



M1 MIASHS

2023/2024 ■ ■

■ CLASSIFICATION SUPERVISÉE ET NON SUPERVISÉE

SVM

MARINE DEMANGEOT

Ce cours s'inspire très largement de l'ouvrage

Introduction au Machine Learning

de Chloé-Agathe Azencott

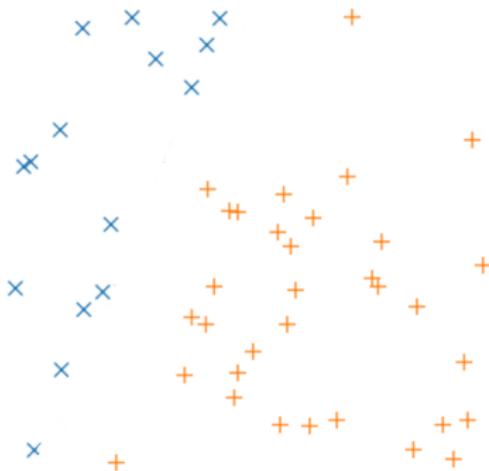
■ Classification binaire



Contexte On observe des **données d'apprentissage**

$$(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \mathcal{Y} = \{-1, 1\}$$

Objectif Apprendre un modèle f à partir de $\{(x_i, y_i)\}_{i \in \{1, \dots, n\}}$ qui prédit (au mieux) l'étiquette y_i pour tout point x_i



■ Quelques classifieurs



► Algorithme des K plus proches voisins

$K \in \mathbb{N}^*$

$$f(x) = \arg \max_{k \in \mathcal{Y}} \sum_{\substack{i \in \{1, \dots, n\} \\ x_i \in \mathcal{N}_K(x)}} \mathbb{1}\{y_i = k\}$$

$\mathcal{N}_K(x)$: K plus proches voisins de x parmi x_1, \dots, x_n

► Classifieurs linéaires $\text{sign}(f(x))$ où $f(x) = \langle \hat{w}, x \rangle + \hat{b}$
 $(\hat{w}, \hat{b}) \in \mathbb{R}^d \times \mathbb{R}$ $\langle \cdot \rangle$: prod. scalaire dans \mathbb{R}^d

- **coût 0/1** (reformulation)

$$(\hat{w}, \hat{b}) = \arg \min_{(w,b) \in \mathbb{R}^d \times \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{[\langle w, x_i \rangle + b] y_i < 0\} \quad (\text{ERM})$$

- **coût logistique** $(\hat{w}, \hat{b}) = \arg \min_{(w,b) \in \mathbb{R}^d \times \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \log [1 + e^{-y_i(\langle w, x_i \rangle + b)}]$ (ERM)

■ Quelques classifieurs

► Algorithme des K plus proches voisins

$K \in \mathbb{N}^*$

$$f(x) = \arg \max_{k \in \mathcal{Y}} \sum_{\substack{i \in \{1, \dots, n\} \\ x_i \in \mathcal{N}_K(x)}} \mathbb{1}\{y_i = k\}$$

$\mathcal{N}_K(x)$: K plus proches voisins de x parmi x_1, \dots, x_n

► Classificateurs linéaires $\text{sign}(f(x))$ où $f(x) = \langle \hat{w}, x \rangle + \hat{b}$

$(\hat{w}, \hat{b}) \in \mathbb{R}^d \times \mathbb{R}$ $\langle \cdot \rangle$: prod. scalaire dans \mathbb{R}^d

- **coût 0/1** (reformulation)

$$(\hat{w}, \hat{b}) = \arg \min_{(w,b) \in \mathbb{R}^d \times \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{[\langle w, x_i \rangle + b] y_i < 0\} \quad (\text{ERM})$$

- **coût logistique** $(\hat{w}, \hat{b}) = \arg \min_{(w,b) \in \mathbb{R}^d \times \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \log [1 + e^{-y_i(\langle w, x_i \rangle + b)}] \quad (\text{ERM})$

- **Machines à vecteurs de support (SVM)** $(\hat{w}, \hat{b}) ?$

■ Données séparables ou non ?



Le choix de l'algorithme SVM utilisé dépend des données. Une première étape est de savoir si les données sont linéairement séparables ou non.

