Chapitre III
Prédicteurs linéaires
Cycle pluridisciplinaire d'études supérieures
Université Paris sciences et lettres

I. Préliminaires

Notation. — Pour $d \geqslant 1$ 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 :

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

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

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

Remarque. — La fonction ϕ étant quelconque, un *prédicteur* linéaire n'est pas nécessairement une *application* linéaire.

Exemple. — *Classifieurs linéaires.* — On note sign la fonction donnant le signe strict d'un nombre réel, c'est-à-dire :

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

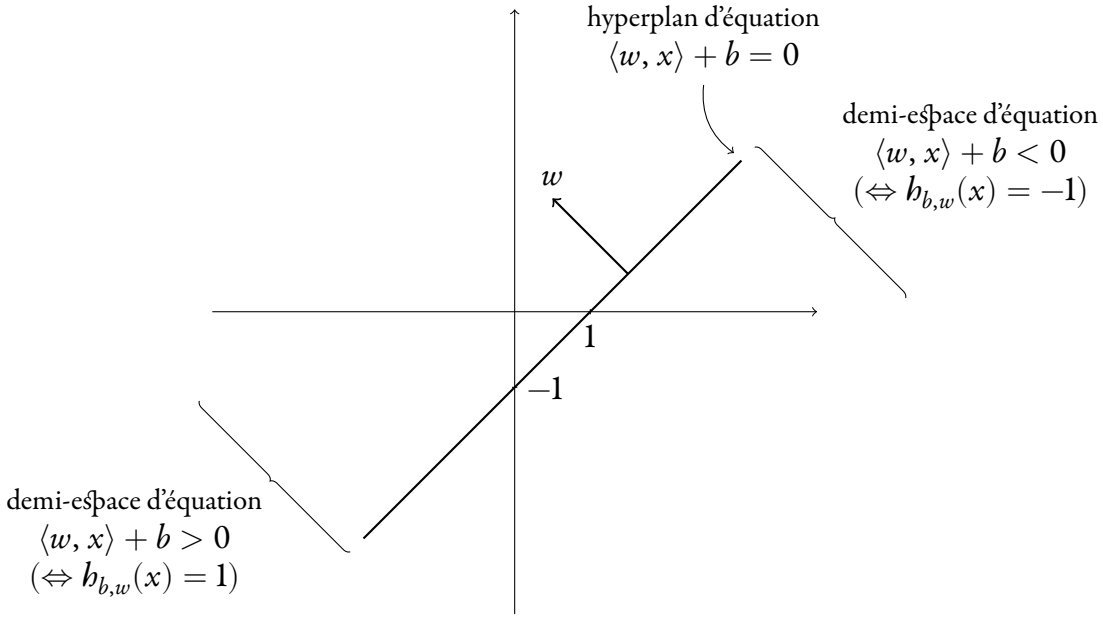
Pour $(w, b) \in \mathbb{R}^d \times \mathbb{R}$, on note $h_{w,b}$ le *classifieur linéaire* associé, qui est défini par :

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

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 lesquel $h_{w,b}$ 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 $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ 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 $n \geqslant 1$ 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 :

$$\widehat{f} = \operatorname*{arg\,min}_{f\in\mathcal{L}_d} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \right\}.$$

Autrement dit, $\widehat{f}$ est le minimiseur du risque empirique [1] associé à la classe $\mathcal{L}_d$ et à la perte quadratique.

1. voir la définition dans le Chapitre I.

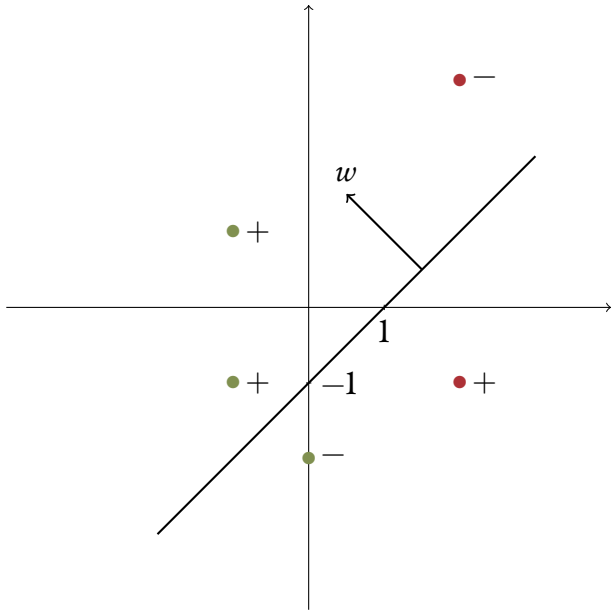
3. CLASSIFICATION BINAIRE : PERCEPTRON

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 :

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

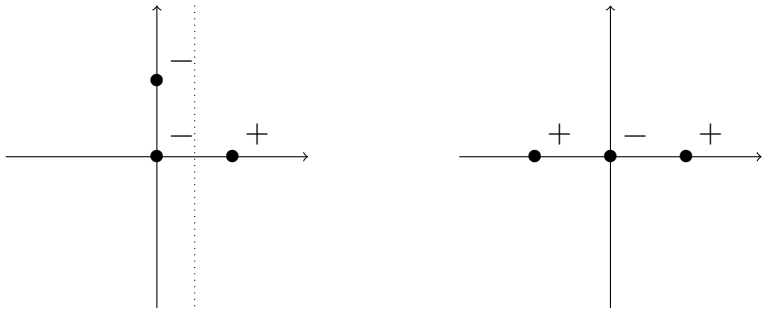
EXEMPLE. — On reprend l'exemple où $\mathcal{X} = \mathbb{R}^2$, $w = (-1, 1)$ et $b = 1$. On représente ci-après un échantillon de 4 exemples où les étiquettes (égales à $+1$ ou -1) sont signalées par des signes $+$ et $-$. Les trois points en vert correspondent aux exemples bien prédits par $h_{b,w}$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



DÉFINITION. — *Séparabilité linéaire.* — Soit $n \geqslant 1$. 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* $t \geqslant 1$, on considère le classifieur (temporaire) $\hat{h}^{(t)} = h_{w^{(t)}, b^{(t)}}$.

— *S'il existe* $i \in [n]$ tel que

$$\hat{h}^{(t)}(x_i) \neq y_i \quad (\Longleftrightarrow \quad y_i \left(\langle w^{(t)}, x_i\rangle + b^{(t)}\right) \leqslant 0),$$

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 $\hat{h} = \hat{h}^{(t)}$.

Remarque. — Une intuition sur le fonctionnement de l'algorithme Perceptron est la suivante. L'algorithme souhaite minimiser le nombre d'exemples mal prédits, c'est-à-dire d'indices $i \in [n]$ tels que :

$$y_i \left(\langle w, x\rangle + b\right) \leqslant 0,$$

où (w, b) désigne ici les paramètres de l'étape courante. S'il existe un tel exemple, on peut voir que la mise à jour des paramètres $w \leftarrow w + y_i x_i$ et $b \leftarrow b + y_i$ fait augmenter la quantité $y_i \left(\langle w, x\rangle + b\right)$, pour essayer de la rendre positive.

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 :

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

Remarque. — La quantité $\log\left(1 + \exp\left(-y_i(\langle w, x_i\rangle + b)\right)\right)$ 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 [2] de régression (voir TD).

2. Un problème *auxiliaire* est un autre problème, en lien avec le problème initial, que l'on considère.