

Machines à vecteurs de support

Bassem Ben Hamed

Control, Optimization and Model Reduction in ML - CIMPA 2025

18 février 2025

Table des matières

1 Classification linéaire

2 Cas séparable

3 Cas non séparable

- On introduit d'abord l'algorithme pour les ensembles de données séparables,
- puis on présente sa version générale pour les ensembles non séparables,
- et enfin on fournit une base théorique des SVMs basée sur la notion de marge.

Table des matières

1 Classification linéaire

2 Cas séparable

3 Cas non séparable

Problème de Classification Linéaire

Soit $X \subset \mathbb{R}^N$ l'espace d'entrée avec $N \geq 1$ et $Y = \{-1, +1\}$ l'espace de sortie. Soit $f : X \rightarrow Y$ la fonction cible.

Le problème de classification binaire consiste à trouver une hypothèse $h \in H$ qui minimise l'erreur de généralisation :

$$R_D(h) = P[h(x) \neq f(x)] \quad (1)$$

L'apprenant reçoit un échantillon S de taille m , tiré i.i.d. de X selon une distribution inconnue D :

$$S = \{(x_1, y_1), \dots, (x_m, y_m)\} \in (X \times Y)^m$$

avec $y_i = f(x_i)$ pour tout $i \in [m]$.

Ensemble d'Hypothèses

Différents ensembles d'hypothèses H peuvent être sélectionnés pour cette tâche. Un ensemble naturel d'hypothèses avec une complexité relativement faible est celui des classificateurs linéaires, ou hyperplans, définis comme suit :

$$H = \{x \mapsto \text{sign}(w \cdot x + b) : w \in \mathbb{R}^N, b \in \mathbb{R}\}. \quad (2)$$

Le problème d'apprentissage est alors appelé problème de classification linéaire.

Hyperplans

L'équation générale d'un hyperplan dans \mathbb{R}^N est donnée par :
 $w \cdot x + b = 0$, où $w \in \mathbb{R}^N$ est un vecteur non nul normal à l'hyperplan et $b \in \mathbb{R}$ est un scalaire. Une hypothèse de la forme $x \mapsto \text{sign}(w \cdot x + b)$ étiquette positivement tous les points d'un côté de l'hyperplan et négativement tous les autres.

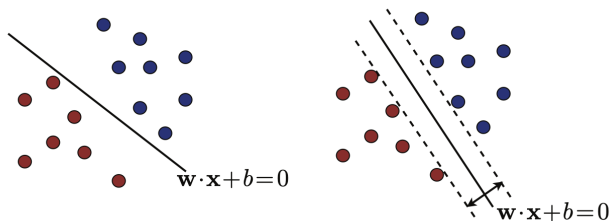


Figure: Deux hyperplans de séparation possibles. La figure de droite montre un hyperplan qui maximise la marge.

Table des matières

1 Classification linéaire

2 Cas séparable

3 Cas non séparable

Hyperplans Séparateurs

On suppose que l'échantillon d'entraînement S est linéairement séparable, c'est-à-dire qu'il existe un hyperplan qui sépare parfaitement les points positifs et négatifs, comme illustré dans la figure précédente.

Cela signifie qu'il existe $(w, b) \in (\mathbb{R}^N \setminus \{0\}) \times \mathbb{R}$ tel que :

$$\forall i \in [m], \quad y_i(w \cdot x_i + b) \geq 0. \quad (3)$$

Toutefois, il existe une infinité d'hyperplans séparateurs. Lequel un algorithme d'apprentissage devrait-il choisir ?

La solution du SVM repose sur la notion de **marge géométrique**.

Definition

La marge géométrique $\rho_h(x)$ d'un classificateur linéaire $h : x \mapsto w \cdot x + b$ en un point x est sa distance euclidienne à l'hyperplan $w \cdot x + b = 0$:

$$\rho_h(x) = \frac{|w \cdot x + b|}{\|w\|_2}. \quad (4)$$

La marge géométrique ρ_h d'un classificateur linéaire h pour un échantillon $S = (x_1, \dots, x_m)$ est la plus petite marge géométrique parmi tous les points de l'échantillon :

$$\rho_h = \min_{i \in [m]} \rho_h(x_i),$$

c'est-à-dire la distance de l'hyperplan définissant h aux points les plus proches de l'échantillon.

