Télécom Paris | **An Introduction to Machine Learning Theory**
APM 5DS27 TP | **Stephan Clémençon - Ekhine Irurozki**

Travaux dirigés nº 3 : Séparateurs linéaires

Stephan Clémençon <stephan.clemencon@telecom-paris.fr>
Ekhine Irurozki <irurozki@telecom-paris.fr>

Exercice 1. On se place dans le cadre de la classification binaire : soient un descripteur aléatoire X à valeurs dans $\mathbb{R}$ muni de sa tribu des Boréliens, et un label aléatoire Y valant 0 ou 1.

Soit $\mathcal{G} := \{g : \mathbb{R} \to \{0,1\}\}$ l'ensemble des classifieurs adaptés à ce contexte. L'erreur de classification est définie comme l'application $\mathrm{L} : g \in \mathcal{G} \mapsto \mathbb{P}(\mathrm{Y} \neq g(\mathrm{X})) \in [0,1]$ et on note $\mathrm{L}^* := \inf_{g \in \mathcal{G}} \mathrm{L}(g)$.

Dans cet exercice, on s'intéresse à la famille $\mathcal{G}_0$ des classifieurs linéaires sur $\mathbb{R}$ de la forme :

$$g_{(x_0,y_0)} : x \in \mathbb{R} \mapsto \begin{cases} y_0 & \text{si } x \leq x_0, \\ 1 - y_0 & \text{sinon,} \end{cases}$$

avec $(x_0, y_0) \in \mathbb{R} \times \{0,1\}$. L'erreur de classification d'un tel $g_{(x_0,y_0)}$ est notée plus simplement $\mathrm{L}(x_0, y_0)$ et on pose $\mathrm{L}_0 := \inf_{(x_0,y_0)} \mathrm{L}(x_0, y_0)$.

1) Exprimer l'erreur de classification d'un élément quelconque de $\mathcal{G}_0$ en fonction des lois conditionnelles de X sachant Y. On utilisera les notations $\mathrm{F}_y(x) := \mathbb{P}\{\mathrm{X} \leq x \mid \mathrm{Y} = y\}$ pour $(x, y) \in \mathbb{R} \times \{0,1\}$ et $p := \mathbb{P}(\mathrm{Y} = 1)$.

2) En considérant les points $(x_0, y_0) = (-\infty, 0)$ et $(x_0, y_0) = (-\infty, 1)$, montrer que $\mathrm{L}_0 \leq \frac{1}{2}$.

3) Montrer que $\mathrm{L}_0 = \frac{1}{2} - \sup_x \left| p\,\mathrm{F}_1(x) - (1-p)\,\mathrm{F}_0(x) - p + \frac{1}{2} \right|$. Simplifier l'expression quand $p = \frac{1}{2}$.

   Indication. Pour tout $(a, b) \in \mathbb{R}^2$ on peut écrire $\min(a, b) = \frac{a + b - |a - b|}{2}$.

4) Montrer que $\mathrm{L}_0 = \frac{1}{2}$ si et seulement si $\mathrm{L}^* = \frac{1}{2}$.

5) Montrer l'inégalité de Chebychev-Cantelli : pour toute variable aléatoire réelle Z et tout $t \geq 0$,

$$\mathbb{P}(\mathrm{Z} - \mathbb{E}(\mathrm{Z}) \geq t) \leq \frac{\mathbb{V}(\mathrm{Z})}{\mathbb{V}(\mathrm{Z}) + t^2}.$$

6) On note respectivement m_y et σ_y^2 l'espérance et la variance de la loi conditionnelle de X sachant $\mathrm{Y} = y$, avec $y \in \{0,1\}$. Montrer que :

$$\mathrm{L}_0 \leq \left(1 + \frac{(m_0 - m_1)^2}{(\sigma_0 + \sigma_1)^2}\right)^{-1}.$$

   Indication. Utiliser l'inégalité démontrée à la question précédente.

7) Discuter de la performance du minimiseur empirique pris dans la classe $\mathcal{G}_0$ et des limites des classifieurs linéaires.