■ Données séparables ou non ?

Le choix de l'algorithme SVM utilisé dépend des données. Une première étape est de savoir si les données sont linéairement séparables ou non.

Linéairement séparables



Non séparables linéairement

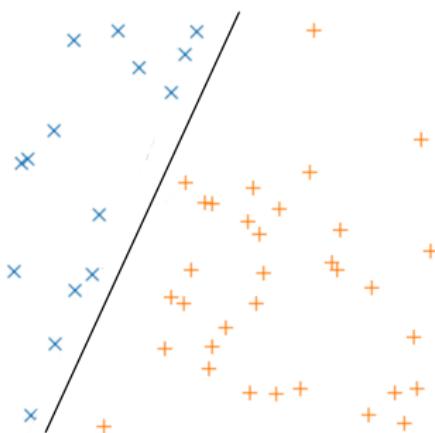


Tirée de l'ouvrage *Introduction au Machine Learning* de Chloé-Agathe Azencott

■ Données séparables ou non ?

Le choix de l'algorithme SVM utilisé dépend des données. Une première étape est de savoir si les données sont linéairement séparables ou non.

Linéairement séparables



Non séparables linéairement



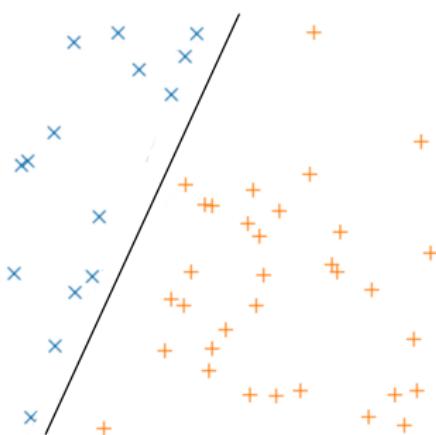
Tirée de l'ouvrage *Introduction au Machine Learning* de Chloé-Agathe Azencott

■ Données séparables ou non ?

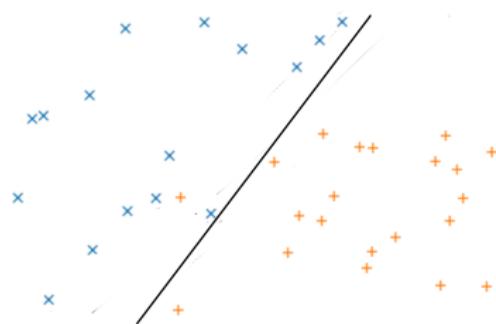


Le choix de l'algorithme SVM utilisé dépend des données. Une première étape est de savoir si les données sont linéairement séparables ou non.

Linéairement séparables



Non séparables linéairement



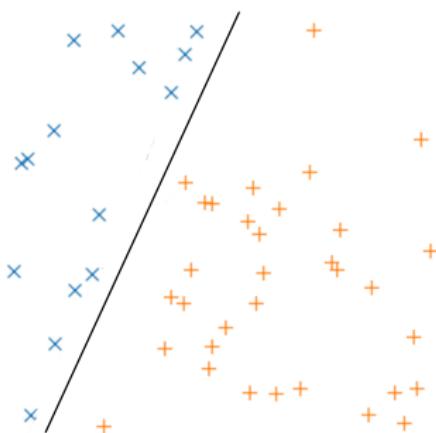
Tirée de l'ouvrage *Introduction au Machine Learning* de Chloé-Agathe Azencott

■ Données séparables ou non ?

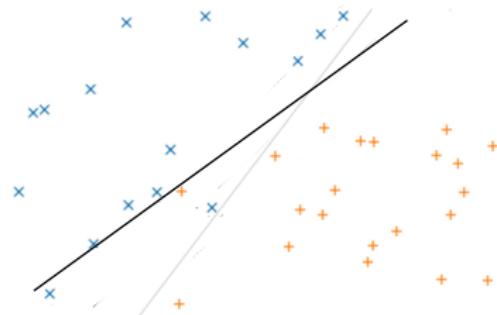


Le choix de l'algorithme SVM utilisé dépend des données. Une première étape est de savoir si les données sont linéairement séparables ou non.

