

CHAPITRE III  
**PRÉDICTEURS LINÉAIRES**  
CYCLE PLURIDISCIPLINAIRE D'ÉTUDES SUPÉRIEURES  
UNIVERSITÉ PARIS SCIENCES ET LETTRES



I. PRÉLIMINAIRES

**NOTATION.** — Pour  $d \geq 1$  un entier et  $x, x' \in \mathbb{R}^d$ , on note  $\langle x, x' \rangle = \sum_{j=1}^d x_j x'_j$  le produit scalaire canonique.

**NOTATION.** — Pour  $w \in \mathbb{R}^d$  et  $b \in \mathbb{R}$ , on note :

$$\begin{aligned} g_{w,b} : \mathbb{R}^d &\longrightarrow \mathbb{R} \\ x &\longmapsto \langle w, x \rangle + b. \end{aligned}$$

Il s'agit des applications affines de  $\mathbb{R}^d$  dans  $\mathbb{R}$ . On note  $\mathcal{L}_d = \{g_{w,b}\}_{\substack{w \in \mathbb{R}^d \\ b \in \mathbb{R}}}$  l'ensemble de ces applications.

**DÉFINITION.** — *Prédicteurs linéaires.* — Soit  $f \in \mathcal{F}(\mathbb{R}^d, \mathbb{R})$ .  $f$  est un *prédicteur linéaire* s'il est de la forme  $f = \phi \circ g$  avec  $g \in \mathcal{L}_d$  et  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  quelconque.

**REMARQUE.** — La fonction  $\phi$  étant quelconque, un *prédicteur* linéaire n'est pas nécessairement une *application* linéaire.

**EXEMPLE.** — *Classifieurs linéaires.* — On note  $\text{sign}$  la fonction donnant le signe strict d'un nombre réel, c'est-à-dire :

$$\forall x \in \mathbb{R}, \quad \text{sign}(x) = \begin{cases} 1 & \text{si } x > 0 \\ 0 & \text{si } x = 0 \\ -1 & \text{si } x < 0. \end{cases}$$

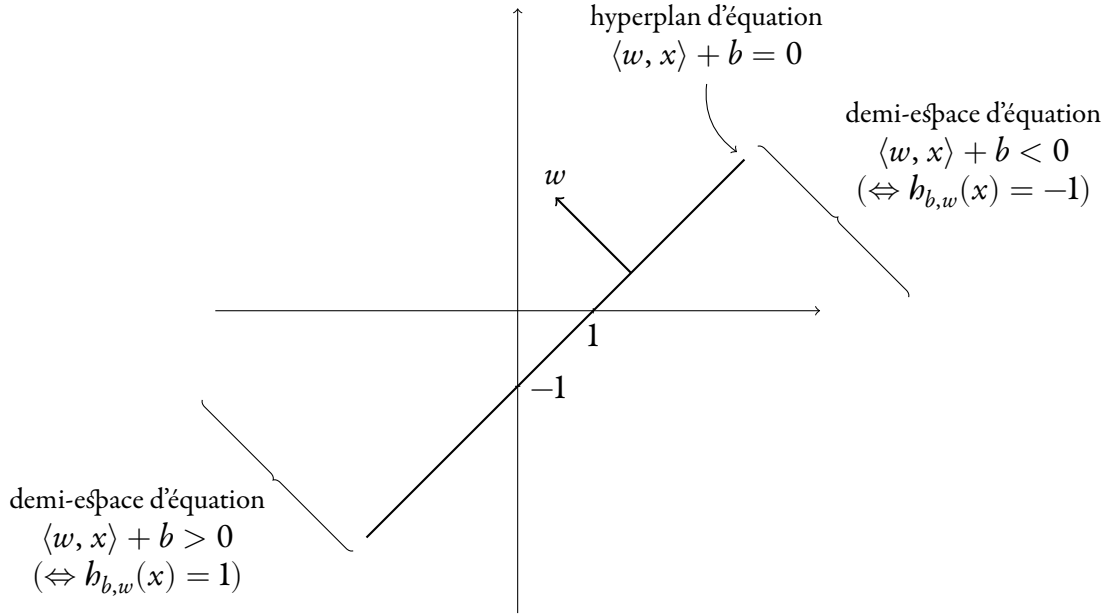
Pour  $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ , on note  $h_{w,b}$  le *classifieur linéaire* associé, qui est défini par :

$$h_{w,b} = \text{sign} \circ g_{w,b}.$$

On note  $\mathcal{H}_d = \{h_{w,b}\}_{\substack{w \in \mathbb{R}^d \\ b \in \mathbb{R}}}$ . Un classifieur linéaire partitionne l'espace en deux demi-espaces séparés par un hyperplan :

