

Éléments de correction de l'examen  
2019 - 2020

Exercice 1 :

1. a)  $L^*$  est le risque associé à la règle de décision optimale  $f^*$  parmi les fonctions  $f: X \rightarrow Y$ .  $f^*$  est appelé prédicteur de Bayes et  $L^*$ , risque de Bayes. C'est un risque "indépassable".

b) Remontrons que  $f^*$  existe (i.e. qu'il existe un minimum) et déterminons sa valeur.

$$\begin{aligned}
 P(f(X) \neq Y) &= \sum_{y=1}^K P(f(X)=y, Y \neq y) \\
 &= \sum_{y=1}^K E[1_{\{f(X)=y\}} 1_{\{Y \neq y\}}] \\
 &= \sum_{y=1}^K E[P(Y \neq y | X). 1_{\{f(X)=y\}}]
 \end{aligned}$$

$$= \sum_{y=1}^K \int_X (1 - P(Y=y | X=x)) \cdot \underbrace{1_{\{f(x)=y\}}}_{(*)} P_X(dx)$$

Or,  $1 - P(Y=y | X=x) \geq 1 - \max_{y=1}^k P(Y=y | X=x)$   
 $\quad \quad \quad =: \eta(x)$

$$\Rightarrow P(f(x) \neq Y) \geq \int_X (1 - \eta(x)) \underbrace{\sum_{y=1}^K 1_{\{f(x)=y\}}}_{=1} P_X(dx)$$

$$\geq E[1 - \eta(x)]$$

On a donc  $L^+ \geq 1 - E[\eta(x)]$ . Pour montrer l'égalité (et prouver qu'il s'agit d'un minimum), on doit trouver une fonction  $f^+$  tq:  $R_{f^+} = 1 - E[\eta(x)]$ . Or, si on pose

$$f^*(x) = \operatorname{Argmax}_{y=1}^K P(Y=y | X=x),$$

on a par construction

$$f(x) = y \Leftrightarrow P(X=x | Y=y) = \eta(x)$$
$$\Leftrightarrow 1 - P(X=x | Y=y) = 1 - \eta(x)$$

de sorte qu'en reprenant (4), on a bien

$$P(f^*(X) = y) = \sum_{y=1}^k \int_X (1 - \eta(x)) \mathbb{1}_{\{f^*(x) = y\}} d\mu_x$$
$$= E[1 - \eta(X)].$$

2. (a) Si  $\mathcal{G}$  augmente  $\inf_{g \in \mathcal{G}} L(g)$  diminue

$\Rightarrow \mathcal{E}_2$  diminue.

( $\mathcal{E}_2$  est tjs positif car  $\mathcal{G} \subset \mathcal{F}$ )

$$\Rightarrow \inf_{g \in \mathcal{G}} L(g) \geq \inf_{f \in \mathcal{F}} L(f).$$

(b) Pour tout  $g$ , on a par la loi de

grands nombres

$$\lim_{n \rightarrow +\infty} L_n(g) = \mathbb{E} [l(g(x), y)]$$

$$\Rightarrow \inf_{g \in \mathcal{G}} \lim_{n \rightarrow +\infty} L_n(g) = \inf_{g \in \mathcal{G}} L(g)$$

(1) comme  $L$  est fini,

$$\begin{aligned} \inf_{g \in \mathcal{G}} \lim_{n \rightarrow +\infty} L_n(g) &= \min_{g \in \mathcal{G}} \lim_{n \rightarrow +\infty} L_n(g) = \lim_{n \rightarrow +\infty} \min_{g \in \mathcal{G}} L_n(g) \\ &= \lim_{n \rightarrow +\infty} L_n(g_n) \text{ (par définition)} \end{aligned}$$

$$\Rightarrow \lim_{n \rightarrow +\infty} L_n(g_n) = \inf_{g \in \mathcal{G}} L(g).$$

$$\Rightarrow \lim_{n \rightarrow +\infty} \mathcal{L}_1 = 0$$

(c) Dans ce cas,  $g^{(n)} = \langle \beta, x \rangle$

$$\hat{\beta}_n = \underset{\beta \in \mathbb{R}^p}{\operatorname{Argmin}} \sum_{i=1}^n (y_i - \langle \beta, x_i \rangle)^2$$

et  $g_n^*(x) = \langle \hat{\beta}_n, x \rangle$ . On rappelle que

que l'on pourra  $\mathbf{x} = \left( \begin{array}{c} \frac{x_1}{i} \\ \vdots \\ \frac{x_n}{i} \end{array} \right)$ ,  
 $\mathbf{y} = \left( \begin{array}{c} y_1 \\ \vdots \\ y_n \end{array} \right)$  et que  $\mathbf{x}^\top \cdot \mathbf{x}$  est inversible alors

