Dans la suite, on va noter $\eta(x) = \mathbb{P}(Y=1 \mid X=x)$

<u>Problème 1</u> : l'erreur de Bayes (risque associé à la perte binaire)

$$R^* = \inf_{g : \mathcal{X} \to \{-1,+1\}} \mathbb{P}[g(X) \neq Y] .$$

Montrons l'équivalence entre les deux définitions de R^* :

a) $R^* = \mathbb{E}\left[\min\left(\eta(X), 1-\eta(X)\right)\right]$

b) $R^* = \frac{1}{2} - \frac{1}{2}\mathbb{E}\left[|2\eta(X) - 1|\right]$.

c) Supposons que pour $y \in \{-1,+1\}$, les variables $(X \mid Y=y)$ ont f_+ et f_- comme densités respectives pour $y=-1$ et $y=+1$, et $\mathbb{P}(Y=1) = \mathbb{P}(Y=-1) = 1/2$. Montrons que :

$$R^* = \frac{1}{2} - \frac{1}{4}\int |f_+(x) - f_-(x)|\, dx .$$

<u>Solutions :</u>

(a) Rappelons que $g^*(x) = \begin{cases} +1 & \text{si } \eta(x) \geq \frac{1}{2} \\ -1 & \text{sinon} \end{cases}$

i.e $R^* = R(g^*) = \mathbb{P}(g^*(X) \neq Y)$.

Nous avons :

$$\mathbb{P}(g^*(X) \neq Y \mid X=x) = \mathbb{P}(g^*(X) = +1, Y=-1 \mid X=x) + \mathbb{P}(g^*(X) = -1, Y=+1 \mid X=x)$$

②

$$= (1-\eta(x))\mathbb{1}\{g^*(x)=+1\} + \eta(x)\mathbb{1}\{g^*(x)=-1\}$$
$$= (1-\eta(x))\mathbb{1}\{\eta(x)\geq 1/2\} + \eta(x)\mathbb{1}\{\eta(x)<1/2\}$$
$$= (1-\eta(x))\mathbb{1}\{\eta(x)\geq 1-\eta(x)\} + \eta(x)\mathbb{1}\{\eta(x)<1-\eta(x)\}$$
$$= \min(\eta(x), 1-\eta(x)).$$

Espérance par rapport à la loi de X donne :

$$R^* = \mathbb{E}\left[\min(\eta(X), 1-\eta(X))\right]$$

b) Nous avons :

$$\min(\eta(x), 1-\eta(x)) = \frac{1}{2} - \left|\eta(x) - \frac{1}{2}\right|$$

$$\begin{cases} \text{si } \eta(x) \geq 1-\eta(x) \iff \frac{1}{2} - \left|\eta(x)-\frac{1}{2}\right| = \frac{1}{2} - \left(\eta(x)-\frac{1}{2}\right) = 1-\eta(x) \\ \text{si } \eta(x) < 1-\eta(x) \iff \frac{1}{2} - \left|\eta(x)-\frac{1}{2}\right| = \frac{1}{2} - \left(\frac{1}{2}-\eta(x)\right) = \eta(x). \end{cases}$$

$$R^* = \mathbb{E}\left[\min(\eta(X), 1-\eta(X))\right] = \mathbb{E}\left[\frac{1}{2} - \left|\eta(X)-\frac{1}{2}\right|\right] = \frac{1}{2} - \frac{1}{2}\mathbb{E}\left[|2\eta(X)-1|\right]$$

