
TRAVAUX DIRIG  S N   5 : Perceptron

Stephan CL  MEN  ON <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.

Correction Exercise 1

Throughout the solution, $\beta \neq 0$ and we write $P = \{i : Y_i = +1\}$, $N = \{i : Y_i = -1\}$.

1) The shortest path to $H := H_{\beta, \beta_0}$ follows the normal direction β : we look for a point of H of the form $x - \lambda\beta$. The condition $\beta^\top (x - \lambda\beta) + \beta_0 = 0$ has the unique solution $\lambda^* = f_{\beta, \beta_0}(x) / \|\beta\|^2$, and we set $p(x) := x - \lambda^*\beta \in H$ (orthogonal projection onto H).

The infimum is therefore attained at the orthogonal projection $p(x)$, and

$$d(x, H_{\beta, \beta_0}) = |\lambda^*| \|\beta\| = \frac{|\beta^\top x + \beta_0|}{\|\beta\|}.$$

In particular, if $\|\beta\| = 1$, $|f_{\beta,\beta_0}(x)|$ is exactly the distance from x to the hyperplane.

2) Writing $z_i := Y_i f_{\beta,\beta_0}(X_i)$ (functional margin) and $\phi(z) := \max(0, -z) = -z \mathbb{1}_{\{z < 0\}}$, the two losses read

$$\ell_{0/1}(f_{\beta,\beta_0}) = \sum_{i=1}^n \mathbb{1}_{\{Y_i \neq \text{sgn}(f_{\beta,\beta_0}(X_i))\}} = \sum_{i=1}^n \mathbb{1}_{\{z_i < 0\}} \quad \ell_H(f_{\beta,\beta_0}) = \sum_{i=1}^n \phi(z_i).$$

(Normalizing by $1/n$ changes neither the minimizers nor the algorithm : the constant is absorbed into the learning rate.)

- The loss $\ell_{0/1}$ counts each mistake as 1. It depends on β only through the signs of the z_i : it is piecewise constant, with zero gradient almost everywhere, non-convex, and minimizing it is NP-hard in general. We would recover it by replacing f with $\text{sgn} \circ f$ in ℓ_H .
- The loss ℓ_H penalizes each mistake by $|z_i| = \|\beta\| d(X_i, H_{\beta,\beta_0})$ (question 1), *i.e.* proportionally to its distance to the boundary. It is continuous and convex in (β, β_0) : ϕ is a maximum of two affine functions, hence convex, and $(\beta, \beta_0) \mapsto z_i$ is linear.

Since ϕ is not differentiable at 0, we will use its subdifferential : $\partial\phi(z) = \{-1\}$ if $z < 0$, $[-1, 0]$ if $z = 0$, $\{0\}$ if $z > 0$.

3) We add a constant coordinate to the features :

$$\tilde{X}_i = \begin{pmatrix} 1 \\ X_i \end{pmatrix} \in \mathbb{R}^{d+1}, \quad \tilde{\beta} = \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix} \in \mathbb{R}^{d+1}, \quad \tilde{\beta}^\top \tilde{X}_i = \beta^\top X_i + \beta_0 = f_{\beta,\beta_0}(X_i).$$

The classifier and the loss are unchanged : looking for an affine hyperplane in $\mathbb{R}^d$ amounts to looking for a linear hyperplane in $\mathbb{R}^{d+1}$. The bound on the norms becomes $\|\tilde{X}_i\| \leq \sqrt{1 + B^2}$, which does not change the argument of question 6. From now on we write $H_\beta := H_{\beta,0}$.

Geometric proof (convex hulls). Let C_P, C_N be the convex hulls of the two classes. A hyperplane separates the classes if and only if it separates C_P and C_N . Take $x^+ \in C_P, x^- \in C_N$ at minimal distance (they exist by compactness). The open ball with diameter $[x^-, x^+]$ meets neither C_P nor C_N (otherwise we would find closer points). For $z \in [x^-, x^+]$, the hyperplane perpendicular to $[x^-, x^+]$ through z separates the data : a positive point on the side of x^- would give a segment $[X_i, x^+] \subset C_P$ crossing the ball. Translating the data by z yields a separation by a linear hyperplane. This is an existence argument, but finding the right z would be too costly.

