Apprentissage: cours 10 (1ère partie)
Théorèmes No Free Lunch

Sylvain Arlot

3 mai 2012

1 No free lunch theorem (cours 7b, 1h)

Référence : Chapitre 7 de [DGL96].

Si la consistance universelle uniforme est possible en classification 0–1 lorsque $\mathcal{X}$ est fini (voir Corollaire 2 du cours "minimisation du risque empirique"), il est sans espoir d'obtenir une règle de classification uniformément universellement consistant en classification lorsque $\mathcal{X}$ est infini, comme le montre le résultat suivant.

Théorème 1. *On considère la perte* $0-1$ $\ell(f;(x,y)) = \mathbb{1}_{f(x)\neq y}$ *en classification binaire supervisée, et l'on suppose que* $\mathcal{X}$ *est infini. Alors, pour tout* $n \in \mathbb{N}$ *et toute règle de classification* $\widehat{f} : (\mathcal{X}\times\mathcal{Y})^n \mapsto \mathcal{F}$*,*

$$\sup_P \left\{ \mathbb{E}_{D_n\sim P^{\otimes n}} \left[\mathcal{R}\left(\widehat{f}\left(D_n\right)\right) - \mathcal{R}\left(f^\star\right) \right] \right\} \geqslant \frac{1}{2} > 0 \ , \tag{1}$$

le sup *étant pris sur l'ensembles des mesures de probabilité sur* $\mathcal{X}\times\mathcal{Y}$*. En particulier, aucune règle de classification ne peut être uniformément universellement consistante lorsque* $\mathcal{X}$ *est infini.*

Démonstration. Soit $n, K \in \mathbb{N}$, $\widehat{f} : (\mathcal{X}\times\mathcal{Y})^n \mapsto \mathcal{F}$ une règle de classification. L'espace $\mathcal{X}$ étant infini, à bijection près, on peut supposer que $\mathbb{N} \subset \mathcal{X}$.

Pour tout $r \in \{0,1\}^K$, notons P_r la distribution de probabilité sur $\mathcal{X}\times\mathcal{Y}$ définie par $\mathbb{P}_{(X,Y)\sim P_r}(X = j \text{ et } Y = r_j) = K^{-1}$ pour tout $j \in \{1,\dots,K\}$. Autrement dit, X est choisi uniformément parmi $\{1,\dots,K\}$, et $Y = r_X$ est une fonction déterministe de X. En particulier, pour tout r, $\mathcal{R}_{P_r}\left(f^\star\right) = 0$.

Pour tout $r \in \{0,1\}^K$ (déterministe), on pose

$$F(r) = \mathbb{E}_{(X_i,Y_i)_{1\leqslant i\leqslant n}\sim P_r^{\otimes n}} \left[\mathcal{R}_{P_r}\left(\widehat{f}\left(D_n\right)\right) \right] \ .$$

La remarque clé est que pour toute distribution de probabilité R sur $\{0,1\}^K$,

$$\sup_{r\in\{0,1\}^K} \{F(r)\} \geqslant \mathbb{E}_{r\sim R}\left[F(r)\right] \ .$$

Notons R la distribution uniforme sur $\{0,1\}^K$, de telle sorte que $r \sim R$ signifie que $r_1, \dots, r_K$ sont indépendantes et de même distribution Bernoulli $\mathcal{B}(1/2)$. Alors,

$$\begin{aligned}
\mathbb{E}_{r\sim R}\left[F(r)\right] &= \mathbb{P}\left(\widehat{f}\left(X; D_n\right) \neq Y\right) \\
&= \mathbb{P}\left(\widehat{f}\left(X; D_n\right) \neq r_X\right) \\
&= \mathbb{E}\left[\mathbb{P}_{(r_j)_{j\notin\{X_1,\dots,X_n\}}}\left(\widehat{f}\left(X; D_n\right) \neq r_X \;\middle|\; X, X_1, \dots, X_n, r_{X_1}, \dots, r_{X_n}\right)\right] \\
&\geqslant \mathbb{E}\left[\mathbb{E}_{(r_j)_{j\notin\{X_1,\dots,X_n\}}}\left(\mathbb{1}_{\widehat{f}(X;D_n)\neq r_X}\mathbb{1}_{X\notin\{X_1,\dots,X_n\}} \;\middle|\; X, X_1, \dots, X_n, r_{X_1}, \dots, r_{X_n}\right)\right] \\
&= \mathbb{E}_{X,X_1,\dots,X_n,r_{X_1},\dots,r_{X_n}}\left[\frac{\mathbb{1}_{X\notin\{X_1,\dots,X_n\}}}{2}\right] \\
&= \frac{1}{2}\left(1 - \frac{1}{K}\right)^n \quad .
\end{aligned}$$

Pour tout $n \in \mathbb{N}$ fixé, cette borne inférieure tend vers $1/2$ lorsque K tend vers 1, d'où le résultat. □

Un défaut du Théorème 1 est que la distribution P faisant échouer une règle de classification arbitraire $\widehat{f}$ change pour chaque taille d'échantillon. On pourrait donc imaginer qu'il est tout de même possible d'avoir une majoration de l'excès de risque de $\widehat{f}$ de la forme $c(P)u_n$ pour une suite $(u_n)_{n\geqslant 1}$ tendant vers 0 et une constante $c(P)$ fonction de la loi des observations. Le résultat suivant montre que ce n'est pas le cas, même avec une suite $(u_n)_{n\geqslant 1}$ tendant très lentement vers zéro.

Théorème 2 (Théorème 7.2 [DGL96], admis)**.** *On considère la perte* $0-1$ $\ell(f;(x,y)) = \mathbb{1}_{f(x)\neq y}$ *en classification binaire supervisée* ($\mathcal{Y} = \{0,1\}$)*, et l'on suppose que* $\mathcal{X}$ *est infini. Soit* $(a_n)_{n\geqslant 1}$ *une suite de réels positifs, décroissante, convergeant vers zéro, et telle que* $a_1 \leqslant 1/16$*. Alors, pour toute règle de classification* $\widehat{f} : \bigcup_{n\geqslant 1}(\mathcal{X}\times\mathcal{Y})^n \mapsto \mathcal{F}$*, il existe une distribution* P *sur* $\mathcal{X}\times\mathcal{Y}$ *telle que pour tout* $n \geqslant 1$*,*

$$\mathbb{E}_{D_n\sim P^{\otimes n}}\left[\mathcal{R}\left(\widehat{f}\left(D_n\right)\right) - \mathcal{R}\left(f^\star\right)\right] \geqslant a_n \quad . \tag{2}$$

Références

[DGL96] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.