Linéairement séparables



Non séparables linéairement



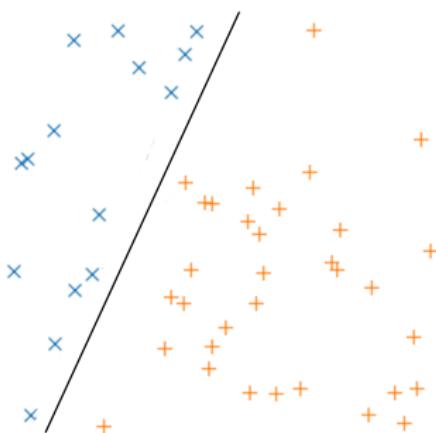
Tirée de l'ouvrage *Introduction au Machine Learning* de Chloé-Agathe Azencott

■ Données séparables ou non ?

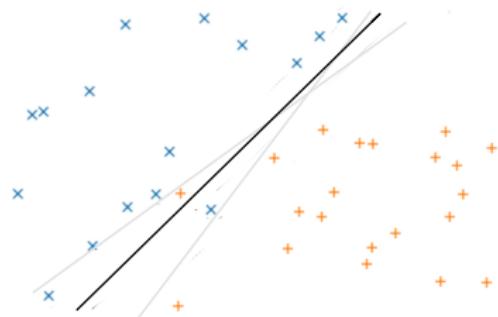


Le choix de l'algorithme SVM utilisé dépend des données. Une première étape est de savoir si les données sont linéairement séparables ou non.

Linéairement séparables



Non séparables linéairement



Tirée de l'ouvrage *Introduction au Machine Learning* de Chloé-Agathe Azencott

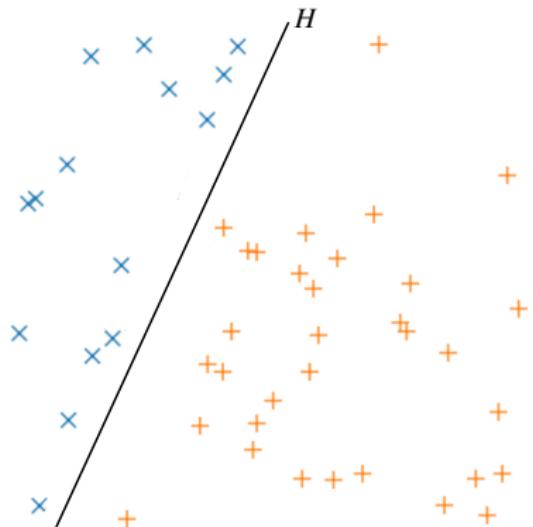
LE CAS LINÉAIREMENT SÉPARABLE : SVM À MARGE RIGIDE

■ Séparabilité linéaire



Définition

Un jeu de données $\{(x_i, y_i)\}_{i=1,\dots,n}$ est **linéairement séparable** s'il existe au moins un hyperplan dans \mathbb{R}^d tel que tous les points x_i positifs ($y_i = 1$) sont d'un côté de cet hyperplan et tous les points x_i négatifs ($y_i = -1$) de l'autre

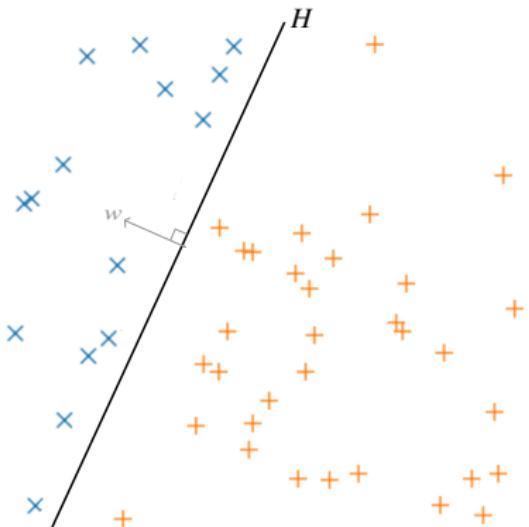


■ Séparabilité linéaire



Définition

Un jeu de données $\{(x_i, y_i)\}_{i=1,\dots,n}$ est **linéairement séparable** s'il existe au moins un hyperplan dans \mathbb{R}^d tel que tous les points x_i positifs ($y_i = 1$) sont d'un côté de cet hyperplan et tous les points x_i négatifs ($y_i = -1$) de l'autre



