

# INTRODUCTION AU MACHINE LEARNING

## SUPPORT VECTOR MACHINE

**Théo Lopès-Quintas**

BPCE Payment Services,  
Université Paris Dauphine

2022 - 2024

## INTRODUCTION

*The shortest path between two truths in the real domain passes through the complex domain*

— Jacques Hadamard (1991)



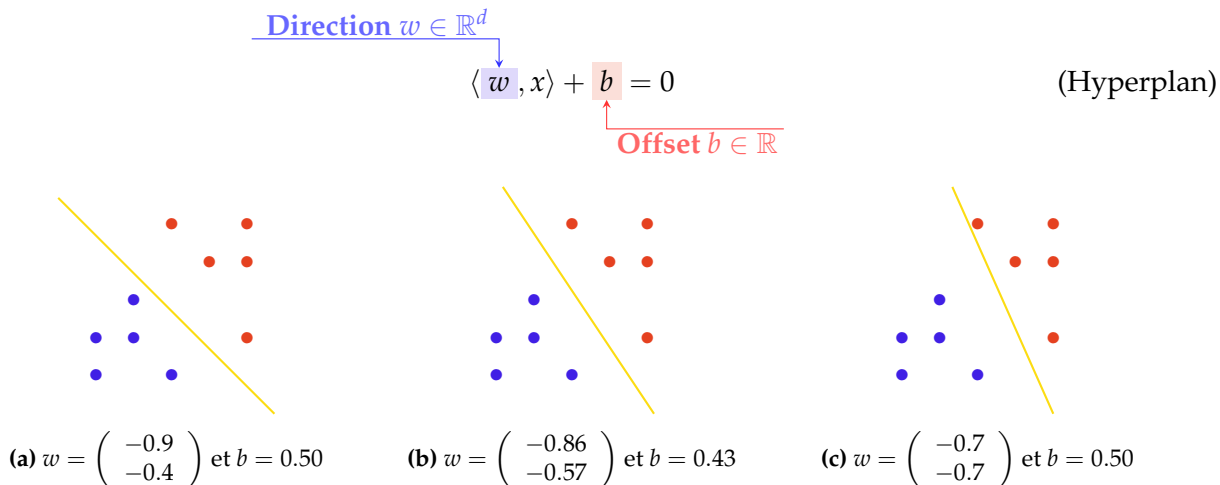
## DANS LE CAS SÉPARABLE

### MARGE

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

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

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



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

## DANS LE CAS SÉPARABLE

### MARGE

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

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

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  $\frac{|\langle w, x \rangle + b|}{\|w\|}$

### 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  $\gamma = \frac{2}{\|w\|_2}$ .

## DANS LE CAS SÉPARABLE

### PROBLÈMES ÉQUIVALENT

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

$$(w, b)^* = \arg \max_{(w, b) \in \mathbb{R}^d \times \mathbb{R}} \frac{2}{\|w\|_2} \\ \text{tel que} \quad \forall i \leq n, y_i(\langle w, x_i \rangle + b) - 1 \geq 0$$

Mais nous préférierions l'écrire comme un problème avec un arg min.

### Exercice 2 (Transformation du problème)

*Montrer que :*

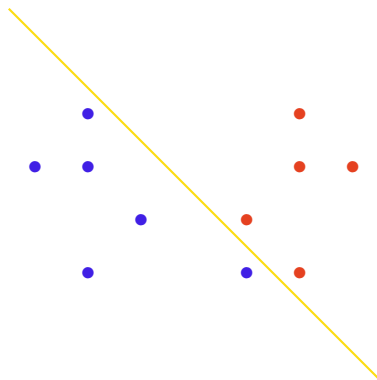
$$\arg \max_{w \in \mathbb{R}^d} \frac{2}{\|w\|_2} = \arg \min_{w \in \mathbb{R}^d} \frac{1}{2} \|w\|_2^2$$

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

$$(w, b)^* = \arg \min_{(w, b) \in \mathbb{R}^d \times \mathbb{R}} \frac{1}{2} \|w\|_2^2 \\ \text{tel que} \quad \forall i \leq n, y_i(\langle w, x_i \rangle + b) - 1 \geq 0$$

