Université Paris Diderot | **Apprentissage statistique**
Master M2MO: 2017/2018 | **Joseph Salmon - Stéphan Clémençon**

Travaux dirigés n° 3 : Séparateurs linéaires

Stéphan Clémençon <stephan.clemencon@telecom-paristech.fr>
Joseph Salmon <joseph.salmon@telecom-paristech.fr>

Exercice 1. On se place dans le modèle de classification binaire $(X, Y) \sim P$ où P est une loi sur $\mathbb{R} \times \{0, 1\}$. On considère la famille $\mathcal{G}_L$ des classifieurs linéaires sur $\mathbb{R}$ de la forme :

$$g(x) = g_{(x_0,y_0)}(x) = \begin{cases} y_0 & \text{si } x \leq x_0 \\ 1 - y_0 & \text{sinon} \end{cases},$$

avec $x_0 \in \mathbb{R}$ et $y_0 \in \{0, 1\}$.

1) Exprimer l'erreur de classification $L(g) = \mathbb{P}\{Y \neq g(X)\}$ pour les éléments de $\mathcal{G}_L$ en fonction des lois conditionnelles de X sachant Y. On utilisera alors la notation $F_y(x) = \mathbb{P}\{X \leq x \mid Y = y\}$ pour $x \in \mathbb{R}$, $y \in \{0, 1\}$.

Pour un élément g_{x_0,y_0}, on note $L(x_0, y_0) = L(g_{x_0,y_0})$ et on pose $L_0 = \inf_{(x_0,y_0)} L(x_0, y_0)$.

2) En considérant les points $(x_0, y_0) = (-\infty, 0)$ et $(x_0, y_0) = (-\infty, 1)$, montrer que $L_0 \leq 1/2$.
3) On rappelle que $\min(a, b) = (a + b - |a - b|)/2$. Montrer que pour $\mathbb{P}\{Y = 1\} = p$ alors

$$L_0 = \frac{1}{2} - \sup_x |pF_1(x) - (1 - p)F_0(x) - p + \frac{1}{2}|)$$

Simplifier l'expression quand $p = 1/2$.

4) On note $L^* = \inf_g L(g)$. Montrer que $L_0 = 1/2$ si et seulement si $L^* = 1/2$. On montrera tout d'abord que $p = 1/2$ puis que $F_0 = F_1$.
5) Montrer l'inégalité de Chebychev-Cantelli : pour toute variable aléatoire Z et $t \geq 0$, on a

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

6) On note respectivement m_y et σ_y^2 l'espérance et la variance de la loi conditionnelle de X sachant $Y = y$. Montrer que :

$$L_0 \leq \left(1 + \frac{(m_0 - m_1)^2}{(\sigma_0 + \sigma_1)^2}\right)^{-1} .$$

Aide : utiliser l'inégalité démontrée à la question précédente.

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