Introduction au Machine Learning
Support Vector Machine

Théo Lopès-Quintas

BPCE Payment Services,
Université Paris Dauphine

2022 - 2025

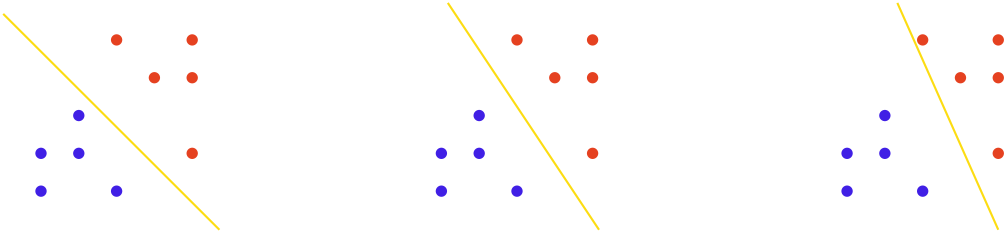
Dans le cas séparable
Trouver un hyperplan séparateur

On considère un problème de classification, donc on a accès à un dataset défini comme :

$$\mathcal{D} = \left\{ (x_i, y_i) \mid \forall i \leqslant n,\ x_i \in \mathbb{R}^d, y_i \in \{-1, 1\} \right\}$$

En supposant que les données soient linéairement séparables, nous aimerions être capables de trouver un hyperplan qui sépare parfaitement les données.

$$\langle \overbrace{\boxed{w}}^{\textbf{Direction } w \in \mathbb{R}^d}, x \rangle + \underbrace{\boxed{b}}_{\textbf{Offset } b \in \mathbb{R}} = 0 \qquad \text{(Hyperplan)}$$



(a) $w = \begin{pmatrix} -0.9 \\ -0.4 \end{pmatrix}$ et $b = 0.50$ **(b)** $w = \begin{pmatrix} -0.86 \\ -0.57 \end{pmatrix}$ et $b = 0.43$ **(c)** $w = \begin{pmatrix} -0.7 \\ -0.7 \end{pmatrix}$ et $b = 0.50$

Figure – Exemple de trois hyperplans possibles pour séparer parfaitement les données

Dans le cas séparable

Trouver un hyperplan séparateur : et la marge ?

Dire que l'on sépare parfaitement les données revient à dire que :

$$\begin{cases} \langle w, x_i\rangle + b \geqslant 0 & \text{pour } y_i = +1 \\ \langle w, x_i\rangle + b < 0 & \text{pour } y_i = -1 \end{cases} \iff y_i(\langle w, x_i\rangle + b) > 0$$

Il nous reste à définir *la plus grande marge*. On peut montrer que pour n'importe quel point x, la distance entre x et l'hyperplan est $\dfrac{|\langle w, x\rangle + b|}{\|w\|}$

Exercice 1 (Marge de valeur 1)

En remarquant que :

$$\forall \lambda > 0,\ \langle(\lambda w), x\rangle + (\lambda b) > 0 \iff \langle w, x\rangle + b > 0$$

Montrer que l'on peut définir la largeur totale de la marge comme $\gamma = \dfrac{2}{\|w\|_2}$.

Dans le cas séparable

Problèmes équivalents

On peut donc écrire le problème initial que l'on veut résoudre comme :

$$
\begin{aligned}
(w,b)^* = \underset{(w,b)\in\mathbb{R}^d\times\mathbb{R}}{\arg\max} \quad & \frac{2}{\|w\|_2} \\
\text{tel que} \quad & \forall i \leqslant n,\ y_i(\langle w, x_i\rangle + b) - 1 \geqslant 0
\end{aligned}
$$

Mais nous préférerions l'écrire comme un problème avec un $\arg\min$.

Exercice 2 (Transformation du problème)

Montrer que :

$$
\underset{w\in\mathbb{R}^d}{\arg\max} \frac{2}{\|w\|_2} = \underset{w\in\mathbb{R}^d}{\arg\min} \frac{1}{2}\|w\|_2^2
$$

Finalement, le problème que l'on cherche à résoudre est :

$$
\begin{aligned}
(w,b)^* = \underset{(w,b)\in\mathbb{R}^d\times\mathbb{R}}{\arg\min} \quad & \frac{1}{2}\|w\|_2^2 \\
\text{tel que} \quad & \forall i \leqslant n,\ y_i(\langle w, x_i\rangle + b) - 1 \geqslant 0
\end{aligned}
$$