## 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

$$(w, b)^* = \arg \min_{(w, b) \in \mathbb{R}^d \times \mathbb{R}} \frac{1}{2} \|w\|_2 + \nu \sum_{i=1}^n \varepsilon_i$$

tel que

$$\forall i \leq n, y_i(\langle w, x_i \rangle + b) \geq 1 - \varepsilon_i$$
$$\forall i \leq n, \varepsilon_i \geq 0$$

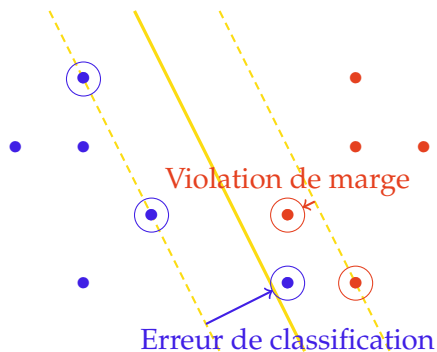
# DANS LE CAS NON SÉPARABLE

## FORMALISATION

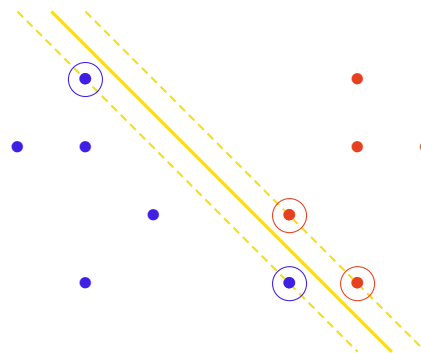
$$(w, b)^* = \arg \min_{(w, b) \in \mathbb{R}^d \times \mathbb{R}} \\ \text{tel que}$$

Scalaire pour moduler la pénalisation

$$\frac{1}{2} \|w\|_2 + \nu \sum_{i=1}^n \varepsilon_i$$
$$\forall i \leq n, y_i(\langle w, x_i \rangle + b) \geq 1 - \varepsilon_i$$
$$\forall i \leq n, \varepsilon_i \geq 0$$



(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 :

$$\begin{array}{c} f : \Omega \rightarrow \mathbb{R} \\ \downarrow \\ x^* = \arg \min_{x \in \Omega} \quad f(x) \\ \text{telque} \quad \forall i \leq m, g_i(x) \leq 0 \end{array}$$

↑  
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  $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}_+^m \rightarrow \mathbb{R}$  :

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

Prenons une application :

$$\begin{array}{c} U \subset \mathbb{R}^d \\ \downarrow \\ L : U \times P \rightarrow \mathbb{R} \end{array}$$

↑  
 $P \subset \mathbb{R}_+^m$

## 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,

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

On peut visualiser un point selle d'une fonction :

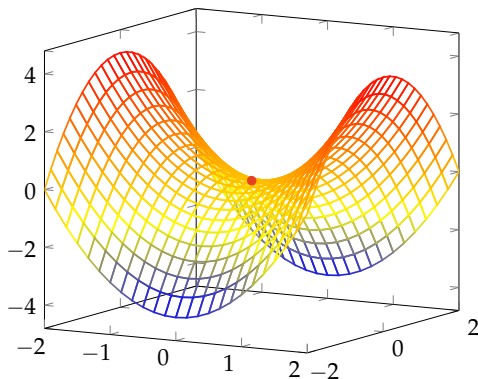


Figure – Point selle de la fonction  $(x, y) \mapsto x^2 - y^2$

## 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  $(w^*, \lambda^*)$  est un point selle de  $\mathcal{L}$  le lagrangien alors  $w^*$  est un minimum global de  $f$  sur  $\Omega$ .