Hyperplan à Marge Maximale

- Un point de test est correctement classé par un hyperplan séparateur de marge ρ tant qu'il reste à une distance inférieure ou égale à ρ des échantillons d'entraînement partageant le même label.
- Pour l'hyperplan SVM, ρ est maximal, garantissant ainsi la meilleure séparation possible.

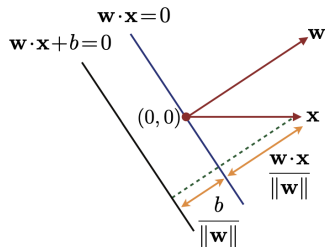


Figure: Illustration de la marge géométrique d'un point \mathbf{x} dans le cas où $\mathbf{w} \cdot \mathbf{x} > 0$ et $b > 0$

Problème d'Optimisation Primal

Les équations et le problème d'optimisation qui définissent la solution SVM :

$$\rho = \max_{w,b} \min_{i \in [m]} \frac{|w \cdot x_i + b|}{\|w\|} = \max_{w,b} \min_{i \in [m]} \frac{y_i(w \cdot x_i + b)}{\|w\|}. \quad (5)$$

En effet, puisque l'échantillon est linéairement séparable, on peut imposer la contrainte : $\forall i \in [m], \quad y_i(w \cdot x_i + b) \geq 0$.

En normalisant de sorte que $\min_{i \in [m]} y_i(w \cdot x_i + b) = 1$, on obtient :

$$\rho = \max_{w,b : \min_{i \in [m]} y_i(w \cdot x_i + b) = 1} \frac{1}{\|w\|} = \max_{w,b : \forall i \in [m], y_i(w \cdot x_i + b) \geq 1} \frac{1}{\|w\|} \quad (6)$$

Hyperplans Marginaux

Hyperplan à marge maximale et les **hyperplans marginaux**

- Ce sont les hyperplans parallèles à l'hyperplan séparateur, passant par les points les plus proches des classes positives et négatives.
- Puisqu'ils sont parallèles à l'hyperplan séparateur, ils possèdent le même vecteur normal w .
- Comme $|w \cdot x + b| = 1$ pour les points les plus proches, leurs équations sont : $w \cdot x + b = \pm 1$.

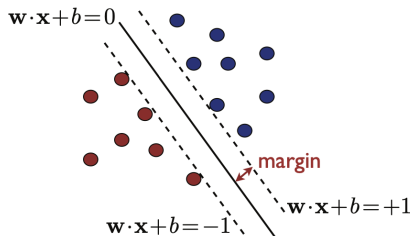


Figure: Solution de l'hyperplan de marge maximale de (6). Les hyperplans marginaux sont représentés par des lignes pointillées

Problème d'Optimisation Convexe

Objectif : Maximiser $\frac{1}{\|w\|}$ équivalent à minimiser $\frac{1}{2}\|w\|^2$

- **Formulation :**

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2 \quad (7)$$

Sous contraintes : $y_i(w \cdot x_i + b) \geq 1, \quad \forall i \in [m]$

- **Propriétés :**

- La fonction objectif $F : w \mapsto \frac{1}{2}\|w\|^2$ est infiniment différentiable
- Gradient : $\nabla F(w) = w$
- Hessienne : $\nabla^2 F(w) = I$ (matrice identité)
- Les valeurs propres de l'Hessienne sont strictement positives ($\nabla^2 F(w) \succ 0$)
- Par conséquent, F est strictement convexe

- **Contraintes :** Définies par des fonctions affines :

$$g_i(w, b) = 1 - y_i(w \cdot x_i + b)$$

- **Conclusion :** Le problème d'optimisation admet une solution unique (propriété non partagée par tous les algorithmes)

Retour sur le problème d'optimisation (7)

- Les contraintes sont affines et donc qualifiées.
- La fonction objectif ainsi que les contraintes affines sont convexes et différentiables.
- **Conditions nécessaires :**
 - Les conditions de Karush-Kuhn-Tucker (KKT) s'appliquent à l'optimum.
 - Ces conditions sont fondamentales pour analyser l'algorithme.
- **Objectif :**
 - Démontrer plusieurs propriétés cruciales de l'algorithme.
 - Dériver le problème d'optimisation dual associé aux SVMs.