(c) Théorème de Bayes :

$$\eta(x) = \mathbb{P}(Y=1 \mid X=x) = \frac{\mathbb{P}(Y=1)f_+(x)}{\mathbb{P}(Y=1)f_+(x) + \mathbb{P}(Y=-1)f_-(x)} = \frac{f_+(x)}{f_+(x)+f_-(x)}$$

donc :
$$2\eta(x) - 1 = \frac{2f_+(x)}{f_+(x)+f_-(x)} - 1 = \frac{f_+(x) - f_-(x)}{f_+(x)+f_-(x)}$$

$$R^* = \frac{1}{2} - \frac{1}{2}\int \left|\frac{f_+(x)-f_-(x)}{f_+(x)+f_-(x)}\right| \left(\frac{1}{2}f_+(x) + \frac{1}{2}f_-(x)\right)dx = \frac{1}{2} - \frac{1}{4}\int |f_+(x) - f_-(x)|\,dx$$

Problème 2 : La divergence de Kullback-Leibler entre deux vecteurs aléatoires X_1 et X_2 de densités f_1 et f_2 sur un support commun $S \subseteq \mathbb{R}^d$ est donnée par :

$$D(f_1 \| f_2) = \int_S f_1(x) \log \frac{f_1(x)}{f_2(x)} \, dx.$$

a) Calculer la divergence de Kullback-Leibler pour $X_i \sim \mathcal{N}(\mu_i, \Sigma_i)$ pour $i = 1, 2$.

b) Supposons que pour $y \in \{-1, +1\}$, $(X|Y=y) \sim \mathcal{N}(\mu_y, \Sigma)$ et $\mathbb{P}(Y=+1) = p$. Montrer que

$$\mathbb{E}\left[\sqrt{\eta(X)(1-\eta(X))}\right] = p(1-p) \exp\left(-\frac{1}{4} D(\mu_{+1} \| \mu_{-1})\right)$$

où $D(\mu_{+1} \| \mu_{-1})$ est la divergence de Kullback-Leibler entre $\mathcal{N}(\mu_{+1}, \Sigma)$ et $\mathcal{N}(\mu_{-1}, \Sigma)$.

Solution :

(a) Supposons : $X \sim f_1 = \mathcal{N}(\mu_1, \Sigma_1)$, nous avons :

$$\mathbb{E}_{f_1}\left[(X-\mu_1)(X-\mu_1)^T\right] = \Sigma_1 \Rightarrow \Sigma_1^{-1} \mathbb{E}_{f_1}\left((X-\mu_1)(X-\mu_1)^T\right) = \Sigma_1^{-1}\Sigma_1 = I_d$$

$$\Rightarrow \mathbb{E}_{f_1}\left[\Sigma_1^{-1}(X-\mu_1)(X-\mu_1)^T\right] = I_d \Rightarrow tr\left(\mathbb{E}_{f_1}\left[\Sigma_1^{-1}(X-\mu_1)(X-\mu_1)^T\right]\right) = tr(I_d) = d$$

donc : $\mathbb{E}_{f_1}\left[(X-\mu_1)^T \Sigma_1^{-1} (X-\mu_1)\right] = d.$

De manière similaire ; $\mathbb{E}_{f_1}\left[(X-\mu_1)^T \Sigma_2^{-1} (X-\mu_1)\right] = tr(\Sigma_2^{-1}\Sigma_1)$

Comme: $(X-\mu_2)^T \Sigma_2^{-1}(X-\mu_2) = \left((X-\mu_1)+(\mu_1-\mu_2)\right)^T \Sigma_2^{-1}\left((X-\mu_1)+(\mu_1-\mu_2)\right)$

$$= (X-\mu_1)^T \Sigma_2^{-1}(X-\mu_1) + 2(\mu_1-\mu_2)^T \Sigma_2^{-1}(X-\mu_1) + (\mu_1-\mu_2)^T \Sigma_2^{-1}(\mu_1-\mu_2),$$

On passe à l'espérance:

$$\mathbb{E}_{f_1}\left[(X-\mu_2)^T \Sigma_2^{-1}(X-\mu_2)\right] = tr\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_1-\mu_2)^T \Sigma_2^{-1}(\mu_2-\mu_2)$$

On note que $D(f_1 \| f_2)$ peut être écrite comme une espérance

$$D(f_1\|f_2) = \mathbb{E}_{f_1}\left[\log \frac{f_1(X)}{f_2(X)}\right],$$ où $f_1 = \mathcal{N}(\mu_1, \Sigma_1)$

