_{N_x>0}] + P(N_X = 0).$$

Solution:

Par l'inegalité de Jensen on a que:

$$E(|\hat{\eta}(x) - \eta(x)||X_1, ..., X_n) \leq \sqrt{E((\hat{\eta}(x) - \eta(x))^2|X_1, ..., X_n)})$$

Par ailleurs on remarque que comme $Y_i \in \{0, 1\}$, alors $E(Y|X = x) = \eta(x) \leq 1$. Ainsi:

- Si $N_x = 0$:
  On a $\hat{\eta}(x) = 0$, ainsi:

  $$E((\hat{\eta}(x)-\eta(x))^2|X_1, ..., X_n) = E(\eta(x)^2|X_1, ..., X_n) \leq Var(Y|X = x) = \eta(x)(1-\eta(x)) \leq \frac{1}{4}$$

  et donc

  $$E(|\hat{\eta}(x) - \eta(x)||X_1, ..., X_n) \leq \frac{1}{2} \leq 1$$

- Si $Nx > 0$:
  On a $\hat{\eta}(x) = \frac{1}{N_x}\sum_{i:X_i=x} Y_i := \bar{Y}|X = x$. Donc $E(\hat{\eta}(x)) = E(\bar{Y}|X = x) = E(Y|X = x) = \eta(x)$, car la moyenne empirique est un estimateur non biaisé de l'espérance. On a donc:

$$\begin{aligned} E((\hat{\eta}(x) - \eta(x))^2|X_1, ..., X_n) &= Var(\hat{\eta}(x)|X_1, ..., X_n) \\ &= \frac{1}{N_x^2} Var(\sum_{i:X_i=x} Y_i|X_1, ..., X_n) \\ &= \frac{1}{N_x^2} N_x Var(Y|X = x) \quad \text{car les } Y_i \text{ sont iid} \\ &= \frac{1}{N_x}\eta(x)(1 - \eta(x)) \quad \text{car les } Y_i \in \{0, 1\} \text{ et } E(Y|X = x) := \eta(x) \\ &\leq \frac{1}{N_x} \times \frac{1}{4} \\ &\leq \frac{1}{N_x} \end{aligned}$$

  Et donc on a que pour $N_x > 0$:

$$\begin{aligned} E(|\hat{\eta}(x) - \eta(x)||X_1, ..., X_n) &\leq \sqrt{E((\hat{\eta}(x) - \eta(x))^2|X_1, ..., X_n)} \\ &\leq \frac{1}{\sqrt{N_x}} \end{aligned}$$

En tout, on a :

$$E(|\hat{\eta}(x) - \eta(x)||X_1, ..., X_n) \leq \frac{1}{\sqrt{N_x}} \mathbb{1}_{N_x>0} + \mathbb{1}_{N_x=0}$$

Et on sait que pour l'excès de risque d'une une règle de classification par plug-in $\hat{f}$, associée à une règle de régression $\hat{\eta}$ (Slides Lecture 2, diapo 4):

$$\begin{aligned} R(\hat{f}) - R^* &= R(\hat{f}) - R(f^*) \\ &\leq 2E(|\hat{\eta}(x) - \eta(x)||X_1, ..., X_n) \\ &\leq 2\left[\frac{1}{\sqrt{N_x}} + \frac{1}{2}\mathbb{1}_{N_x=0}\right] \\ &= 2\frac{1}{\sqrt{N_x}} + \mathbb{1}_{N_x=0} \end{aligned}$$

Et alors:

$$E[R(\hat{f}) - R^*] \leq 2E[\frac{1}{\sqrt{N_x}}\mathbb{1}_{N_x>0}] + P(N_X = 0)$$

2. Démontrer que pour tout x tel que $P(X_1 = x) > 0$ on a $N_x \to \infty$ p.s lorsque $n \to \infty$ puis en déduire que les deux termes du membre de droite dans l'inégalité ci-dessus tendent vers 0.