Dans le cas séparable

Motivation pour le cas non séparable

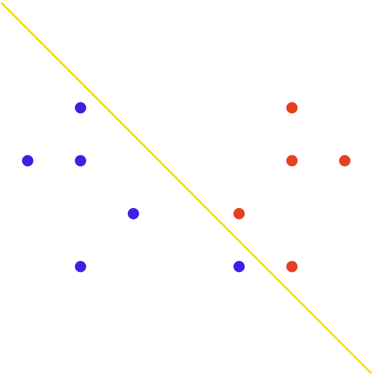


Figure – Cas séparable où il serait préférable d'avoir une erreur

Dans le cas non séparable

Formalisation

Scalaire pour moduler la pénalisation

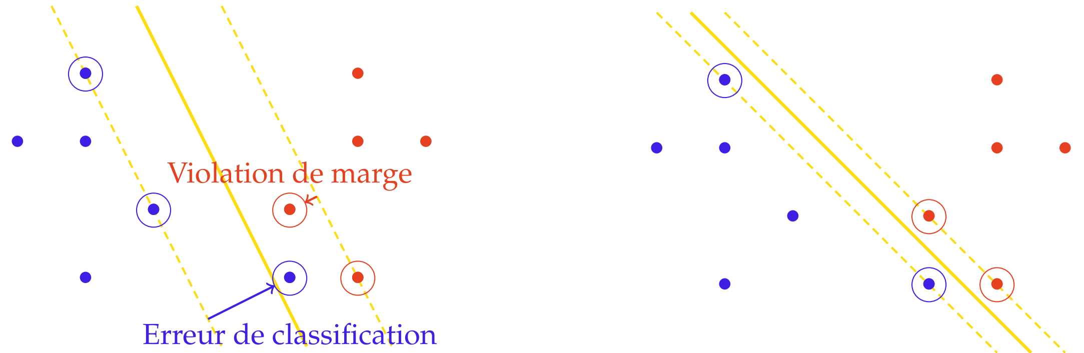
$$
\begin{aligned}
(w,b)^* = \underset{(w,b)\in\mathbb{R}^d\times\mathbb{R}}{\arg\min} \quad & \frac{1}{2}\|w\|_2 + \nu \sum_{i=1}^{n} \varepsilon_i \\
\text{tel que} \quad & \forall i \leqslant n,\ y_i(\langle w, x_i\rangle + b) \geqslant 1 - \varepsilon_i \\
& \forall i \leqslant n,\ \varepsilon_i \geqslant 0
\end{aligned}
$$

Dans le cas non séparable

Formalisation

Scalaire pour moduler la pénalisation

$$
\begin{aligned}
(w, b)^* = \underset{(w,b)\in\mathbb{R}^d\times\mathbb{R}}{\arg\min} \quad & \frac{1}{2}\|w\|_2 + \nu \sum_{i=1}^{n} \varepsilon_i \\
\text{tel que} \quad & \forall i \leqslant n,\ y_i(\langle w, x_i\rangle + b) \geqslant 1 - \varepsilon_i \\
& \forall i \leqslant n,\ \varepsilon_i \geqslant 0
\end{aligned}
$$



(a) Avec $\nu = 1$

(b) Avec $\nu = 10$

Figure – Différence d'apprentissage pour deux valeurs de pénalisations différentes

Dans le cas non séparable

Lagrangien

On peut écrire notre problème comme :

$f : \Omega \to \mathbb{R}$

$$
\begin{aligned}
x^* = \underset{x \in \Omega}{\arg\min} \quad & \boxed{f(x)} \\
\textit{telque} \quad & \forall i \leqslant \boxed{m} \,,\ g_i(x) \leqslant 0
\end{aligned}
$$

Nombre de contraintes d'inégalités g

Definition 1 (Lagrangien)

Pour un problème d'optimisation sous contraintes d'inégalités comme défini précédemment, on définit le lagrangien du problème comme l'application $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}_+^m \to \mathbb{R}$:

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$$

Prenons une application :

$U \subset \mathbb{R}^d$

$$L : \boxed{U} \times \boxed{P} \to \mathbb{R}$$

$P \subset \mathbb{R}_+^m$

DANS LE CAS NON SÉPARABLE

POINT SELLE

Definition 2 (Point selle)