$f_1 = \mathcal{N}(\mu_1, \Sigma_1)$ et $f_2 = \mathcal{N}(\mu_2, \Sigma_2)$

$$\mathbb{E}_{f_1}\left[\log \frac{f_1(X)}{f_2(X)}\right] = \mathbb{E}_{f_1}\left[\log \frac{|\Sigma_2|^{1/2} \exp\left\{-\frac{1}{2}(X-\mu_1)^T \Sigma_1^{-1}(X-\mu_1)\right\}}{|\Sigma_1|^{1/2} \exp\left\{-\frac{1}{2}(X-\mu_2)^T \Sigma_2^{-1}(X-\mu_2)\right\}}\right]$$

$$= \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} + \frac{1}{2}\mathbb{E}_{f_1}\left[(X-\mu_2)^T \Sigma_2^{-1}(X-\mu_2) - (X-\mu_1)^T \Sigma_1^{-1}(X-\mu_1)\right]$$

$$= \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_2|} + tr\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_1-\mu_2)\Sigma_2^{-1}(\mu_1-\mu_2) - d\right]$$

(b) On note $f_+ = \mathcal{N}(\mu_+, \Sigma)$ et $f_- = \mathcal{N}(\mu_-, \Sigma)$, alors:

$f = p f_+ + (1-p) f_+$; densité marginale en X.

$$\mathbb{E}\left(\sqrt{\eta(X)(1-\eta(X))}\right) = \int \sqrt{\eta(x)(1-\eta(x))}\, f(x)\, dx$$

$$= \int \sqrt{\frac{p f_+(x)}{f(x)} \frac{(1-p) f_-(x)}{f(x)}}\, f(x)\, dx = \sqrt{p(1-p)} \int \sqrt{f_+(x) f_-(x)}\, dx$$

$$= \sqrt{p(1-p)} \int \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{4} h(x)\right\} dx$$

où $h(x) = (x-\mu_{+1})^T \Sigma^{-1} (x-\mu_{+1}) + (x-\mu_{-1})^T \Sigma^{-1} (x-\mu_{-1})$

on note $\bar{\mu} = \dfrac{(\mu_{+1}+\mu_{-1})}{2}$

$$h(x) = 2(x-\mu)^T \Sigma^{-1} (x-\bar{\mu}) + \frac{1}{2}(\mu_{+1}-\mu_{-})^T \Sigma^{-1} (\mu_{+1}+\mu_{-1})$$
$$= 2(x-\bar{\mu}) \Sigma^{-1} (x-\bar{\mu}) + D(\mu_{+1} \| \mu_{-1})$$

on obtient :

$$\mathbb{E}\left(\sqrt{\eta(x)(1-\eta(x))}\right) = \sqrt{p(1-p)} \exp\left\{-\frac{1}{4} D(\mu_{+1} \| \mu_{-1})\right\}$$
$$\int \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\bar{\mu})^T \Sigma^{-1} (x-\bar{\mu})\right\} dx$$
$$= \sqrt{p(1-p)} \exp\left\{-\frac{1}{4} D(\mu_{+1} \| \mu_{-1})\right\}.$$

Problème ③ : Dans cet exercice, nous explorons quelques questions liées aux données de grande dimension.
Pour $i = 1, \dots, n$, soit $X^{(i)} \in \mathbb{R}^d$ des vecteurs aléatoires i.i.d de loi uniforme $[0,1]^d$. Soit X une nouvelle observation, on définit la espérance de la distance à la donnée la plus proche

$$\delta_\infty(d,n) = \mathbb{E}\left[\min_{i=1,\dots,n} \|X^{(i)} - X\|_\infty\right].$$

L'objectif de cet exercice est de comprendre le comportement de $\delta_\infty(d,n)$ lorsque n et d sont grands.

