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

Joon Kwon

mercredi 23 mars 2022

Soit $d \geqslant 1$ un entier. Pour $(w, b) \in \mathbb{R}^d \times \mathbb{R}$, on note $g_{w,b}$ la fonction définie par :

$$\forall x \in \mathbb{R}^d, \quad g_{w,b}(x) = \langle w, x \rangle + b.$$

On se place dans un cadre de classification binaire. Soient $\mathscr{X} = \mathbb{R}^d$ et $\mathscr{Y} = \{-1, 1\}$ les ensembles d'entrée et de sortie, et $\mathrm{S} = (x_i, y_i)_{i \in [n]}$ l'échantillon d'apprentissage. Soit $(\hat{w}, \hat{b}) \in \mathbb{R}^d \times \mathbb{R}$ tel que $\hat{f} = \operatorname{sign} \circ g_{\hat{w},\hat{b}}$ soit le classifieur donné par la régression logistique.

1) Rappeler la définition de $\hat{w}$ et $\hat{b}$.

Le but de l'exercice est de montrer que la régression logistique peut être vue comme une minimisation de risque empirique dans un problème auxiliaire de *régression*.

Soit $\phi_{\mathrm{sig}} : \mathbb{R} \to]0, 1[$ la fonction définie par

$$\forall z \in \mathbb{R}, \quad \phi_{\mathrm{sig}}(z) = \frac{1}{1 + e^{-z}}.$$

2) Tracer l'allure de ϕ_{sig}.

3) Montrer que ϕ_{sig} admet une fonction réciproque et en donner une expression.

On considère un problème auxiliaire avec pour ensembles d'entrée et de sortie $\mathscr{X}_0 = \mathbb{R}^d$ et $\mathscr{Y}_0 = [0, 1]$, et pour classe de prédicteurs :

$$\mathscr{F}_0 = \left\{\phi_{\text{sig}} \circ g_{w,b}\right\}_{\substack{w\in\mathbb{R}^d \\ b\in\mathbb{R}}}.$$

Soit $S_0 = \left(x_i, \frac{y_i+1}{2}\right)_{i\in[n]}$.

4) Soit $\hat{f}_0 = \phi_{\text{sig}} \circ g_{\hat{w},\hat{b}}$. Montrer que $\hat{f}_0$ est le prédicteur donné par la minimisation du risque empirique dans ce problème auxiliaire avec l'échantillon d'apprentissage S_0 et pour une certaine fonction de perte ℓ_0 à déterminer.

5) Montrer que $\hat{f} = \text{sign} \circ (\hat{f}_0 - \frac{1}{2})$.