

Chapitre 9 : Support-Vector Machines (SVM)

26 février 2026

Plan du chapitre

Une infinité de séparateurs

Géométrie de la Marge

Optimisation sous contraintes

Optimisation SVM par Dualité

Marge Souple

Le problème des multiples séparateurs

Pour un jeu de données $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ linéairement séparable ($y_i \in \{-1, 1\}$), il existe une **infinité** d'hyperplans qui annulent l'erreur empirique.

Quel est le "meilleur" séparateur au sens de la généralisation ?

La théorie de l'apprentissage statistique (Vapnik-Chervonenkis) montre que tous les hyperplans ne se valent pas. Minimiser la complexité de l'espace d'hypothèses réduit la borne d'erreur en généralisation.

Le principe des SVM (Boser, Guyon, Vapnik - 1992) :

Choisir le séparateur qui ménage la **plus grande marge de sécurité** (espace vide) avec les exemples d'apprentissage les plus proches.

Distance algébrique à un Hyperplan

Soit l'hyperplan $P_{\mathbf{w},b}$ défini par $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$. La distance de x_i à l'hyperplan vaut :

$$d(\mathbf{x}_i, P_{\mathbf{w},b}) = |\lambda| = \frac{|\langle \mathbf{w}, \mathbf{x}_i \rangle + b|}{\|\mathbf{w}\|}$$

Preuve par projection orthogonale

Soit $\mathbf{q}$ le projeté orthogonal d'un point $\mathbf{x}_i$ sur $P_{\mathbf{w},b}$. Le vecteur $\mathbf{x}_i - \mathbf{q}$ est donc colinéaire au vecteur normal $\mathbf{w}$. Posons $\mathbf{x}_i = \mathbf{q} + \lambda \frac{\mathbf{w}}{\|\mathbf{w}\|}$, de sorte que la distance cherchée soit exactement $d = |\lambda|$.

1. $\mathbf{q}$ appartient à l'hyperplan $\implies \langle \mathbf{w}, \mathbf{q} \rangle + b = 0$.
2. Injection de $\mathbf{q}$: $\langle \mathbf{w}, \mathbf{x}_i - \lambda \frac{\mathbf{w}}{\|\mathbf{w}\|} \rangle + b = \langle \mathbf{w}, \mathbf{x}_i \rangle - \lambda \|\mathbf{w}\| + b = 0$.
3. On isole λ pour obtenir la distance :

$$d(\mathbf{x}_i, P_{\mathbf{w},b}) = |\lambda| = \frac{|\langle \mathbf{w}, \mathbf{x}_i \rangle + b|}{\|\mathbf{w}\|}$$

Définition de la Marge

La Marge géométrique (γ)

La marge est la distance entre l'hyperplan séparateur et l'exemple le plus proche (le plus défavorable) du jeu de données $\mathcal{S}$:

$$\gamma = \min_{i \in [n]} \frac{1}{\|\mathbf{w}\|} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b|$$

L'objectif d'un SVM est de trouver les paramètres $(\mathbf{w}, b)$ qui **maximisent** cette marge γ sous contrainte de séparation.

Marge Fonctionnelle et Invariance d'Échelle

Pour une observation bien classée, $\langle \mathbf{w}, \mathbf{x}_i \rangle + b$ et le label $y_i \in \{-1, 1\}$ ont le même signe.

La quantité

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)$$

est strictement positive et s'appelle **marge fonctionnelle**.

Levée de l'indétermination géométrique

L'équation de l'hyperplan est invariante par multiplication par un scalaire $c > 0$. On peut donc *imposer arbitrairement* l'échelle de $\mathbf{w}$ pour que la marge fonctionnelle des vecteurs les plus proches de l'hyperplan vaille exactement 1 :

$$\min_{i=1 \dots n} y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1$$

Le Problème SVM Primal

Grâce à cette normalisation, la distance géométrique de l'hyperplan aux points les plus proches (la demi-marge) vaut exactement $\gamma = \frac{1}{\|\mathbf{w}\|}$.

La largeur totale de la marge séparatrice est $\frac{2}{\|\mathbf{w}\|}$.

Maximiser cette marge revient mathématiquement à **minimiser** $\frac{1}{2}\|\mathbf{w}\|^2$.

Formulation Primale (Hard-Margin QP)

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{sous contraintes} \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1, \quad \forall i \in \{1, \dots, n\} \end{aligned}$$

C'est un Programme Quadratique (QP) **strictement convexe** sous contraintes affines. Il admet un unique minimum global, et la dualité forte s'applique (Théorème de Slater).

Méthode de Lagrange

Pour maximiser la marge, on doit résoudre un problème de minimisation sous contraintes. La méthode est celle des **multiplicateurs de Lagrange**.

Problème primal (P)

$$\min_{\mathbf{u} \in \mathcal{U}} f(\mathbf{u}) \text{ sous les contraintes } g_i(\mathbf{u}) \leq 0, i = 1, \dots, m$$

Toutes les fonctions sont à valeurs réelles et de classe C^1 .

