Examen de
Machine learning
Cycle pluridisciplinaire d'études supérieures
Université Paris sciences et lettres

mercredi 2 juin 2021

Les exercices sont indépendants. Toute affirmation devra être justifiée, à l'exception du premier exercice. Il n'est pas interdit d'admettre certains éléments de démonstration (voire des questions entières) afin de ne pas rester bloqué, mais cela doit être mentionné. Il n'est pas nécessaire de traiter l'ensemble du sujet pour obtenir une bonne note.

Exercice 1 (*Questions de cours*). — Répondre aux questions sans justification.

1) Rappeler la définition de la minimisation du risque empirique. Donner un exemple vu dans le cours.
2) Rappeler la définition de la minimisation du risque empirique régularisée et donner un exemple vu dans le cours.
3) Parmi les algorithmes suivants, indiquer ceux qui peuvent s'écrire comme une minimisation du risque empirique (éventuellement régularisée) : k plus proches voisins, Soft-SVM, Hard-SVM, régression linéaire aux moindres carrés, régression logistique.
4) Lorsqu'on utilise un algorithme de minimisation du risque empirique régularisée, comment peut-on détecter une situation de sous-apprentissage (resp. de sur-apprentissage) ? Comment peut-on y remédier ?

Exercice 2 (*Noyau polynomial*). — Soit $d, m \geqslant 1$ des entiers et $\mathrm{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ définie par :

$$\forall x, x' \in \mathbb{R}^d, \quad \mathrm{K}(x, x') = \langle x, x' \rangle^m .$$

Montrer que K est un noyau.

Exercice 3 (*Noyaux définis positifs*). — Pour chaque application $\mathrm{K}_1, \dots, \mathrm{K}_4$ définie ci-après, déterminer s'il s'agit d'un noyau, ou non.

1) Pour tous $x, y \in \mathbb{R}$, $\mathrm{K}_1(x, y) = 10^{xy}$.
2) Pour tous $x, y \in \mathbb{R}$, $\mathrm{K}_2(x, y) = 10^{x+y}$.
3) Pour tous $x, y \in [0, 1[$, $\mathrm{K}_3(x, y) = -\ln(1 - xy)$.
4) Soit $\mathscr{X}$ un ensemble quelconque, et $f, g\colon \mathscr{X} \to \mathbb{R}_+$ deux applications. Pour tous $x, y \in \mathscr{X}$, $\mathrm{K}_4(x, y) = \min\{f(x)g(y), f(y)g(x)\}$.

Exercice 4 (*Formulation ERM du Soft-SVM*). — Soit $n, d \geqslant 1$ deux entiers, $(x_i, y_i)_{i\in[n]} \in \mathscr{S}(\mathbb{R}^d, \{-1, 1\})$ un échantillon, et $\lambda > 0$. On considère les deux problèmes d'optimisation suivants, dont il s'agira d'établir l'équivalence.

$$\begin{aligned} &\text{minimiser} && \lambda \|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \xi_i \\ &\text{soumis aux contraintes} && w \in \mathbb{R}^d,\ b \in \mathbb{R} \\ & && \forall i \in [n], \quad \xi_i \geqslant 0 \\ & && \forall i \in [n], \quad y_i(\langle w, x_i\rangle + b) \geqslant 1 - \xi_i. \end{aligned} \tag{$*$}$$

$$\begin{aligned} &\text{minimiser} && \lambda \|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i\left(\langle w, x_i\rangle + b\right)) \\ &\text{soumis aux contraintes} && w \in \mathbb{R}^d,\ b \in \mathbb{R}. \end{aligned} \tag{$**$}$$

On note F et G les fonctions à minimiser dans les problèmes d'optimisation $(*)$ et $(**)$ respectivement :

$$\forall w \in \mathbb{R}^d,\ \forall \xi_1, \dots, \xi_n \in \mathbb{R}, \quad \mathrm{F}(w, \xi_1, \dots, \xi_n) = \lambda \|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \xi_i,$$

$$\forall w \in \mathbb{R}^d,\ b \in \mathbb{R}, \quad \mathrm{G}(w, b) = \lambda \|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i\left(\langle w, x_i\rangle + b\right)).$$

1) On suppose dans cette question que $(\hat{w}, \hat{b}, \hat{\xi}_1, \dots, \hat{\xi}_n)$ est une solution du problème $(*)$.
   a) Montrer que $\mathrm{G}(\hat{w}, \hat{b}) \leqslant \mathrm{F}(\hat{w}, \hat{\xi}_1, \dots, \hat{\xi}_n)$.
   b) Soit $(w, b) \in \mathbb{R}^d \times \mathbb{R}$. Exprimer $\mathrm{G}(w, b)$ à l'aide de F.
   c) Montrer que $\mathrm{F}(\hat{w}, \hat{\xi}_1, \dots, \hat{\xi}_n) \leqslant \mathrm{G}(w, b)$.
   d) Qu'en déduire ?
2) On suppose dans cette question que $(\hat{w}, \hat{b})$ est une solution du problème $(**)$. Montrer qu'il existe $\hat{\xi}_1, \dots, \hat{\xi}_n \in \mathbb{R}$ tels que $(\hat{w}, \hat{b}, \hat{\xi}_1, \dots, \hat{\xi}_n)$ est solution de $(*)$.