Équation de $H : \langle w, x \rangle + b = 0$

$(w, b) \in \mathbb{R}^d \times \mathbb{R}$ avec $\vec{w} \perp H$

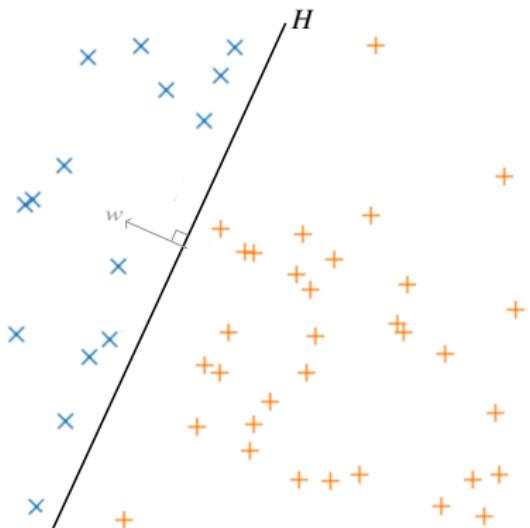
On choisit (w, b) tel que l'angle entre \vec{w} et tout point positif est inférieur à 90°

■ Séparabilité linéaire



Définition

Un jeu de données $\{(x_i, y_i)\}_{i=1,\dots,n}$ est **linéairement séparable** s'il existe au moins un hyperplan dans \mathbb{R}^d tel que tous les points x_i positifs ($y_i = 1$) sont d'un côté de cet hyperplan et tous les points x_i négatifs ($y_i = -1$) de l'autre



Équation de $H : \langle w, x \rangle + b = 0$

$(w, b) \in \mathbb{R}^d \times \mathbb{R}$ avec $\vec{w} \perp H$

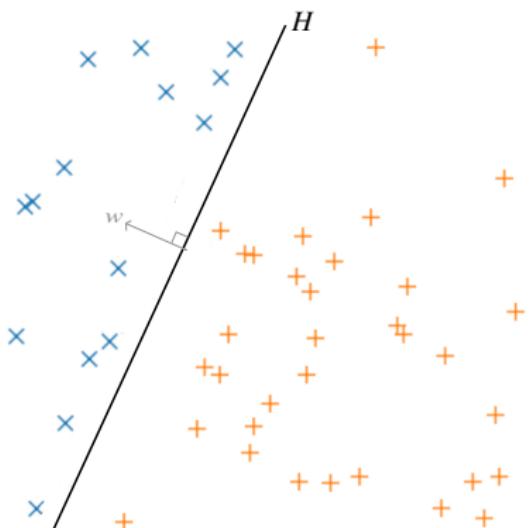
On choisit (w, b) tel que l'angle entre \vec{w} et tout point positif est inférieur à 90°

Pour tout nouveau point x , on estime son label y par sign $[\langle w, x \rangle + b]$

■ Séparabilité linéaire

Définition

Un jeu de données $\{(x_i, y_i)\}_{i=1,\dots,n}$ est **linéairement séparable** s'il existe au moins un hyperplan dans \mathbb{R}^d tel que tous les points x_i positifs ($y_i = 1$) sont d'un côté de cet hyperplan et tous les points x_i négatifs ($y_i = -1$) de l'autre



Équation de $H : \langle w, x \rangle + b = 0$

$(w, b) \in \mathbb{R}^d \times \mathbb{R}$ avec $\vec{w} \perp H$

On choisit (w, b) tel que l'angle entre \vec{w} et tout point positif est inférieur à 90°

Pour tout nouveau point x , on estime son label y par sign $[\langle w, x \rangle + b]$

