
TRAVAUX DIRIGÉS N° 5 : Perceptron

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

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

EXERCICE 1. We work in the framework of binary classification, as in Tutorial 4. This time we observe an n -sample $\mathcal{D}_n = \{(X_i, Y_i)\}_{1 \leq i \leq n}$, made of n independent copies of (X, Y) . For parameters $\beta \in \mathbb{R}^d$ and $\beta_0 \in \mathbb{R}$ we write $f_{\beta, \beta_0} : x \in \mathbb{R}^d \mapsto \beta^\top x + \beta_0$ and H_{β, β_0} for the (affine) hyperplane defined by $H_{\beta, \beta_0} = \{x \in \mathbb{R}^d : f_{\beta, \beta_0}(x) = 0\}$. We consider classifiers of the form $\text{sgn} \circ f_{\beta, \beta_0}$ and assume that the data are linearly separable.

- 1) Let $x \in \mathbb{R}^d$. Compute $d(x, H_{\beta, \beta_0})$, where d is the Euclidean distance.
- 2) We propose the following error measure associated with a function f_{β, β_0} (Hinge loss) :

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

Compare with the expression of the usual $\ell_{0/1}$ loss.

- 3) Check that one can always reduce to the case $\beta_0 = 0$. From now on we assume this is the case.
- 4) The perceptron algorithm consists in optimizing the loss ℓ_H with respect to β using a stochastic (sub)gradient descent approach. One may introduce a learning rate parameter $\eta > 0$, and we denote by β^k the coefficient obtained after the k -th mistake.

Discuss possible initializations for β^0 , and propose the successive steps to update β given the data.

We will prove the following result : “If the two classes are strictly separable by a hyperplane on the training sample, then the perceptron stops after a finite number of steps”. In what follows we assume that the X_i have (Euclidean) norm almost surely bounded by $B > 0$.

- 5) For every $\beta \in \mathbb{R}^d$ define $\gamma(\beta) := \min_{1 \leq i \leq n} Y_i \beta^\top X_i$. Interpret this quantity and deduce an explicit formulation of the linear separability assumption. Let $\gamma^* := \max_{\|\beta\|=1} \gamma(\beta)$. Check that it exists. What is its sign ?
- 6) For simplicity, initialize the algorithm with $\beta^0 = 0$. Show that the perceptron then stops after at most $\left(\frac{B}{\gamma^*}\right)^2$ updates.

Hint. Bound $\|\beta^k\|$ for $k \geq 1$ before the last update.

Finally, for more details on the perceptron, a good reference is the book by Ripley : [Ripley, 1996].

Références

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