Solution:

- $N_x \xrightarrow[n\to\infty]{} \infty$ p.s. ?
  En considérant les variables aléatoires $\mathbb{1}(X_i = x)$, on peut montrer par la loi des grands nombres que

$$\frac{N_x}{N} = \sum_{i=1}^{N} \frac{\mathbb{1}(X_i = X)}{N} \xrightarrow[n\to\infty]{} P(X_1 = x) \quad \text{p.s..}$$

  Or, $P(N_x \xrightarrow[n\to\infty]{} \infty) \geq P\big(N_x/N \xrightarrow[n\to\infty]{} P(X_1 = x)\big) = 1$, ce qui implique que $N_x \xrightarrow[n\to\infty]{} \infty$ p.s..

- Alors $2E[\frac{1}{\sqrt{N_x}}\mathbb{1}_{N_x>0}] \xrightarrow[n\to\infty]{} 0$ et $P(N_X = 0) \xrightarrow[n\to\infty]{} 0$?
  Pour tout β positif,

$$E[\frac{1}{\sqrt{N_x}}\mathbb{1}_{N_x>0}] \leq E[\frac{1}{\sqrt{N_x}}] \leq 1 \times P(N_x \leq \beta) + \frac{1}{\sqrt{\beta}} \times P(N_x \geq \beta) \leq P(N_x \leq \beta) + \frac{1}{\sqrt{\beta}}.$$

  On a $N_x \xrightarrow[n\to\infty]{} \infty$ en probabilité, alors pour tout ϵ positif on peut trouver un β tel que $P(N_x \leq \beta) \leq \epsilon/2$ et $1/\sqrt{\beta} \leq \epsilon/2$. On en déduit que pour tout ϵ positif, $E[\frac{1}{\sqrt{N_x}}\mathbb{1}_{N_x>0}] \leq \epsilon$ et donc $2E[\frac{1}{\sqrt{N_x}}\mathbb{1}_{N_x>0}] \xrightarrow[n\to\infty]{} 0$.
  D'ailleurs, $P(N_x = 0) = P(X_1 \neq x)^N$ ce qui converge vers 0 quand N tend vers l'infini.

3. Conclure: démontrer que $\hat{f}$ est universellement consistante.
Solution:

On a

$$E[R(\hat{f}) - R^*] \leq 2E[\frac{1}{\sqrt{N_x}}\mathbb{1}_{N_x>0}] + P(N_X = 0)$$

$$\lim_{n\to\infty} E[R(\hat{f}) - R^*] \leq \lim_{n\to 0}[2E[\frac{1}{\sqrt{N_x}}\mathbb{1}_{N_x>0}] + P(N_X = 0)] = 0$$

et on sait que $R(\hat{f}) - R^* = R(\hat{f}) - R(f^*) \geq 0$ car $f^* = argmin_f R(f)$ par définition (optimal/Bayes predictor).
Ainsi

$$E[R(\hat{f}) - R^*] \xrightarrow[n\to\infty]{} 0$$

pour toute loi P sur $\mathcal{X} \times \mathcal{Y}$, et comme $R(\hat{f})$ converge en distribution vers un constant R^*, $R(\hat{f})$ converge également en probabilité vers R^*, ce qui montre que $\hat{f}$ est universellement consistante.

Exercice 2

On considère une variable aléatoire X tirée selon une loi P_θ , où le paramètre réel θ suit quant à lui une distribution π. Pour estimer θ, on considère la fonction de perte:

$$L(\theta, a) = \begin{cases} k_2(\theta - a) & si \quad \theta > a \\ k_1(a - \theta) & sinon \end{cases}$$

Montrer que l'estimateur de Bayes est un fractile (à déterminer) de la loi de θ sachant X.

Solution:

Rappel: z est un fractile d'ordre k d'une loi de probabilité P_z, si $P(Z < z) = k$.
On a $X|\theta \sim P_\theta$, et $\theta \sim \pi$.
L'estimateur de Bayes est l'argument a qui minimise

$$E(L(\theta, a)|X) = \int_\Theta L(\theta, a) dP_{\theta|X=x}$$

On suppose que la loi de probabilité $P_{\theta|X=x}$ admet une densité $p(\theta|x)$ par rapport à la mesure de Lebesgue. Ainsi:

$$\frac{\partial}{\partial a}\int_\Theta L(\theta, a) dP_{\theta|X=x} = 0$$

$$
\begin{aligned}
&\iff \frac{\partial}{\partial a}\left[\int_a^{+\infty} k_2(\theta-a)dP_{\theta|X=x} + \int_{-\infty}^a k_1(a-\theta)dP_{\theta|X=x}\right] = 0\\
&\iff \frac{\partial}{\partial a}\left[\int_{-\infty}^{+\infty} k_2(\theta-a)dP_{\theta|X=x} + \int_{-\infty}^a (k_1+k_2)(a-\theta)dP_{\theta|X=x}\right] = 0\\
&\iff \frac{\partial}{\partial a}\left[\int_{-\infty}^a (k_1+k_2)(a-\theta)dP_{\theta|X=x}\right] = k_2\\
&\iff \frac{\partial}{\partial a}\left[\int_{-\infty}^a (a-\theta)dP_{\theta|X=x}\right] = \frac{k_2}{k_1+k_2}\\
&\iff P_{\theta|X=x}(\theta<a) + a\frac{\partial}{\partial a}\int_{-\infty}^a p(\theta|x)d\theta - \frac{\partial}{\partial a}\int_{-\infty}^a \theta p(\theta|x)d\theta = \frac{k_2}{k_1+k_2}
\end{aligned}
$$

En supposant que $\int_{-\infty}^a \theta p(\theta|x)d\theta$ converge:

$$
\begin{aligned}
&P_{\theta|X=x}(\theta<a) + ap(a|x) - ap(a|x) = \frac{k_2}{k_1+k_2}\\
&\iff P_{\theta|X=x}(\theta<a) = \frac{k_2}{k_1+k_2}
\end{aligned}
$$

Donc l'estimateur de Bayes a est le fractile $\frac{k_2}{k_1+k_2}$ de la loi de $\theta|X$.

Exercice 3

On suppose que (X,Y) est un couple de la loi $\mathbb{P}$ déterminée par

- $Y \sim Be(p)$ est une Bernoulli
- pour $y \in \{0,1\}, X|Y=y$ a pour densité $f_y(.)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$

1. Exprimer la fonction de régression $\eta(x) = \mathbb{P}\{Y=1|X=x\}$ en fonction de f_0, f_1 et p
Solution :

$$
\begin{aligned}
\eta(x) = \mathbb{P}\{Y=1|X=x\} &= \frac{f_{X|Y=1}(x)\mathbb{P}\{Y=1\}}{f_X(x)}\\
&= \frac{f_{X|Y=1}(x)\mathbb{P}\{Y=1\}}{f_{X|Y=1}(x)\mathbb{P}\{Y=1\} + f_{X|Y=0}(x)\mathbb{P}\{Y=0\}}\\
&= \frac{f_0(x)p}{f_0(x)p + f_1(x)(1-p)}
\end{aligned}
$$

2. En déduire le classifieur optimal f^*
Solution : D'après le cours on sait que $f^*(x) = \mathbb{1}(\eta(x) \geq \frac{1}{2})$, et d'où

$$f^*(x) = \mathbb{1}(f_1(x)p \geq f_0(x)(1-p)).$$

3. Déterminer f^* et son risque $L^* = \mathbb{P}\{Y \neq f^*(X)\}$ dans les cas suivants

a) $f_0(x) = \frac{1}{\theta_0}\mathbb{1}\{x \in [0,\theta_0]\}$ et $f_1(x) = \frac{1}{\theta_1}\mathbb{1}\{x \in [0,\theta_1]\}$ pour $\theta_1 > \theta_0 > 0$

Solution : De manière générale, nous avons

$$L^* = \mathbb{P}\{Y \neq f^*(X)\} = \mathbb{E}[\mathbb{1}\{Y \neq f^*(x)\}] = \mathbb{E}\big[\mathbb{E}[\mathbb{1}\{Y \neq f^*(x)\}|X]\big],$$

et d'ailleurs

$$
\begin{aligned}
\mathbb{P}\{Y \neq f^*(X)|X\} &= \mathbb{P}\{Y=0, f^*(X)=1|X\} + \mathbb{P}\{Y=1, f^*(X)=0|X\}\\
&= \mathbb{1}\{f^*(x)=1\}\mathbb{P}\{Y=0|X\} + \mathbb{1}\{f^*(x)=0\}\mathbb{P}\{Y=1|X\}\\
&= \mathbb{1}\{f^*(x)=1\}(1-\eta(x)) + \mathbb{1}\{f^*(x)=0\}\eta(x).
\end{aligned}
$$