On dit que le couple $(u^*, p^*) \in U \times P$ est un point selle de L si, et seulement si,

$$\forall (u,p) \in U \times P,\ L(u^*, p) \leqslant L(u^*, p^*) \leqslant L(u, p^*)$$

On peut visualiser un point selle d'une fonction :

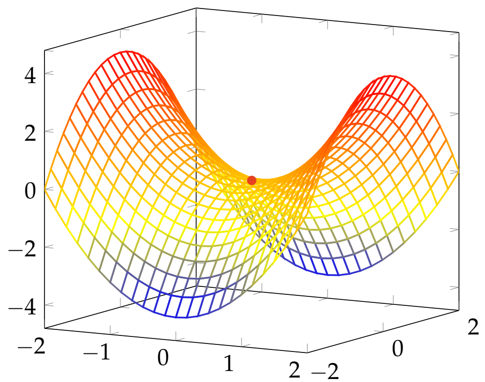


Figure – Point selle de la fonction $(x,y) \mapsto x^2 - y^2$

DANS LE CAS NON SÉPARABLE

PROBLÈME PRIMAL ET DUAL (1/2)

Definition 3 (Problème primal et dual)

Avec les notations précédentes, on définit les fonctions :

$$\mathcal{I}(u) = \sup_{p \in P} L(u,p) \text{ et } \mathcal{G}(p) = \inf_{u \in U} L(u,p)$$

On appelle problème **primal** le problème de minimisation :

$$\inf_{u \in U} \mathcal{I}(u) = \inf_{u \in U} \sup_{p \in P} L(u,p)$$

On appelle problème **dual** le problème de maximisation :

$$\sup_{p \in P} \mathcal{G}(p) = \sup_{p \in P} \inf_{u \in U} L(u,p)$$

On a défini deux problèmes qui se ressemblent beaucoup, mais qui sont différents. Le résultat suivant nous montre que ces deux problèmes sont intimement liés au point selle (u^*, p^*).

Dans le cas non séparable

Problème primal et dual (2/2)

Théorème 1 (Dualité)

Le point $(u^*, p^*) \in U \times P$ *est un point selle de L si, et seulement si,*

$$L(u^*, p^*) = \mathcal{I}(u^*) = \mathcal{G}(p^*)$$

Autrement dit, un point selle résout à la fois le problème primal et dual !

Proposition 1

Avec les notations précédentes, si (w^*, λ^*) *est un point selle de* $\mathcal{L}$ *le lagrangien alors* w^* *est un minimum global de* f *sur* Ω.
De plus, si f *et les contraintes* $(g_i)_{i \leqslant m}$ *sont de classe* $\mathcal{C}^1$ *et convexe, alors :*

$$\exists \lambda^* \in \mathbb{R}_+^m, \begin{cases} \nabla f(w^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(w^*) & = 0 \\ \forall i \leqslant m, \quad \lambda_i^* \geqslant 0 \\ \forall i \leqslant m, \quad \lambda_i^* g_i(w^*) = 0 \\ \forall i \leqslant m, \quad g_i(w^*) \leqslant 0 \end{cases}$$

Résolution
Problèmes

$$(w,b)^* = \underset{(w,b)\in\mathbb{R}^d\times\mathbb{R}}{\arg\min} \quad \frac{1}{2}\|w\|_2 + \nu\sum_{i=1}^{n}\varepsilon_i$$
$$\text{tel que} \quad \forall i \leqslant n,\ y_i(\langle w, x_i\rangle + b) \geqslant 1-\varepsilon_i$$
$$\forall i \leqslant n,\ \varepsilon_i \geqslant 0$$

$$\mathcal{L}\left((w,b,\varepsilon_i),(\mu_i,\delta_i)\right) = \frac{1}{2}\|w\|^2 + \nu\sum_{i=1}^{n}\varepsilon_i - \left(\sum_{i=1}^{n}\mu_i\left[y_i(\langle w, x_i\rangle + b) - (1-\varepsilon)\right] + \sum_{i=1}^{n}\delta_i\varepsilon_i\right) \quad \text{(Lagrangien)}$$

$$\begin{cases}
\dfrac{\partial\mathcal{L}}{\partial w}((w,b,\varepsilon_i),(\mu_i,\delta_i)) & = 0\\
\dfrac{\partial\mathcal{L}}{\partial b}((w,b,\varepsilon_i),(\mu_i,\delta_i)) & = 0\\
\forall i \leqslant n,\ \dfrac{\partial\mathcal{L}}{\partial \varepsilon_i}((w,b,\varepsilon_i),(\mu_i,\delta_i)) & = 0\\
\forall i \leqslant n,\ \mu_i & \geqslant 0\\
\forall i \leqslant n,\ \delta_i & \geqslant 0\\
\forall i \leqslant n,\ \mu_i\left[y_i(\langle w, x_i\rangle + b) - (1-\varepsilon_i)\right] & = 0\\
\forall i \leqslant n,\ \delta_i\varepsilon_i & = 0
\end{cases}$$