$$\hat{\beta}_n = (\mathbf{x}^\top \cdot \mathbf{x})^{-1} \cdot \mathbf{x}^\top \cdot \mathbf{y} \text{ de sorte que}$$

$$\text{que } \langle \hat{\beta}_n, \mathbf{x} \rangle = \langle \mathbf{x}, \hat{\beta}_n \rangle = \mathbf{x}^\top \cdot (\mathbf{x}^\top \cdot \mathbf{x})^{-1} \mathbf{x}^\top \cdot \mathbf{y}.$$

$$\inf_{g \in \mathcal{G}} L(g) = \underset{\beta \in \mathbb{R}^p}{\operatorname{Argmin}} \underset{\epsilon \in \mathbb{R}}{\mathbb{E}} \left[ \underset{\epsilon' \in \mathbb{R}^p}{\left( Y - \langle \beta, \mathbf{x} \rangle \right)^2} \right]$$

La encore, cette quantité est calculable. Posons

$$F(\beta) = \mathbb{E}[(Y - \mathbf{x} \cdot \beta)^2], \text{ où } \mathbf{x} = (x^{(1)}, \dots, x^{(n)}).$$

$$\triangleright F(\beta) = \mathbb{E} \left[ \mathbf{x}^\top \cdot (\mathbf{x} \cdot \beta - Y) \right]$$

$$= \mathbb{E} \left[ \underbrace{\mathbf{x}^\top \mathbf{x}}_{p \times p} \beta - \mathbf{x}^\top \cdot Y \right]$$

$$\Rightarrow \triangleright F(\beta) = 0 \text{ssi } \beta^* = \left( \mathbb{E}[\mathbf{x}^\top \mathbf{x}] \right)^{-1} \cdot \mathbb{E}[\mathbf{x}^\top \cdot Y]$$

$$\Rightarrow \inf_{g \in \mathcal{G}} L(g) \text{ est atteint en } g^*(\mathbf{x}) = \langle \beta^*, \mathbf{x} \rangle.$$

Complément

(d) Dans ce cadre, les fonctions  $g$  privilégiées sont les fonctions  $g(x) = \langle p, x \rangle$  avec  $p$  "sparsé".

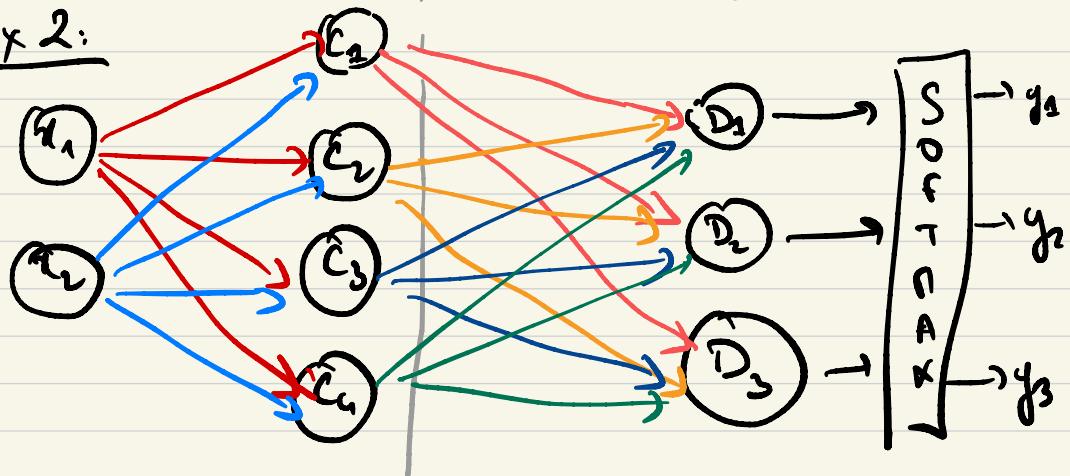
(e) Augmentation de la profondeur  $\Rightarrow$  augmentation de la taille  $P_2 \Rightarrow E_2$  diminuée.

En revanche,  $E_1$  augmente "moralement" car on a plus de flexibilité donc besoin de plus d'observations pour bien estimer le modèle optimal.

(f) Dans le cadre des réseaux de neurones, on utilise des descentes de gradient stochastique pour estimer  $g_\theta$  (avec calcul du gradient par "rétropropagation").  
Plus précisément, on utilise des variantes de la descente de gradient stochastique telles que ADAGR ou RMS PRO.

$$\text{Relu: } \phi(x) = \max(0, x)$$

Ex 2:



Précisions: 'sgd': descente de gradient stochastique.

