Chapitre V
Machines à vecteurs de support
Cycle pluridisciplinaire d'études supérieures
Université Paris sciences et lettres

Dans leur forme la plus basique, les SVM (*support vector machines*) sont des algorithmes de classification binaire donnant des classifieurs linéaires, à l'instar de la régression logistique.

I. Cadre et rappels

On se place dans un cadre de classification binaire. Soit $\mathscr{X} = \mathbb{R}^d$ l'ensemble d'entrées et $\mathscr{Y} = \{-1, 1\}$ l'ensemble de sorties.

Pour $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ et $x \in \mathbb{R}^d$, on rappelle qu'on note :

$$\begin{aligned} g_{w,b}(x) &= \langle w, x \rangle + b, \\ h_{w,b}(x) &= \operatorname{sign}(g_{w,b}(x)), \end{aligned}$$

et $\mathscr{F} = \{h_{w,b}\}_{\substack{w \in \mathbb{R}^d \\ b \in \mathbb{R}}}$ la classe des *classifieurs linéaires*. On dit que le classifieur linéaire $h_{w,b}$ *prédit correctement* un exemple $(x_0, y_0) \in \mathbb{R}^d \times \{-1, 1\}$ si, et seulement si $h_{w,b}(x_0) = y_0$, ce qui est aussi équivalent à :

$$y_0(\langle w, x_0 \rangle + b) > 0.$$

On appelle *hyperplan* (w, b) l'ensemble :

$$\left\{x \in \mathbb{R}^d \mid \langle w, x \rangle + b = 0\right\}.$$

Une notion importante dans la définition des SVM est la *marge*.

Définition. — *Marge.* — Soit $(w, b) \in \mathbb{R}^d$. On appelle *marge de l'hyperplan* (w, b) l'ensemble :

$$\left\{x \in \mathbb{R}^d \,\middle|\, -1 < \langle w, x\rangle + b < 1\right\}.$$

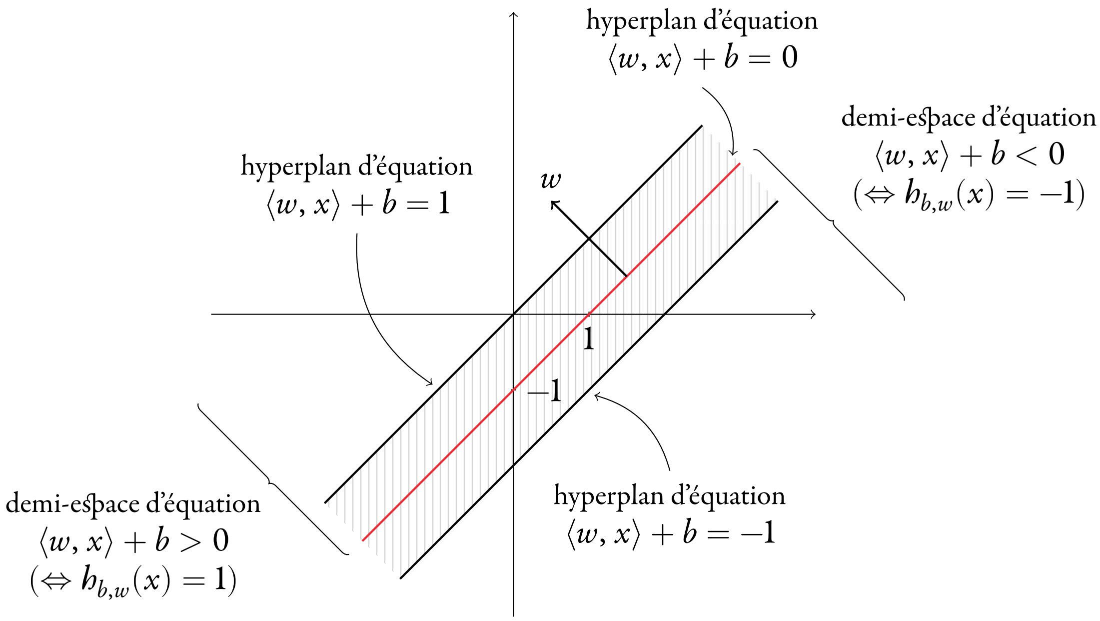

Figure 1. — Exemple avec $w = (-1, 1)$ et $b = 1$. La marge est représentée par les hachures.

Remarque. — Plus $\|w\|_2$ est grand, plus la marge est étroite.

2. Cas séparable : Hard-SVM

On introduit d'abord l'algorithme Hard-SVM qui n'est défini que pour des échantillons d'apprentissage *linéairement séparables* dont on rappelle la définition.

Définition. — Soit $n \geqslant 1$. Un échantillon $\mathrm{S} = (x_i, y_i)_{i\in[n]} \in \mathcal{S}(\mathcal{X}, \mathcal{Y})$ est *linéairement séparable* s'il existe $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ tel que :

$$\forall i \in [n], \qquad y_i = h_{w,b}(x_i).$$

$$(\iff \forall i \in [n], \quad y_i(\langle w, x_i \rangle + b) > 0)\,.$$

On dit alors que $h_{w,b}$ (ou (w, b)) est un hyperplan séparateur pour S.

REMARQUE. — Si S est linéairement séparable, il existe une infinité d'hyperplans séparateurs. Le Hard-SVM va choisir celui qui maximise la distance au point x_i le plus proche.