KERNEL TRICK

INTUITION

Espace de départ de dimension d

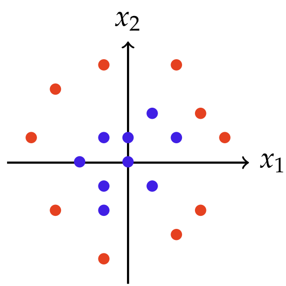
$$\phi : \boxed{\mathbb{R}^d} \to \boxed{\mathbb{R}^{d'}}$$

Espace d'arrivée de dimension $d' > d$

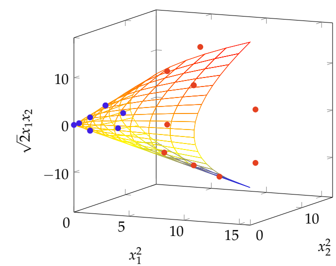
Definition 4 (Noyau)

Avec les notations précédentes, on appelle K le noyau associé à ϕ, l'application définie par :

$$K(x, y) = \langle \phi(x), \phi(y) \rangle$$



(a) Dans l'espace initial

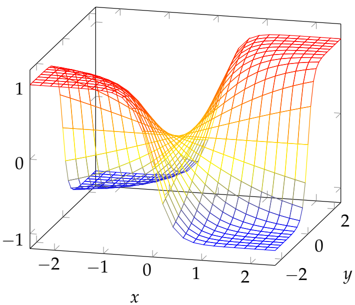


(b) Dans l'espace de dimension supérieure

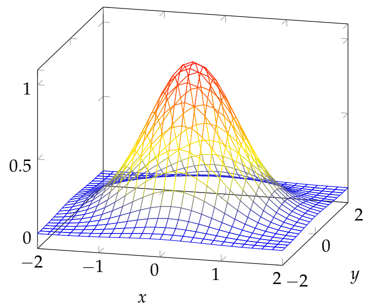
Figure – Application d'un noyau pour classifier deux groupes

En pratique
Kernel classique

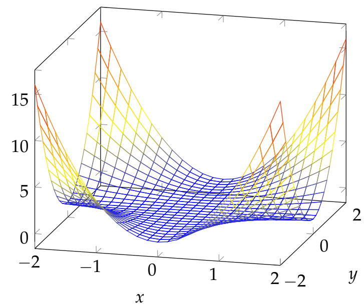
- **Noyau linéaire** : $K(x_1, x_2) = \langle x_1, x_2 \rangle$
- **Noyau polynomial** : $K(x_1, x_2) = (\gamma \langle x_1, x_2 \rangle + r)^d$
- **Noyau Radial Basis Function** : $K(x_1, x_2) = \exp\left\{-\frac{\|x_1 - x_2\|^2}{\sigma^2}\right\} = \exp\left\{-\gamma \|x_1 - x_2\|^2\right\}$
- **Noyau sigmoid** : $K(x_1, x_2) = \tanh\left(\gamma \langle x_1, x_2 \rangle + r\right)$



(a) Noyau sigmoid



(b) Noyau RBF



(c) Noyau polynomial

Figure – Représentation des différents noyaux classiques (pour $\gamma = 1$, $r = 0$ et $d = 2$)

En pratique

Principaux Hyperparamètres

- **Paramétrer le noyau**
  - kernel : pour définir le noyau avec lequel on veut travailler
  - degree : degré du noyau polynômial si kernel = 'poly', ignorer sinon
  - gamma : coefficient du noyau pour les noyaux 'poly', 'rbf' ou 'sigmoid'.
  - coef0 : terme indépendant (r) dans le noyau polynomial ou sigmoid, ignorer sinon.
- **Paramétrer l'algorithme**
  - C : la pénalité ν
  - max_iter : le nombre maximum d'itérations du solveur numérique pour résoudre le problème
  - tol : tolérance pour le critère d'arrêt du solveur