Variables de Lagrange et Lagrangien

Variables de Lagrange $\alpha_i \geq 0$, $i \in [m]$, associées aux m contraintes et notons α le vecteur $(\alpha_1, \dots, \alpha_m)^\top$.

Le Lagrangien peut alors être défini pour tout $w \in \mathbb{R}^N$, $b \in \mathbb{R}$, et $\alpha \in \mathbb{R}^m$ par :

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^m \alpha_i [y_i(w \cdot x_i + b) - 1] \quad (8)$$

Conditions de Karush-Kuhn-Tucker (KKT)

Les conditions KKT sont obtenues en posant le gradient du Lagrangien par rapport aux variables primitives w et b égal à zéro et en écrivant les conditions de complémentarité :

$$\nabla_w \mathcal{L} = 0 \Rightarrow w = \sum_{i=1}^m \alpha_i y_i x_i \quad (9)$$

$$\nabla_b \mathcal{L} = 0 \Rightarrow \sum_{i=1}^m \alpha_i y_i = 0 \quad (10)$$

$$\forall i, \alpha_i [y_i (w \cdot x_i + b) - 1] = 0 \Rightarrow \alpha_i = 0 \text{ ou } y_i (w \cdot x_i + b) = 1 \quad (11)$$

Rôle des Vecteurs de Support (1)

Par l'équation (9), le vecteur de poids w à la solution du problème SVM est une combinaison linéaire des vecteurs d'apprentissage x_1, \dots, x_m .

- Un vecteur x_i apparaît dans cette combinaison si et seulement si $\alpha_i \neq 0$.
- Ces vecteurs sont appelés **vecteurs de support**.
- Par les conditions de complémentarité (11), si $\alpha_i \neq 0$, alors $y_i(w \cdot x_i + b) = 1$.
- Ainsi, les vecteurs de support se trouvent sur les hyperplans marginaux : $w \cdot x_i + b = \pm 1$.

Rôle des Vecteurs de Support (2)

- Les vecteurs de support définissent entièrement l'hyperplan à maximum de marge, justifiant ainsi le nom de l'algorithme.
- Les vecteurs qui ne se trouvent pas sur les hyperplans marginaux n'affectent pas la définition de ces hyperplans.
- Bien que la solution w soit unique, les vecteurs de support ne le sont pas.
- En dimension N , $N + 1$ points suffisent pour définir un hyperplan.
- Ainsi, lorsque plus de $N + 1$ points se trouvent sur un hyperplan marginal, plusieurs choix sont possibles pour les $N + 1$ vecteurs de support.

Pour dériver la forme duale du problème d'optimisation (7), on substitue la définition de w en fonction des variables duales (9) dans le Lagrangien et on applique la contrainte (10). On obtient

$$\mathcal{L} = \frac{1}{2} \left\| \sum_{i=1}^m \alpha_i y_i x_i \right\|^2 - \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^m \alpha_i y_i b + \sum_{i=1}^m \alpha_i \quad (12)$$

Ainsi

$$\mathcal{L} = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (13)$$

Problème d'optimisation dual dans le cas séparable :

$$\max_{\alpha} \quad \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (14)$$

Sous contraintes : $\alpha_i \geq 0$ et $\sum_{i=1}^m \alpha_i y_i = 0, \quad \forall i \in [m].$

Propriétés de la Fonction Objectif et Problème d'Optimisation

La fonction objectif $G : \alpha \mapsto \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$ est infiniment différentiable.