LEMME. — *Soit* $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ *et* $x \in \mathbb{R}^d$. *Si* $\|w\|_2 = 1$*, alors :*

$$\min_{\substack{x' \in \mathbb{R}^d \\ \langle w, x' \rangle + b = 0}} \|x - x'\|_2 = |\langle w, x \rangle + b|\,.$$

Démonstration. — Voir TD. □

DÉFINITION. — *Hard-SVM.* — Soit $S = (x_i, y_i)_{i \in [n]} \in \mathscr{S}(\mathscr{X}, \mathscr{Y})$ un échantillon linéairement séparable. L'algorithme *Hard-SVM* donne le classifieur linéaire $h_{\hat{w},\hat{b}}$ où $(\hat{w}, \hat{b})$ est solution de :

$$\begin{aligned} \text{maximiser en } (w, b) \quad & \min_{i \in [n]} |\langle w, x_i \rangle + b| \\ \text{soumis aux contraintes} \quad & \|w\|_2 = 1 \\ & \forall i \in [n],\ y_i(\langle w, x_i \rangle + b) > 0. \end{aligned} \tag{$*$}$$

REMARQUE. — Si S n'est pas linéairement séparable, le problème $(*)$ n'a pas de solution, car par définition de la séparabilité, il n'existe pas de paramètres $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ qui satisfont la seconde contrainte.

PROPOSITION. — *Formulation équivalente du Hard-SVM.* — *Le Hard-SVM donne le classifieur linéaire* $h_{\tilde{w},\tilde{b}}$ *où* $(\tilde{w}, \tilde{b})$ *est solution de :*

$$\begin{aligned} \textit{minimiser en } (w, b) \quad & \|w\|_2 \\ \textit{soumis aux contraintes} \quad & \forall i \in [n], \quad y_i(\langle w, x_i \rangle + b) \geqslant 1. \end{aligned}$$

Démonstration. — Voir TD. □

3. CAS GÉNÉRAL : SOFT-SVM

Le Hard-SVM n'est défini que lorsque l'échantillon d'apprentissage est linéairement séparable, il donne alors un hyperplan séparateur. On présente dans ce paragraphe le Soft-SVM qui est lui toujours bien défini, et qui donne un classifieur linéaire en tolérant que des points x_i se retrouvent du mauvais côté de l'hyperplan.

DÉFINITION. — Soit $n \geqslant 1$ et $S = (x_i, y_i)_{i \in [n]} \in \mathscr{S}(\mathscr{X}, \mathscr{Y})$ un échantillon. Soit $\lambda > 0$ un réel. L'algorithme *Soft-SVM* (avec hyperparamètre λ) donne le classifieur linéaire $h_{\widehat{w},\widehat{b}}$ où $(\widehat{w}, \widehat{b}, \widehat{\xi}_1, \dots, \widehat{\xi}_n)$ est solution de :

$$
\begin{array}{rll}
\text{minimiser en } (w, b, \xi_1, \dots, \xi_n) & \lambda \|w\|_2^2 + \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \xi_i & \\
\text{soumis aux contraintes} & \forall i \in [n], & y_i(\langle w, x_i \rangle + b) \geqslant 1 - \xi_i \\
 & \forall i \in [n], & \xi_i \geqslant 0.
\end{array}
$$

REMARQUE. — Contrairement au Hard-SVM, le Soft-SVM autorise des *violations* (mesurées par les ξ_i), c'est-à-dire des points x_i qui se retrouvent soit dans la marge, soit du mauvais côté de l'hyperplan.

REMARQUE. — La quantité minimisée comporte deux termes : $\lambda \|w\|_2^2$, qui encourage les grandes marges ; et $\frac{1}{n}\sum_{i=1}^{n} \xi_i$ qui pénalise les violations. Le choix de l'hyperparamètre λ contrôle l'importance relative des ces deux termes.

REMARQUE. — De façon similaire à la minimisation du risque empirique régularisée, lorsque l'hyperparamètre λ croît, la complexité de l'algorithme diminue et son biais augmente.

PROPOSITION. — *Formulation équivalente du Soft-SVM. — Le Soft-SVM (avec hyperparamètre λ) donne le classifieur $h_{\widehat{w},\widehat{b}}$ où :*

$$
(\widehat{w}, \widehat{b}) = \operatorname*{arg\,min}_{(w,b) \in \mathbb{R}^d \times \mathbb{R}} \left\{ \lambda \|w\|_2^2 + \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i (\langle w, x_i \rangle + b)) \right\}
$$

REMARQUE. — La proposition précédente montre que le Soft-SVM peut être vu comme une minimisation du risque empirique avec régularisation Ridge dans un problème auxiliaire de régression défini par : l'ensemble d'entrées $\widetilde{\mathscr{X}} = \mathbb{R}^d$, l'ensemble de sorties $\tilde{\mathscr{Y}} = \mathbb{R}$, le même échantillon d'apprentissage $\tilde{\mathrm{S}} = \mathrm{S}$, la classe de prédicteurs :

$$\tilde{\mathscr{F}} = \left\{ g_{w,b} \colon x \mapsto \langle w, x \rangle + b \right\}_{\substack{w \in \mathbb{R}^d \\ b \in \mathbb{R}}},$$

ainsi que la fonction de perte dite *charnière* (*hinge loss* en anglais) :

$$\forall y, y' \in \tilde{\mathscr{Y}}, \quad \tilde{\ell}(y, y') = \max(0,\ 1 - yy').$$