Enfin

$$\begin{aligned}
\mathbb{P}\{Y \neq f^*(X)\} &= \mathbb{E}\big[\mathbb{1}\{f^*(x)=1\}(1-\eta(x)) + \mathbb{1}\{f^*(x)=0\}\eta(x)\big] \\
&= \mathbb{E}\big[\min\big(\eta(x), 1-\eta(x)\big)\big] \\
&= \int \min(\big(\eta(x), 1-\eta(x)\big) f_X(x)dx \\
&= \int \min\Big(\frac{f_0(x)p}{f_X(x)}, \frac{f_1(x)(1-p)}{f_X(x)}\Big) f_X(x)dx \\
&= \int \min\big(f_0(x)p, f_1(x)(1-p)\big)dx.
\end{aligned}$$

En l'appliquant sur (a) on obtient

$$\begin{aligned}
L^* &= \int_0^{\theta_1} \min\big(f_0(x)p, f_1(x)(1-p)\big)dx \\
&= \int_0^{\theta_1} \min\Big(\frac{p}{\theta_0}, \frac{1-p}{\theta_1}\Big)dx \\
&= \int_0^{\theta_0} \min\Big(\frac{p}{\theta_0}, \frac{1-p}{\theta_1}\Big)dx + \int_{\theta_0}^{\theta_1} \min\Big(0, \frac{1-p}{\theta_1}\Big)dx \\
&= \int_0^{\theta_0} \min\Big(\frac{p}{\theta_0}, \frac{1-p}{\theta_1}\Big)dx \\
&= \theta_0 \min\Big(\frac{p}{\theta_0}, \frac{1-p}{\theta_1}\Big) \\
&= \min(p, (\theta_0/\theta_1)(1-p)).
\end{aligned}$$

b) $f_0(x) = \frac{1}{\theta_0}\mathbb{1}\big\{x \in [0,\theta_0]\big\}$ et $f_1(x) = \frac{1}{\theta_1-\theta_0}\mathbb{1}\big\{x \in [\theta_0,\theta_1]\big\}$ pour $\theta_1 > \theta_0 > 0$

Solution :

$$\begin{aligned}
L^* &= \int_0^{\theta_1} \min\big(f_0(x)p, f_1(x)(1-p)\big)dx \\
&= \int_0^{\theta_0} \min\Big(\frac{p}{\theta_0}, 0)dx + \int_{\theta_0}^{\theta_1} \min\Big(0, \frac{1-p}{\theta_1-\theta_0}\Big)dx \\
&= 0.
\end{aligned}$$

c) $f_0(x) = \theta_0 e^{-\theta_0 x}\mathbb{1}\big\{x \in \mathbb{R}_+\big\}$ et $f_1(x) = \theta_1 e^{-\theta_1 x}\mathbb{1}\big\{x \in \mathbb{R}_+\big\}$ pour $\theta_1 > \theta_0 > 0$

Solution: On peut montrer qu'on a $f_0(x)p > f_1(x)(1-p)$ ssi $x > \frac{\log\big((1-p)\theta_1/p\theta_0\big)}{\theta_1-\theta_0}$. Alors en posant $a = \frac{\log\big((1-p)\theta_1/p\theta_0\big)}{\theta_1-\theta_0}$, on obtient

$$\begin{aligned}
L^* &= \int_0^a f_0(x)pdx + \int_a^\infty f_1(x)(1-p)dx \\
&= p - pe^{-a\theta_0} + (1-p)e^{-a\theta_1}.
\end{aligned}$$

d) $f_0(x) = \frac{1}{\sqrt{2\pi\sigma_0^2}}e^{-\frac{(x-\theta_0)^2}{2\sigma_0^2}}$ et $f_1(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{(x-\theta_1)^2}{2\sigma_1^2}}$ pour $\theta_1 > \theta_0 > 0$ et $\sigma_0, \sigma_1 > 0$

Solution :