- Son Hessien est donné par $\nabla^2 G = -A$, avec $A = y_i x_i \cdot y_j x_j$.
- A est la matrice de Gram associée aux vecteurs $y_1 x_1, \dots, y_m x_m$ et est donc positive semi-défini.
- Cela montre que $\nabla^2 G \leq 0$ et que G est une fonction concave.
- Les contraintes étant affines et convexes, le problème de maximisation (14) est un problème d'optimisation convexe.
- Puisque G est une fonction quadratique de α , ce problème d'optimisation dual est également un problème de programmation quadratique (QP).
- Tant pour le cas primal que dual, des solveurs QP généraux et spécialisés peuvent être utilisés pour obtenir la solution.

Équivalence des Problèmes Primal et Dual

Les problèmes primal et dual sont équivalents, c'est-à-dire que la solution α du problème dual (14) peut être utilisée directement pour déterminer l'hypothèse retournée par les SVMs, en utilisant l'équation (9) :

$$h(x) = \text{sgn}(w \cdot x + b) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i (x_i \cdot x) + b \right). \quad (15)$$

Puisque les vecteurs de support se trouvent sur les hyperplans marginaux, pour tout vecteur de support x_i , on a : $w \cdot x_i + b = y_i$, et ainsi, b peut être obtenu via :

$$b = y_i - \sum_{j=1}^m \alpha_j y_j (x_j \cdot x_i). \quad (16)$$

Le problème d'optimisation dual (14) et les expressions (15) et (16) révèlent une propriété importante des SVMs :

- La solution de l'hypothèse dépend uniquement des produits scalaires entre les vecteurs et non des vecteurs eux-mêmes.

Cette observation est clé, et son importance est cruciale pour les méthodes de noyau.

Expression du Marge Géométrique

L'équation (16) peut maintenant être utilisée pour dériver une expression simple du marge géométrique ρ en termes de α .

En utilisant (16), qui est valide pour tout i avec $\alpha_i > 0$, et en multipliant les deux côtés par $\alpha_i y_i$, on obtient :

$$\sum_{i=1}^m \alpha_i y_i b = \sum_{i=1}^m \alpha_i y_i^2 - \sum_{i=1}^m \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (17)$$

En utilisant le fait que $y_i^2 = 1$ avec l'équation (9), on obtient :

$$0 = \sum_{i=1}^m \alpha_i - \|w\|^2 \quad (18)$$

En notant que $\alpha_i > 0$, on obtient l'expression suivante du marge ρ en termes de norme L_1 de α :

$$\rho^2 = \frac{1}{\|w\|_2^2} = \frac{1}{\sum_{i=1}^m \alpha_i} = \frac{1}{\|\alpha\|_1} \quad (19)$$

Table des matières

1 Classification linéaire

2 Cas séparable

3 Cas non séparable

Problème : Dans la plupart des cas pratiques, les données d'entraînement ne sont pas linéairement séparables.

- Pour tout hyperplan $w \cdot x + b = 0$, il existe $x_i \in S$ tel que :

$$y_i(w \cdot x_i + b) \not\geq 1. \quad (20)$$

- Les contraintes du cas linéairement séparable ne peuvent pas toutes être satisfaites simultanément.

Solution : Une version relaxée des contraintes peut être imposée :

- Introduction des variables de relaxation $\xi_i \geq 0$.
- Pour chaque $i \in [m]$, la contrainte devient :

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i. \quad (21)$$

- Permet une meilleure gestion des erreurs de classification.

Les Variables de Relaxation

Definition

Les variables ξ_i sont appelées variables de relaxation et sont utilisées pour assouplir les contraintes d'optimisation.

- ξ_i mesure la distance par laquelle x_i viole l'inégalité $y_i(w \cdot x_i + b) \geq 1$.
- Un vecteur x_i avec $\xi_i > 0$ est considéré comme un outlier.
- Si $0 < y_i(w \cdot x_i + b) < 1$, x_i est bien classé mais reste un outlier.