. 'Dropout' sur la première couche : un neurone aleatoirement les neurones de la couche cachée avec proba 0,6.

. forme de la fonction finale:

$$f(x) = \text{Softmax}\left(\left(\langle v_j, (\phi(\langle w_i, x \rangle + b_i)) \rangle_{i=1}^n + b_j\right)\right)$$

$$\text{où } v_j \in \mathbb{R}^4, w_i \in \mathbb{R}^2, b_i, b_j \in \mathbb{R}, j \in \{1, 2, 3\}$$

pas indispensable

2. Il s'agit d'un modèle de machine à vecteur support à noyau polynomial de degré 2.

La fonction polynomiale s'écrit  $k(x, x') = (b + \gamma \langle x, x' \rangle)^2$   
 Ici,  $\gamma = 0,5$ . Enfin, "cost" correspond à la contrainte " $c$ " de flexibilité du cours.

### Ejercicio 3:

1. Son gradient barien défini car  $L$  convexe.

$$\begin{aligned} 2. \quad \partial_{\theta_1} L(\theta) &= \frac{1}{N} \sum_{k=1}^N x_k (\theta_1 x_k + \theta_2 - y_k) + \lambda \theta_1 \\ &\quad + \frac{\lambda}{2} \begin{cases} \operatorname{sgn}(\theta_1) & \text{si } \theta_1 \neq 0 \\ [-1, 1] & \text{si } \theta_1 = 0. \end{cases} \end{aligned}$$

$$\begin{aligned} \partial_{\theta_2} L(\theta) &= \frac{1}{N} \sum_{k=1}^N (\theta_1 x_k + \theta_2 - y_k) + \lambda \theta_2 \\ &\quad + \frac{\lambda}{2} \begin{cases} \operatorname{sgn}(\theta_2) & \text{si } \theta_2 \neq 0 \\ [-1, 1] & \text{si } \theta_2 = 0. \end{cases} \end{aligned}$$

$$3. \quad \partial_{\theta_1} L(\theta) = 0 \iff$$

$$\theta_1 \left( \frac{1}{N} \sum_{k=1}^N x_k^2 + \lambda \right) + \theta_2 \left( \frac{1}{N} \sum_{k=1}^N x_k \right) - \frac{1}{N} \sum x_k y_k$$

$$\begin{cases} = \lambda_1 \operatorname{sgn}(\theta_1) & \text{si } \theta_1 \neq 0 \\ \in [-\lambda_1, \lambda_2] & \text{si } \theta_1 = 0 \end{cases}$$

$$\partial_{\theta_2} L(\theta) = 0$$

$$\Leftrightarrow \theta_2(1+\lambda) + \theta_4 \left( \frac{1}{N} \sum x_n \right) - \frac{1}{N} \sum y_n$$

$$\begin{cases} \in \frac{1}{2} \operatorname{sgn}(\theta_2) & \text{si } \theta_2 \neq 0 \\ \in [-\lambda_1, \lambda_1] & \text{si } \theta_2 = 0. \end{cases}$$

Un point  $\theta = (\theta_1, \theta_2)$  est point critique de  $L$  ssi il est solution du système  $\begin{cases} \partial_{\theta_1} L(\theta) = 0 & \text{d'après ci-dessus.} \\ \partial_{\theta_2} L(\theta) = 0 \end{cases}$

Exercice 4:

Corrigé dans le TD 1.

Exercice 5:

Corrigé dans le TD 1.

Exercice 6 :

Corrigé dans le TD 5.

$$\begin{aligned} \hookrightarrow D(P||Q) &= \sum_{k=1}^m \frac{P(k)}{Q(k)} \log \left( \frac{P(k)}{Q(k)} \right) \cdot Q(k) \\ &= \sum_{k=1}^m \phi \left( \frac{P(k)}{Q(k)} \right) \cdot Q(k) \end{aligned}$$

où  $\phi(x) = x \log x$ . Formellement,

$$D(P||Q) = E_Q \left[ \phi\left(\frac{P}{Q}\right) \right]$$

Par l'inégalité de Jensen, il vient

$$D(P||Q) \geq \phi\left(E_Q\left(\frac{P}{Q}\right)\right)$$

$$\geq \phi\left(\sum_{k=1}^m \frac{P(k)}{Q(k)} \cdot Q(k)\right)$$

$$\geq \phi(1) = 0.$$

Comme  $x \rightarrow x \log x$  est STRICTEMENT CONVexe,  
l'inégalité est stricte dès que  $k \rightarrow \frac{P(k)}{Q(k)}$  n'est  
pas constante  $\Rightarrow \{ D(P||Q) = 0 \iff P = Q \}$ .