Lagrangien

On note $\boldsymbol{\alpha}^\top = [\alpha_1, \dots, \alpha_m] \in \mathbb{R}^m$, $\mathbf{g}^\top = [g_1, \dots, g_m]$. Le Lagrangien est défini par

$$L(\mathbf{u}, \boldsymbol{\alpha}) = f(\mathbf{u}) + \boldsymbol{\alpha}^\top \mathbf{g}(\mathbf{u})$$

Les α_i sont appelés multiplicateurs de Lagrange ou variables duales.

Problème dual

Fonction duale de Lagrange.

C'est la fonction concave définie par

$$Q(\alpha) = \inf_{\mathbf{u} \in \mathcal{U}} L(\mathbf{u}, \alpha)$$

Si $p_\star$ est le minimum du problème primal, on a $Q(\alpha) \leq p_\star$ pour tout $\alpha_i \geq 0$; ainsi, il est naturel de s'intéresser au maximum de Q .

Problème dual (D)

On cherche

$$\max_{\alpha \in \mathbb{R}_+^m} Q(\alpha) \text{ sous les contraintes } \alpha_i \geq 0, i = 1, \dots, m$$

Ce problème d'optimisation sous contraintes est convexe. On note $d_\star$ le maximum du problème dual.

Dualité faible et forte

Dualité faible.

On appelle dualité faible le fait qu'une solution d_* du problème dual vérifie $d_* \leq p_*$.

Dualité forte.

Parfois $d_* = p_*$, on parle alors de dualité forte et, dans ce cas, il est possible de résoudre le problème primal en passant par le problème dual.

La condition de Slater (contraintes qualifiables)

La condition de Slater est une condition suffisante de dualité forte :
si le problème primal est convexe (f, g_i sont convexes) et qu'il existe au moins un point admissible en lequel les contraintes inégalités non affines sont vérifiées strictement alors il y a dualité forte.

Conditions KKT

Dans le cas dualité forte, si toutes les fonctions sont différentiables, la vérification des conditions KKT est une condition suffisante d'optimalité de la solution.

Les conditions KKT au point $(\mathbf{u}_\star, \alpha_\star)$ sont :

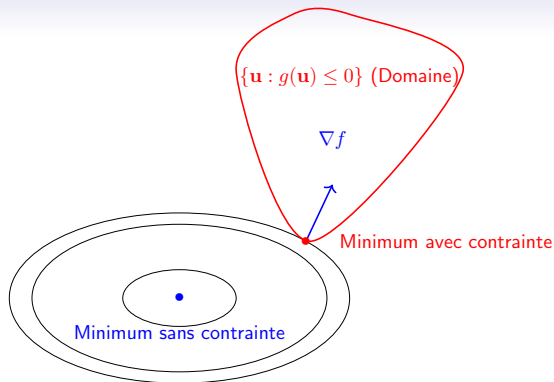
Admissibilité primale : $g_i(\mathbf{u}_\star) \leq 0, i = 1, \dots, m$

Admissibilité duale : $\alpha_i^\star \geq 0, i = 1, \dots, m.$

Complémentarité des contraintes : $\alpha_i^\star g_i(\mathbf{u}_\star) = 0, i = 1, \dots, m.$

Stationnarité : $\nabla f(\mathbf{u}_\star) + \alpha_\star^\top \nabla \mathbf{g}(\mathbf{u}_\star) = 0.$

Si ces conditions sont vérifiées alors $\mathbf{u}_\star$ est un point de minimum du problème primal et $\alpha_\star$ est un point de maximum du problème dual.



Lorsque le minimum sans contrainte n'est pas dans le domaine $\{\mathbf{u} : g(\mathbf{u}) \leq 0\}$ des contraintes non affines, le minimum sous contrainte est à sa frontière : à ce point, la frontière du domaine des contraintes est tangente à une ligne de niveau et les gradients de f et de g sont colinéaires et de directions opposées : $\nabla f = -\alpha \nabla g$ avec $\alpha \geq 0$, c'est-à-dire que l'on a la condition de stationarité $\nabla f + \alpha \nabla g = 0$ avec $\alpha \neq 0$ (contrainte active). Cette condition vaut aussi lorsque le minimum est dans le domaine des contraintes, avec $\alpha = 0$ (contrainte inactive).

Le Lagrangien Primal-Dual

Pour intégrer les contraintes dans la fonction objectif et s'affranchir de la dimension d , on introduit les multiplicateurs de Lagrange $\alpha_i \geq 0$:

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i \left[y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \right]$$

On cherche le point-selle :

$$\min_{\mathbf{w}, b} \max_{\boldsymbol{\alpha} \geq 0} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}).$$

À l'optimum, le gradient de la fonction objectif et celui des contraintes actives sont colinéaires.