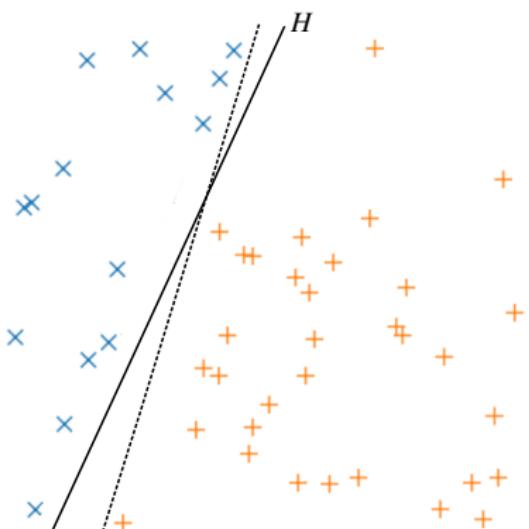
Un hyperplan séparateur H ne fait aucune erreur de classification sur les données d'apprentissage

► (w, b) minimise le risque empirique pour le coût $0/1 \rightarrow$ candidat pour (\hat{w}, \hat{b})

■ Séparabilité linéaire

Définition

Un jeu de données $\{(x_i, y_i)\}_{i=1,\dots,n}$ est **linéairement séparable** s'il existe au moins un hyperplan dans \mathbb{R}^d tel que tous les points x_i positifs ($y_i = 1$) sont d'un côté de cet hyperplan et tous les points x_i négatifs ($y_i = -1$) de l'autre



Équation de $H : \langle w, x \rangle + b = 0$

$(w, b) \in \mathbb{R}^d \times \mathbb{R}$ avec $\vec{w} \perp H$

On choisit (w, b) tel que l'angle entre \vec{w} et tout point positif est inférieur à 90°

Pour tout nouveau point x , on estime son label y par sign $[\langle w, x \rangle + b]$

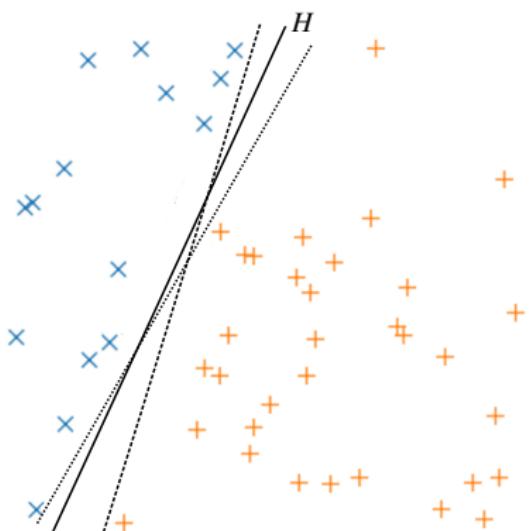
Un hyperplan séparateur H ne fait aucune erreur de classification sur les données d'apprentissage

► (w, b) minimise le risque empirique pour le coût $0/1 \rightarrow$ candidat pour (\hat{w}, \hat{b})

■ Séparabilité linéaire

Définition

Un jeu de données $\{(x_i, y_i)\}_{i=1,\dots,n}$ est **linéairement séparable** s'il existe au moins un hyperplan dans \mathbb{R}^d tel que tous les points x_i positifs ($y_i = 1$) sont d'un côté de cet hyperplan et tous les points x_i négatifs ($y_i = -1$) de l'autre



Équation de $H : \langle w, x \rangle + b = 0$

$(w, b) \in \mathbb{R}^d \times \mathbb{R}$ avec $\vec{w} \perp H$

On choisit (w, b) tel que l'angle entre \vec{w} et tout point positif est inférieur à 90°

Pour tout nouveau point x , on estime son label y par sign $[\langle w, x \rangle + b]$

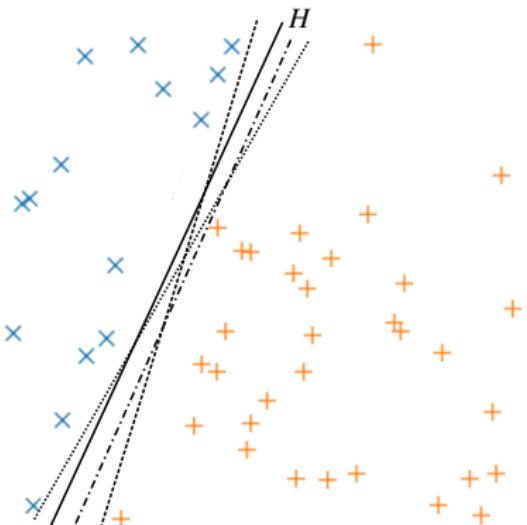
Un hyperplan séparateur H ne fait aucune erreur de classification sur les données d'apprentissage