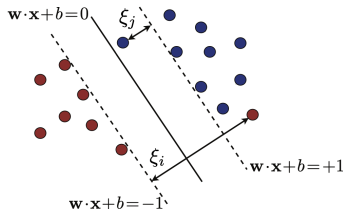


Figure: Hyperplan de séparation avec le point x_i mal classé et le point x_j correctement classé, mais avec une marge inférieure à 1

Problème d'optimisation :

- Deux objectifs contradictoires :

- ① Minimiser l'erreur empirique en réduisant la somme des ξ_i :

$$\sum_{i=1}^m \xi_i, \quad \text{ou plus généralement} \quad \sum_{i=1}^m \xi_i^p \quad \text{pour un } p \geq 1.$$

- ② Maximiser la marge $\rho = \frac{1}{\|w\|}$.

- Un compromis est nécessaire :

- Un hyperplan avec une marge plus large favorise une meilleure généralisation.
- Cependant, une marge trop grande peut engendrer plus d'outliers, augmentant ainsi les valeurs des ξ_i .
- L'optimisation repose donc sur un équilibre entre minimisation des erreurs et maximisation de la marge.

Formulation mathématique du SVM non-séparable :

$$\min_{w,b,\xi} \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i^p \quad (22)$$

sous les contraintes :

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i \in [m].$$

- $C \geq 0$ contrôle le compromis entre la maximisation de la marge et la pénalité des erreurs.
- Typiquement, C est déterminé via validation croisée.

- Comme dans le cas séparable, le problème (22) est une optimisation convexe :
 - Les contraintes sont affines, donc convexes.
 - La fonction objectif est convexe pour tout $p \geq 1$.
- La norme $\|\xi\|_p = \sum_{i=1}^m \xi_i^p$ est convexe.
- Différentes valeurs de p influencent la pénalisation des erreurs :
 - $p = 1$: perte hinge (hinge loss).
 - $p = 2$: perte hinge quadratique (quadratic hinge loss).

Perte Hinge (Hinge Loss)

Définition : La perte hinge est une fonction de perte utilisée pour l'entraînement des SVMs.

$$L_{\text{hinge}}(y, f(x)) = \max(0, 1 - yf(x))$$

- $y \in \{-1, 1\}$: label réel.
- $f(x) = w \cdot x + b$: prédiction du modèle.

Interprétation :

- Si $yf(x) \geq 1$, la classification est correcte, la perte est nulle.
- Si $0 \leq yf(x) < 1$, la classification est correcte mais proche de la frontière, une pénalité est appliquée.
- Si $yf(x) < 0$, la classification est incorrecte, et la perte augmente linéairement.

Avantage : La perte hinge favorise une classification avec une marge large, améliorant ainsi la généralisation.

Fonctions de Perte pour SVM

- Les pertes hinge et hinge quadratique sont des bornes convexes de la perte zéro-un.
- Elles sont adaptées à l'optimisation.
- La perte hinge ($p = 1$) est la plus utilisée en pratique.

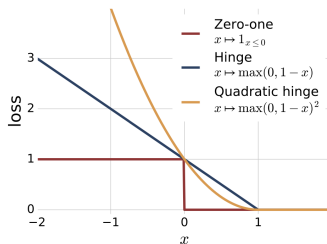


Figure: La perte de hinge et la perte de hinge quadratique fournissent toutes deux des limites supérieures convexes à la perte binaire zéro-un.

Propriétés :

- Comme dans le cas séparable, les contraintes sont affines.
- La fonction objectif ainsi que les contraintes affines sont convexes et différentiables.
- On peut appliquer les conditions KKT à l'optimum.

Définition : On introduit les variables de Lagrange :

- $\alpha_i \geq 0, i \in [m]$, associées aux m premières contraintes.
- $\mu_i \geq 0, i \in [m]$, associées aux contraintes de non-négativité des variables de relaxation.
- Notations :
 - $\alpha = (\alpha_1, \dots, \alpha_m)^T$.
 - $\mu = (\mu_1, \dots, \mu_m)^T$.