De plus, si  $f$  et les contraintes  $(g_i)_{i \leq m}$  sont de classe  $\mathcal{C}^1$  et convexe, alors :

$$\exists \lambda^* \in \mathbb{R}_+^m, \left\{ \begin{array}{l} \nabla f(w^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(w^*) = 0 \\ \forall i \leq m, \lambda_i^* \geq 0 \\ \forall i \leq m, \lambda_i^* g_i(w^*) = 0 \\ \forall i \leq m, g_i(w^*) \leq 0 \end{array} \right.$$

# RÉSOLUTION

## PROBLÈMES

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

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

$$\left\{ \begin{array}{ll}
 \frac{\partial \mathcal{L}}{\partial w}((w, b, \varepsilon_i), (\mu_i, \delta_i)) & = 0 \\
 \frac{\partial \mathcal{L}}{\partial b}((w, b, \varepsilon_i), (\mu_i, \delta_i)) & = 0 \\
 \forall i \leq n, \ \frac{\partial \mathcal{L}}{\partial \varepsilon_i}((w, b, \varepsilon_i), (\mu_i, \delta_i)) & = 0 \\
 \forall i \leq n, \ \mu_i & \geq 0 \\
 \forall i \leq n, \ \delta_i & \geq 0 \\
 \forall i \leq n, \ \mu_i [y_i(\langle w, x_i \rangle + b) - (1 - \varepsilon_i)] & = 0 \\
 \forall i \leq n, \ \delta_i \varepsilon_i & = 0
 \end{array} \right.$$

# KERNEL TRICK

## INTUITION

Espace de départ de dimension  $d$

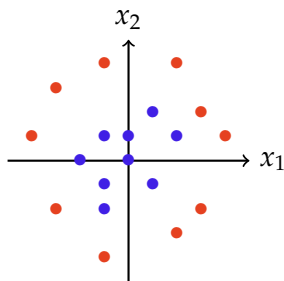
$$\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$$

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

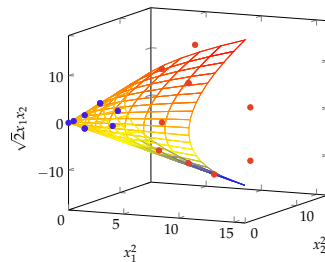
### Definition 4 (Noyau)

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

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



(a) Dans l'espace initial



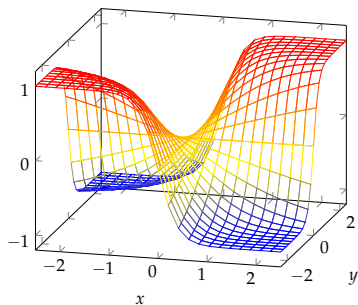
(b) Dans l'espace de dimension supérieure

**Figure** – Application d'un noyau pour classer deux groupes

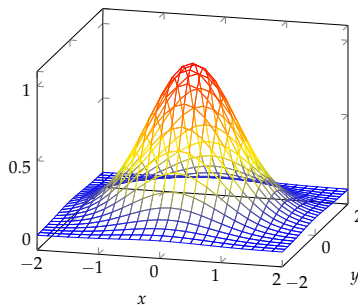
## EN PRATIQUE

### KERNEL CLASSIQUE

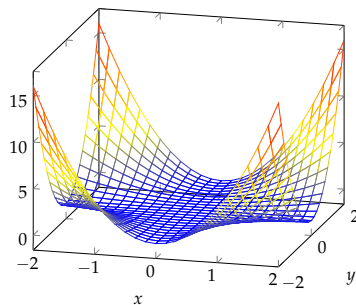
- ▶ **Noyau linéaire** :  $K(x_1, x_2) = \langle x_1, x_2 \rangle$
- ▶ **Noyau polynomial** :  $K(x_1, x_2) = (\gamma \langle x_1, x_2 \rangle + r)^d$
- ▶ **Noyau Radial Basis Function** :  $K(x_1, x_2) = \exp \left\{ -\frac{\|x_1 - x_2\|^2}{\sigma^2} \right\} = \exp \left\{ -\gamma \|x_1 - x_2\|^2 \right\}$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



(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  $\nu$
- max\_iter : le nombre maximum d'itérations du solveur numérique pour résoudre le problème
- tol : tolérance pour le critère d'arrêt du solveur