- l'hyperplan est constitué de l'ensemble des points  $x \in \mathbb{R}^d$  tels que  $\langle w, x \rangle + b = 0$  (ce qui équivaut à  $h_{w,b}(x) = 0$ ) ; on dit également que l'hyperplan est d'équation  $\langle w, x \rangle + b = 0$  ;
- l'un des demi-espaces est d'équation  $\langle w, x \rangle + b > 0$  et correspond aux points  $x \in \mathbb{R}^d$  pour lesquels  $h_{w,b}(x)$  prédit l'étiquette +1 ;
- l'autre demi-espace est d'équation  $\langle w, x \rangle + b < 0$  et correspond aux points  $x \in \mathbb{R}^d$  pour lesquels  $h_{w,b}$  prédit l'étiquette -1.

La figure ci-après représente l'hyperplan et les deux demi-espaces dans le cas simple où  $\mathcal{X} = \mathbb{R}^2$ ,  $w = (-1, 1)$  et  $b = 1$ . On peut voir que  $w$  s'interprète comme un vecteur normal à l'hyperplan d'équation  $\langle w, x \rangle + b = 0$  et indique (par son sens) l'hyperplan d'équation  $\langle w, x \rangle + b > 0$ .



## 2. RÉGRESSION LINÉAIRE

On considère dans ce paragraphe  $\mathcal{X} = \mathbb{R}^d$ ,  $\mathcal{Y} = \mathbb{R}$ , et pour classe de prédicteurs  $\mathcal{F} = \mathcal{L}_d$  (l'ensemble des applications affines). La fonction de perte  $\ell: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  est à choisir ; le plus souvent, il s'agira de la perte quadratique  $\ell(y, y') = (y - y')^2$ , et on rencontre également la perte norme absolue  $\ell(y, y') = |y - y'|$ .

**DÉFINITION.** — *Régression linéaire aux moindres carrés.* — Soit  $n \geq 1$  et  $(x_i, y_i)_{i \in [n]} \in \mathcal{S}(\mathcal{X}, \mathcal{Y})$  un échantillon d'apprentissage. L'algorithme de régression linéaire aux moindres carrés donne le prédicteur :

$$\hat{f} = \arg \min_{f \in \mathcal{L}_d} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \right\}.$$

Autrement dit,  $\hat{f}$  est le minimiseur du risque empirique<sup>1</sup> associé à la classe  $\mathcal{L}_d$  et à la perte quadratique.

1. voir la définition dans le Chapitre II.

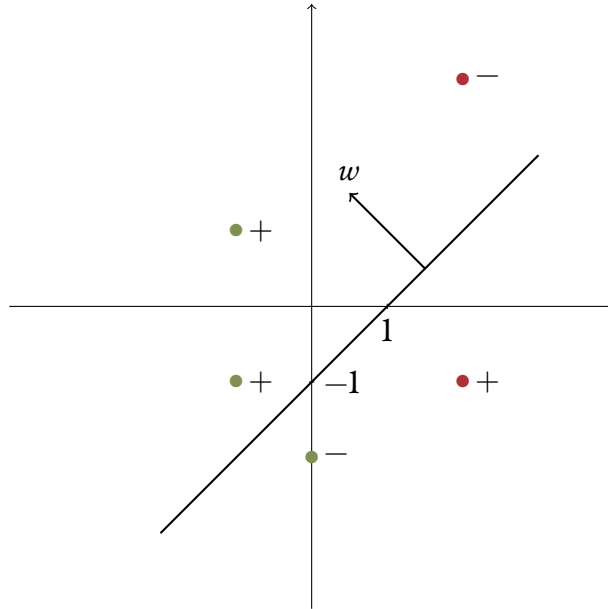
### 3. CLASSIFICATION BINAIRE : PERCEPTRON

On considère dans ce paragraphe  $\mathcal{X} = \mathbb{R}^d$  et  $\mathcal{Y} = \{-1, +1\}$ . Il s'agit donc d'un cadre de classification binaire. La classe de prédicteur considérée est celle des classifieurs linéaires :  $\mathcal{F} = \mathcal{H}_d$ .