Formulation du Lagrangien :

$$\mathcal{L}(w, b, \xi, \alpha, \mu) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i [y_i (w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^m \mu_i \xi_i. \quad (23)$$

- Cette formulation permet d'analyser l'algorithme et de démontrer ses propriétés essentielles.
- Elle conduit à la formulation du problème dual d'optimisation des SVMs.

Les conditions KKT sont obtenues en annulant le gradient du Lagrangien par rapport aux variables primales w , b et ξ_i , et en écrivant les conditions de complémentarité.

$$\nabla_w \mathcal{L} = w - \sum_{i=1}^m \alpha_i y_i x_i = 0 \implies w = \sum_{i=1}^m \alpha_i y_i x_i \quad (24)$$

$$\nabla_b \mathcal{L} = - \sum_{i=1}^m \alpha_i y_i = 0 \implies \sum_{i=1}^m \alpha_i y_i = 0 \quad (25)$$

$$\nabla_{\xi_i} \mathcal{L} = C - \alpha_i - \mu_i = 0 \implies \alpha_i + \mu_i = C \quad (26)$$

$$\forall i, \alpha_i [y_i (w \cdot x_i + b) - 1 + \xi_i] = 0 \implies \alpha_i = 0 \vee y_i (w \cdot x_i + b) = 1 - \xi_i \quad (27)$$

$$\forall i, \mu_i \xi_i = 0 \implies \mu_i = 0 \vee \xi_i = 0 \quad (28)$$

Definition

- Le vecteur de poids w à la solution du problème SVM est une combinaison linéaire des vecteurs d'entraînement x_1, \dots, x_m .
- Un vecteur x_i apparaît si et seulement si $\alpha_i \neq 0$.
- Ces vecteurs sont appelés **vecteurs de support**.

Deux catégories de vecteurs de support :

- Si $\xi_i = 0$, alors $y_i(w \cdot x_i + b) = 1$ et x_i se trouve sur un hyperplan marginal (comme dans le cas séparable).
- Si $\xi_i \neq 0$, alors x_i est un **outlier** et vérifie $\alpha_i = C$.

Remarque : Bien que le vecteur de poids w soit unique, les vecteurs de support ne le sont pas nécessairement.

Dérivation de la Forme Duale :

- On remplace la définition de w en termes des variables duales (24) dans le Lagrangien.
- On applique ensuite la contrainte donnée par l'équation (25).
- Cela permet d'obtenir la formulation duale du problème d'optimisation contraint (22).

$$\mathcal{L} = \frac{1}{2} \left\| \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \right\|^2 - \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^m \alpha_i y_i b + \sum_{i=1}^m \alpha_i \quad (29)$$

Objectif : Maximiser la fonction suivante :

$$\mathcal{L} = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (30)$$

Sous contraintes :

- $0 \leq \alpha_i \leq C$ pour tout $i \in [m]$.
- $\sum_{i=1}^m \alpha_i y_i = 0$.

Différence avec le cas séparable : La contrainte supplémentaire $\alpha_i \leq C$ permet de gérer les erreurs de classification.

Détermination de l'hypothèse :

- La solution du problème dual permet de déterminer l'hypothèse retournée par le SVM :

$$h(x) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i (x_i \cdot x) + b \right)$$

- b peut être obtenu à partir de n'importe quel vecteur de support x_i situé sur un hyperplan marginal, c'est-à-dire vérifiant $0 < \alpha_i < C$:

$$b = y_i - \sum_{j=1}^m \alpha_j y_j (x_j \cdot x_i)$$

Importance des Produits Scalaires :

Comme dans le cas séparable, la solution du problème dual et les expressions de $h(x)$ et b montrent une propriété essentielle :

- L'hypothèse finale dépend uniquement des produits scalaires entre les vecteurs et non des vecteurs eux-mêmes.
- Cette propriété est fondamentale pour l'utilisation des noyaux en SVM.