4) We minimize $\ell_H(\beta) = \sum_i \ell_i(\beta)$ with $\ell_i(\beta) = \phi(Y_i \beta^\top X_i)$. Stochastic subgradient descent uses a single term at a time : $\beta \leftarrow \beta - \eta g_i$ with $g_i \in \partial \ell_i(\beta)$. By composition with the linear map $\beta \mapsto Y_i \beta^\top X_i$, we have $g_i = s Y_i X_i$ with $s \in \partial \phi(Y_i \beta^\top X_i)$:

- correctly classified point ($Y_i \beta^\top X_i > 0$) : $s = 0$, β is unchanged ;
- misclassified point or point on the boundary ($Y_i \beta^\top X_i \leq 0$) : we choose $s = -1$, hence $\beta \leftarrow \beta + \eta Y_i X_i$.

The choice $s = -1$ at $z = 0$ (test ≤ 0 rather than < 0) is essential. With $s = 0$, $\beta = 0$ would be a fixed point, although it classifies nothing. With $s = -1$, the algorithm stops only if $\gamma(\beta) > 0$, *i.e.* on a strict separation, which question 6 relies on.

Algorithm 1 : Perceptron

Input : data $(X_i, Y_i)_{1 \leq i \leq n}$, learning rate $\eta > 0$, initialization β^0
 $\beta \leftarrow \beta^0, k \leftarrow 0$;
repeat
 for each $i \in \{1, \dots, n\}$ (in any or random order) **do**
 if $Y_i \beta^\top X_i \leq 0$ **then**
 $\beta \leftarrow \beta + \eta Y_i X_i$;
 $k \leftarrow k + 1$; // $\beta^k = \beta$: k -th mistake
until a full pass over the data triggers no update;
Output : β

Initialization. The simplest choice is $\beta^0 = 0$. One can also initialize randomly, but with a constant step size, a nonzero initialization may cost extra updates (see question 6). Moreover, one has to choose the distribution used for the initialization and its parameters. One can also vary η , for instance with the number of mistakes.

5) $\gamma(\beta) = \min_i Y_i \beta^\top X_i = (\min_{i \in P} \beta^\top X_i) \wedge (-\max_{i \in N} \beta^\top X_i)$ is the functional margin of the worst point. Since $\gamma(c\beta) = c\gamma(\beta)$ for $c > 0$ while the hyperplane is unchanged, we restrict to $\|\beta\| = 1$. In that case, $|\beta^\top X_i| = d(X_i, H_\beta)$ (question 1), and :

- if $\gamma(\beta) > 0$, all points are strictly correctly classified and $\gamma(\beta)$ is the distance from the closest point to the hyperplane (*geometric margin*);
- if $\gamma(\beta) = 0$, no point is misclassified but at least one lies on the hyperplane;
- if $\gamma(\beta) < 0$, at least one point is misclassified, and $-\gamma(\beta)$ is the distance to H_β of the worst-classified point.

The linear separability assumption therefore reads : *there exists $\beta \in \mathbb{R}^d$, $\|\beta\| = 1$, such that $\gamma(\beta) > 0$.*

Existence of γ^* . The unit sphere $\{\|\beta\| = 1\}$ is compact (closed and bounded in finite dimension) and γ is continuous (finite minimum of linear functions) : γ attains its maximum on it at some unit vector β^* .

Sign. If the data are separable, $\gamma^* \geq \gamma(\beta) > 0$. Otherwise, $\gamma(\beta) \leq 0$ for all β , so $\gamma^* \leq 0$. Hence the data are linearly separable if and only if $\gamma^* > 0$, and γ^* is the largest margin achievable by a separating hyperplane. By Cauchy-Schwarz, $\gamma^* \leq B$.