**REMARQUE.** — Soit  $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ . Un exemple  $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$  est *correctement prédit* par  $h_{b,w}$  si et seulement si :

$$\begin{aligned} h_{b,w}(x) = y &\iff (\langle w, x \rangle + b) \text{ et } y \text{ sont du même signe strict} \\ &\iff y(\langle w, x \rangle + b) > 0. \end{aligned}$$

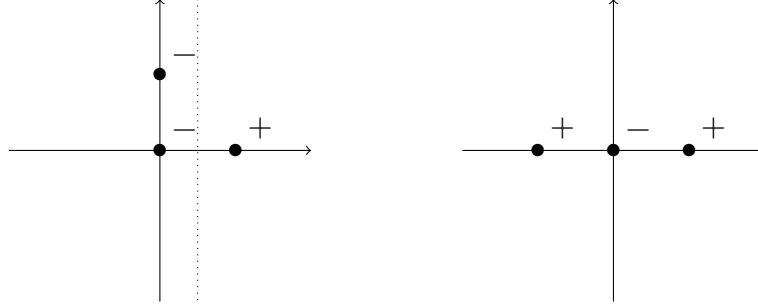
**EXEMPLE.** — On reprend l'exemple où  $\mathcal{X} = \mathbb{R}^2$ ,  $w = (-1, 1)$  et  $b = 1$ . On représente ci-après un échantillon de 4 exemples où les étiquettes (égales à  $+1$  ou  $-1$ ) sont signalées par des signes  $+$  et  $-$ . Les trois points en vert correspondent aux exemples bien prédits par  $h_{b,w}$  et les points rouges à ceux qui sont mal prédits.



**DÉFINITION.** — *Séparabilité linéaire.* — Soit  $n \geq 1$ . Un échantillon  $(x_i, y_i)_{i \in [n]} \in \mathcal{S}(\mathbb{R}^d, \{-1, 1\})$  est *linéairement séparable* s'il existe un classifieur linéaire  $h \in \mathcal{H}_d$  tel que :

$$\forall i \in [n], \quad h(x_i) = y_i.$$

**EXEMPLE.** — L'échantillon de gauche est linéairement séparable tandis que celui de droite ne l'est pas.



**DÉFINITION.** — *Perceptron.* — Soit  $(x_i, y_i)_{i \in [n]} \in \mathcal{S}(\mathbb{R}^d, \{-1, 1\})$  un échantillon d'apprentissage. L'algorithme *Perceptron* est itératif.

- *Initialisation* :  $w^{(1)} = 0 \in \mathbb{R}^d$  et  $b^{(1)} = 0$ .
- À chaque étape  $t \geq 1$ , on considère le classifieur (temporaire)  $\hat{h}^{(t)} = h_{w^{(t)}, b^{(t)}}$ .
- S'il existe  $i \in [n]$  tel que

$$\hat{h}^{(t)}(x_i) \neq y_i \quad (\Leftrightarrow \quad y_i (\langle w^{(t)}, x_i \rangle + b^{(t)}) \leq 0),$$

alors on pose :

$$w^{(t+1)} = w^{(t)} + y_i x_i \quad \text{et} \quad b^{(t+1)} = b^{(t)} + y_i.$$

- Sinon, s'algorithme s'arrête et renvoie le classifieur  $\hat{h} = \hat{h}^{(t)}$ .

**THÉORÈME.** — Si l'échantillon d'apprentissage est linéairement séparable, l'algorithme *Perceptron* s'arrête. Alors, le classifieur renvoyé prédit correctement chaque exemple de l'échantillon d'apprentissage.

**REMARQUE.** — Si l'échantillon d'apprentissage n'est pas linéairement séparable, l'algorithme ne s'arrête pas. En pratique, on peut forcer l'arrêt à l'aide d'une *condition d'arrêt*, comme par exemple un nombre maximal d'étapes.

#### 4. CLASSIFICATION BINAIRE : RÉGRESSION LOGISTIQUE

On se place dans le cadre du paragraphe précédent, c'est-à-dire  $\mathcal{X} = \mathbb{R}^d$ ,  $\mathcal{Y} = \{-1, 1\}$  et la classe de prédicteurs considérée est celle des classifieurs linéaires :  $\mathcal{F} = \mathcal{H}_d$ .

**DÉFINITION.** — *Régression logistique.* — Soit  $(x_i, y_i)_{i \in [n]} \in \mathcal{S}(\mathbb{R}^d, \{-1, 1\})$  un échantillon d'apprentissage. La régression logistique donne le classifieur  $h_{\hat{w}, \hat{b}}$  où :

$$(\hat{w}, \hat{b}) = \arg \min_{(w, b) \in \mathbb{R}^d \times \mathbb{R}} \left\{ \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-y_i(\langle w, x_i \rangle + b))) \right\}.$$

**REMARQUE.** — La quantité  $\log(1 + \exp(-y_i(\langle w, x_i \rangle + b)))$  pénalise d'autant plus l'exemple  $(x_i, y_i)$  qu'il est loin dans le mauvais demi-espace.

**REMARQUE.** — La régression logistique peut être vue comme une minimisation du risque empirique dans un problème auxiliaire de régression (voir TD).

