
TRAVAUX DIRIGÉS N° 5 : Perceptron

Stephan CLÉMENÇON <stephan.clemencon@telecom-paris.fr>

Quentin BOUNIOT <quentin.bouniot@telecom-paris.fr>

EXERCICE 1. On se place dans le cadre de la classification binaire, comme au TD 4. On observe cette fois-ci un n -échantillon $\mathcal{D}_n = \{(X_i, Y_i)\}_{1 \leq i \leq n}$, représentant n copies indépendantes du (X, Y) . Pour des paramètres $\beta \in \mathbb{R}^d$ et $\beta_0 \in \mathbb{R}$ on note $f_{\beta, \beta_0} : x \in \mathbb{R}^d \mapsto \beta^\top x + \beta_0$ puis H_{β, β_0} l'hyperplan (affine) défini par $H_{\beta, \beta_0} = \{x \in \mathbb{R}^d : f_{\beta, \beta_0}(x) = 0\}$. On considère les classifieurs de la forme $\text{sgn} \circ f_{\beta, \beta_0}$ et on suppose que les données sont linéairement séparables.

- 1) Soit $x \in \mathbb{R}^d$. Calculer $d(x, H_{\beta, \beta_0})$ où d est la distance euclidienne.
- 2) On propose comme mesure d'erreur associée à une fonction f_{β, β_0} (Hinge loss) :

$$\ell_H(f_{\beta, \beta_0}) = - \sum_{\{i: Y_i f_{\beta, \beta_0}(X_i) < 0\}} Y_i f_{\beta, \beta_0}(X_i).$$

Comparer avec l'écriture de la perte $\ell_{0/1}$ classique utilisée habituellement.

- 3) Vérifier que l'on peut toujours se ramener à la situation où $\beta_0 = 0$. On supposera dorénavant que c'est le cas.
- 4) L'algorithme du perceptron consiste à optimiser en β la perte ℓ_H avec une approche de type descente de (sous-)gradient stochastique. On pourra introduire un paramètre de taux d'apprentissage ρ (learning rate), et l'on notera β^k , le coefficient obtenu lorsque l'on a commis la k^e erreur.
Discuter d'initialisations possibles pour β^0 , et proposer les étapes successives pour mettre à jour β au vue des données.

On va prouver le résultat suivant : "Si les deux classes sont strictement séparables par un hyperplan sur l'échantillon d'apprentissage alors le perceptron s'arrête en un nombre fini d'étapes". On supposera dans la suite que les X_i sont presque-sûrement de norme (euclidienne) bornée par $B > 0$.

- 5) Pour tout $\beta \in \mathbb{R}^d$ on définit $\gamma(\beta) := \min_{1 \leq i \leq n} Y_i \beta^\top X_i$. Interpréter ce paramètre et en déduire une formulation explicite de l'hypothèse de séparabilité linéaire. On pose $\gamma^* := \max_{\|\beta\|=1} \gamma(\beta)$. Vérifier son existence. Quel est son signe ?
- 6) Par simplicité on initialise l'algorithme avec $\beta^0 = 0$. Montrer alors que le perceptron s'arrête en au plus $\left(\frac{B}{\gamma^*}\right)^2$ mises à jour.

Indication. Borner $\|\beta^k\|$ pour $k \geq 1$ avant la dernière mise à jour.

Enfin, pour plus de détails sur le perceptron, une bonne référence est le livre de Ripley : [Ripley, 1996].

Références

[Ripley, 1996] Ripley, B. (1996). *Pattern recognition via neural networks*. Cambridge University Press. 1