► (w, b) minimise le risque empirique pour le coût $0/1 \rightarrow$ candidat pour $(\hat{w}, \hat{b})_5$

■ Séparabilité linéaire

Définition

Un jeu de données $\{(x_i, y_i)\}_{i=1,\dots,n}$ est **linéairement séparable** s'il existe au moins un hyperplan dans \mathbb{R}^d tel que tous les points x_i positifs ($y_i = 1$) sont d'un côté de cet hyperplan et tous les points x_i négatifs ($y_i = -1$) de l'autre



Équation de $H : \langle w, x \rangle + b = 0$

$(w, b) \in \mathbb{R}^d \times \mathbb{R}$ avec $\vec{w} \perp H$

On choisit (w, b) tel que l'angle entre \vec{w} et tout point positif est inférieur à 90°

Pour tout nouveau point x , on estime son label y par sign $[\langle w, x \rangle + b]$

Un hyperplan séparateur H ne fait aucune erreur de classification sur les données d'apprentissage

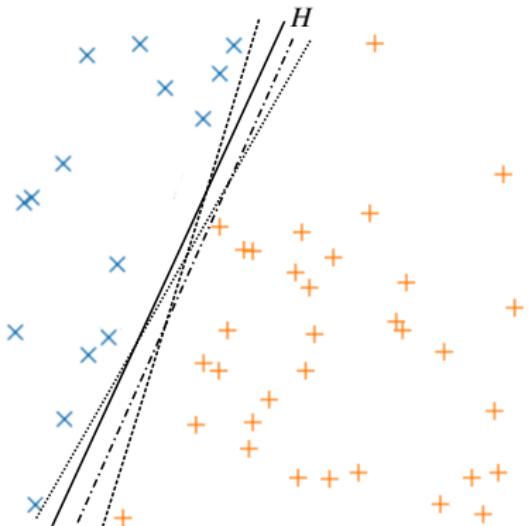
► (w, b) minimise le risque empirique pour le coût $0/1 \rightarrow$ candidat pour $(\hat{w}, \hat{b})_5$

■ Séparabilité linéaire



Définition

Un jeu de données $\{(x_i, y_i)\}_{i=1,\dots,n}$ est **linéairement séparable** s'il existe au moins un hyperplan dans \mathbb{R}^d tel que tous les points x_i positifs ($y_i = 1$) sont d'un côté de cet hyperplan et tous les points x_i négatifs ($y_i = -1$) de l'autre

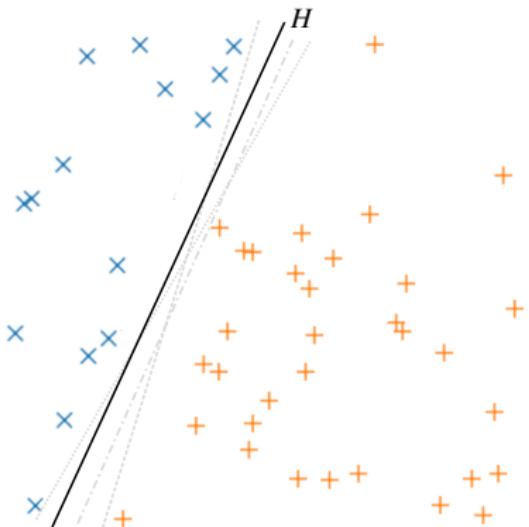


Il existe une infinité d'hyperplans séparateurs

■ Séparabilité linéaire

Définition

Un jeu de données $\{(x_i, y_i)\}_{i=1,\dots,n}$ est **linéairement séparable** s'il existe au moins un hyperplan dans \mathbb{R}^d tel que tous les points x_i positifs ($y_i = 1$) sont d'un côté de cet hyperplan et tous les points x_i négatifs ($y_i = -1$) de l'autre