Conditions KKT et Annulation des Gradients

Le point-selle optimal implique que les dérivées partielles par rapport aux variables primales ($\mathbf{w}$ et b) s'annulent :

Conditions de stationnarité

$$\begin{aligned}\nabla_{\mathbf{w}} \mathcal{L} = 0 &\implies \mathbf{w} - \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i = \mathbf{0} \implies \mathbf{w}^* = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \\ \partial_b \mathcal{L} = 0 &\implies \sum_{i=1}^n \alpha_i y_i = 0\end{aligned}$$

Théorème de Représentation : Le vecteur normal à l'hyperplan optimal $\mathbf{w}^*$ est une pure **combinaison linéaire** des données d'apprentissage.

Dérivation du Problème Dual de Wolfe

En réinjectant $\mathbf{w}^*$ et la condition sur b dans le Lagrangien, on élimine définitivement les variables primales :

Le Problème Dual (Maximisation)

$$\begin{aligned} \max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ \text{sous contraintes} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \quad \text{et} \quad \alpha_i \geq 0 \quad \forall i \in \{1, \dots, n\} \end{aligned}$$

Pourquoi cette formulation est-elle décisive ?

1. La complexité de l'optimisation dépend de n (nombre d'exemples), et non plus de d (dimension des *features*).
2. Les données n'interviennent **que via leurs produits scalaires** par paires.

Parcimonie : Les Vecteurs de Support

Les conditions de Karush-Kuhn-Tucker (KKT) imposent à l'optimum :

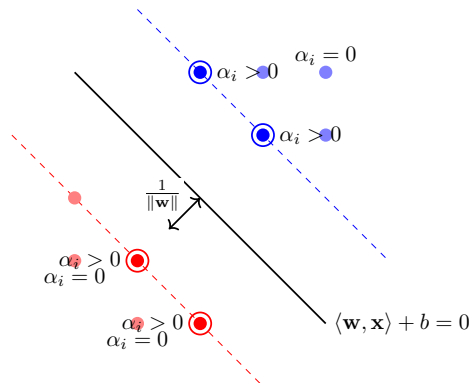
$$\alpha_i \left[y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \right] = 0 \quad \forall i$$

Preuve de la parcimonie :

- Si un point est au-delà de la marge stricte, le crochet est $\neq 0$, donc obligatoirement $\alpha_i = 0$. **Il n'influence pas le modèle.**
- Si $\alpha_i > 0$, alors la marge fonctionnelle vaut exactement 1. Le point repose sur la marge : c'est un **Vecteur de Support**.

Prédiction :

$$f(\mathbf{x}) = \text{sign} \left(\sum_{\alpha_i > 0} \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b^* \right)$$



Les Vecteurs de Support

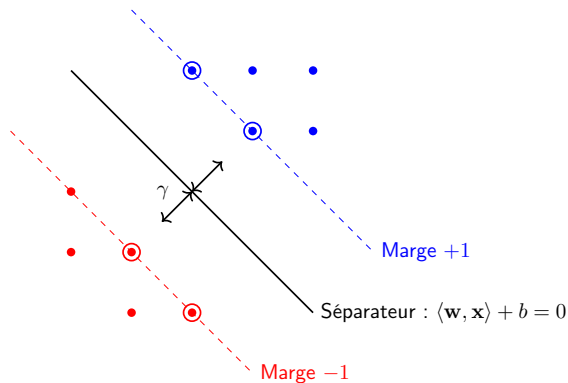


Figure – Seuls les points cerclés (Vecteurs de Support) déterminent l'hyperplan.

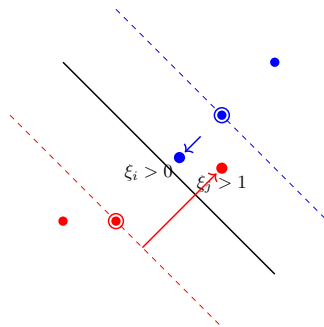
Tolérance aux erreurs (Marge Souple / C-SVM)

Face au chevauchement de classes (bruit), le domaine des contraintes strictes est vide. On autorise des violations géométriques via des variables d'écart (*slack variables*) $\xi_i \geq 0$.

La contrainte devient :

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i$$

- $\xi_i = 0$: Bien classé, hors de la marge.
- $0 < \xi_i \leq 1$: Bien classé, dans la marge.
- $\xi_i > 1$: Mal classé.



Le C-SVM et la Perte Hinge (Charnière)

On modifie l'objectif Primal pour pénaliser linéairement les violations via un paramètre $C > 0$.
(Le problème Dual reste identique, sauf que les multiplicateurs sont bornés : $0 \leq \alpha_i \leq C$).

Formulation Marge Souple (C-SVM Primal)

$$\min_{\mathbf{w}, b, \xi} \quad \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i$$

$$\text{sous contraintes} \quad y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \quad \text{et} \quad \xi_i \geq 0$$

À l'optimum, $\xi_i^* = \max(0, 1 - y_i f(\mathbf{x}_i))$.

Le SVM minimise une perte **Hinge** sous régularisation de Tikhonov :

$$\arg \min_{\mathbf{w}, b} \underbrace{\sum_{i=1}^n \max(0, 1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b))}_{\text{Hinge Loss}} + \underbrace{\frac{1}{2C} \|\mathbf{w}\|^2}_{\text{Régularisation } L_2}$$