(a) Montrer que pour tout $t > 0$

$$\mathbb{P}\left[\min_{i=1,\dots,n} \|X^{(i)} - X\|_\infty > t\right] \geq 1 - n(2t)^d.$$

(b) Supposons qu'on veut être sûr que X est une distance $1/4$ à au moins une donnée $X^{(1)}, X^{(2)}, \dots, X^{(n)}$ avec une probabilité d'au moins $1/2$. Donnez la borne inférieure de la taille du jeu de données n en fonction d.

$$\mathbb{P}\left[\min_{i=1,\dots,n} \|X^{(i)} - X\|_\infty > t\right] = \mathbb{P}\left[\bigcap_{i=1}^{n} \left\{\|X^{(i)} - X\|_\infty > t\right\}\right]$$

$$= 1 - \mathbb{P}\left[\overline{\bigcap_{i=1}^{n} \left\{\|X^{(i)} - X\|_\infty > t\right\}}\right] = 1 - \mathbb{P}\left[\bigcup_{i=1}^{n} \overline{\left\{\|X^{(i)} - X\|_\infty > t\right\}}\right]$$

$$= 1 - \mathbb{P}\left(\bigcup_{i=1}^{n} \left\{\|X^{(i)} - X\|_\infty \leq t\right\}\right)$$

$$\geq 1 - \sum_{i=1}^{n} \mathbb{P}\left(\left\{\|X^{(i)} - X\|_\infty \leq t\right\}\right)$$

$$= 1 - n\,\mathbb{P}\left(\left\{\|X^{(1)} - X\|_\infty \leq t\right\}\right)$$

$$= 1 - n\,\mathbb{P}\left[\bigcap_{j=1,\dots,d} \left\{|X_j^{(1)} - X_j| \leq t\right\}\right]$$

$$= 1 - n\left(\mathbb{P}\left[|X_1^{(1)} - X_1| \leq t\right]\right)^d$$

$$= 1 - n\left(1 - (1-t)^2\right)^d$$

$$= 1 - n\left(2t - t^2\right)^d$$

$$\geq 1 - n(2t)^d \quad \text{si } 0 < t < 1.$$

(b) On s'intéresse : $\mathbb{P}\left[\min_{i=1,\dots,n} \|X^{(i)} - X\|_\infty < 1/4\right]$.

$$\mathbb{P}\left[\min_{i=1,\dots,n} \|X^{(i)} - X\|_\infty \leq 1/4\right] = 1 - \mathbb{P}\left[\min_{i=1,\dots,n} \|X^{(i)} - X\|_\infty > \tfrac{1}{4}\right]$$

$$\leq 1 - \left(1 - n\left(\tfrac{1}{2}\right)^d\right)$$

$$= n\left(\tfrac{1}{2}\right)^d.$$

pour garantir que cette probabilité $\geq \frac{1}{2}$, on doit avoir

$$\frac{n}{2^d} \geq \frac{1}{2} \iff n \geq 2^{d-1}.$$

(c) En utilisant (a), montrer que $\rho_\infty(d,n) \geq \dfrac{d}{2(d+1)} n^{-\frac{1}{d}}$.

Hint : $\left[Z \geq 0 ;\ \mathbb{E}(Z) = \int_0^{+\infty} \mathbb{P}(Z > t)\, dt\right]$

$$\rho_\infty(d,n) = \int_0^{+\infty} \mathbb{P}\left[\min_{1\leq i\leq n} \|X^{(i)} - X\|_\infty > t\right] dt$$

$$\geq \int_0^{+\infty} \left(1 - n(2t)^d\right)_+ dt = \int_0^{\frac{1}{2}n^{-1/d}} \left(1 - n(2t)^d\right) dt$$

$$= \boxed{\frac{d}{2(d+1)} n^{-1/d}}$$

(d) Calculer cette quantité pour $d \in \{2, 10, 20\}$ et $n \in \{100, 1000, 10000, 100000\}$