6) Let β^* be a unit vector such that $\gamma(\beta^*) = \gamma^* > 0$, so that $Y_i \beta^{*\top} X_i \geq \gamma^*$ for all i . Let i_k be the point used at the k -th update : $Y_{i_k} \beta^{k-1\top} X_{i_k} \leq 0$ and $\beta^k = \beta^{k-1} + \eta Y_{i_k} X_{i_k}$. The idea is to bound $\|\beta^k\|$ from both sides.

Lower bound. $\beta^{k\top} \beta^* = \beta^{k-1\top} \beta^* + \eta Y_{i_k} \beta^{k-1\top} X_{i_k} \geq \beta^{k-1\top} \beta^* + \eta \gamma^*$. Iterating from $\beta^0 = 0$, then by Cauchy-Schwarz ($\|\beta^*\| = 1$) :

$$\|\beta^k\| \geq \beta^{k\top} \beta^* \geq k\eta\gamma^*.$$

Upper bound. Since updates happen only on mistakes :

$$\|\beta^k\|^2 = \|\beta^{k-1}\|^2 + 2\eta \underbrace{Y_{i_k} \beta^{k-1\top} X_{i_k}}_{\leq 0} + \eta^2 \underbrace{\|X_{i_k}\|^2}_{\leq B^2} \leq \|\beta^{k-1}\|^2 + \eta^2 B^2, \quad \text{hence} \quad \|\beta^k\|^2 \leq k\eta^2 B^2.$$

Conclusion. $k^2 \eta^2 (\gamma^*)^2 \leq \|\beta^k\|^2 \leq k\eta^2 B^2$, so any number k of updates performed satisfies $k \leq (B/\gamma^*)^2$. The perceptron stops after a finite number of steps, on a β that strictly separates the data. ■

The rate η cancels out : with $\beta^0 = 0$, all the β^k are proportional to η . The difficulty is measured by the radius/margin ratio B/γ^* , independently of n and d (and $(B/\gamma^*)^2 \geq 1$ since $\gamma^* \leq B$). This is a worst-case guarantee. If the data are not separable, the algorithm never stops.

General case : arbitrary β^0 and varying rate η_k . Assume $\gamma_0 := \gamma(\beta^0) \leq 0$ (otherwise no update happens) and reason on $\beta^k - \beta^0$. The same lower bound gives $\|\beta^k - \beta^0\| \geq \gamma^* \sum_{\ell \leq k} \eta_\ell$. For the upper bound,

$$\|\beta^k - \beta^0\|^2 = \|\beta^{k-1} - \beta^0\|^2 + \eta_k^2 \|X_{i_k}\|^2 + 2\eta_k \underbrace{Y_{i_k} \beta^{k-1 \top} X_{i_k}}_{\leq 0} - 2\eta_k \underbrace{Y_{i_k} \beta^0 \top X_{i_k}}_{\geq \gamma_0},$$

hence $\|\beta^k - \beta^0\|^2 \leq B^2 \sum_{\ell \leq k} \eta_\ell^2 + 2|\gamma_0| \sum_{\ell \leq k} \eta_\ell$. Combining both bounds :

$$1 \leq \left(\frac{B}{\gamma^*}\right)^2 \frac{\sum_{\ell \leq k} \eta_\ell^2}{(\sum_{\ell \leq k} \eta_\ell)^2} + \frac{2|\gamma_0|}{(\gamma^*)^2} \frac{1}{\sum_{\ell \leq k} \eta_\ell}.$$

- $\eta_k = \eta, \beta^0 = 0 : k \leq (B/\gamma^*)^2$, we recover the previous result.
- $\eta_k = \eta, \beta^0 \neq 0 : k \leq (B/\gamma^*)^2 + 2|\gamma_0|/(\eta(\gamma^*)^2)$. A bad initialization costs extra updates, all the more so as η is small.
- $\sum_\ell \eta_\ell \rightarrow \infty$ and $\sum_\ell \eta_\ell^2 / (\sum_\ell \eta_\ell)^2 \rightarrow 0$: the right-hand side goes to 0, so the inequality is violated beyond some index k_0 , and the algorithm performs at most k_0 updates.

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. 4