Il existe une infinité d'hyperplans séparateurs

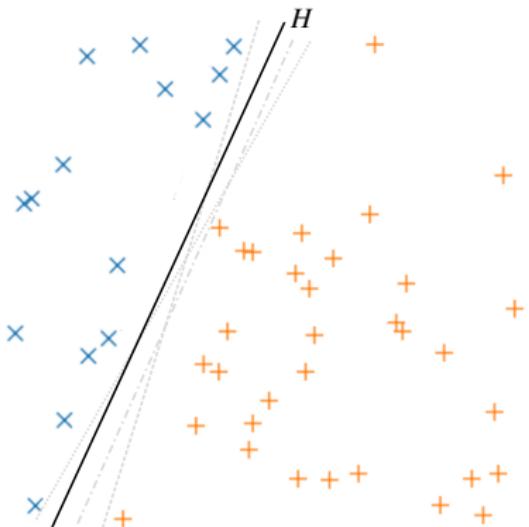


On cherche l'hyperplan dont la distance à l'observation la plus proche est la plus grande possible

■ Séparabilité linéaire

Définition

Un jeu de données $\{(x_i, y_i)\}_{i=1,\dots,n}$ est **linéairement séparable** s'il existe au moins un hyperplan dans \mathbb{R}^d tel que tous les points x_i positifs ($y_i = 1$) sont d'un côté de cet hyperplan et tous les points x_i négatifs ($y_i = -1$) de l'autre



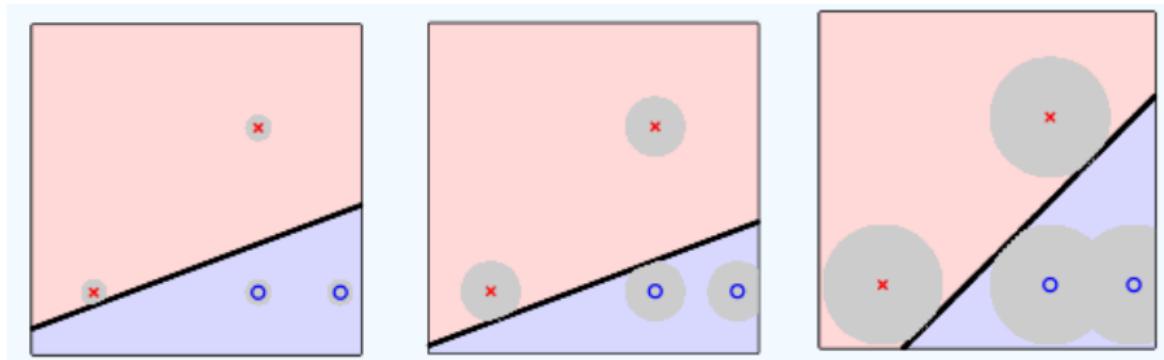
Il existe une infinité d'hyperplans séparateurs



On cherche l'hyperplan dont la distance à l'observation la plus proche est la plus grande possible

Cela revient à trouver l'hyperplan équidistant des observations positive et négative les plus proches

L'hyperplan équidistant des observations positive et négative les plus proches permet de ne pas être sensible à un petit changement dans les données (permet de mieux gérer les erreurs de mesure par exemple).



Tirée de cet ouvrage

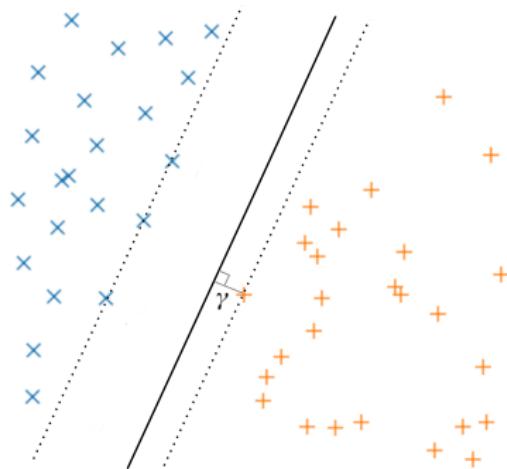
■ Marges et vecteurs de support



Définition

La marge γ d'un hyperplan séparateur est la distance de cet hyperplan à l'observation du jeu d'apprentissage la plus proche.

Objectif Chercher l'hyperplan séparateur qui est à une distance γ d'au moins une observation négative et une observation positive *i.e.* chercher l'hyperplan **qui maximise sa marge γ**



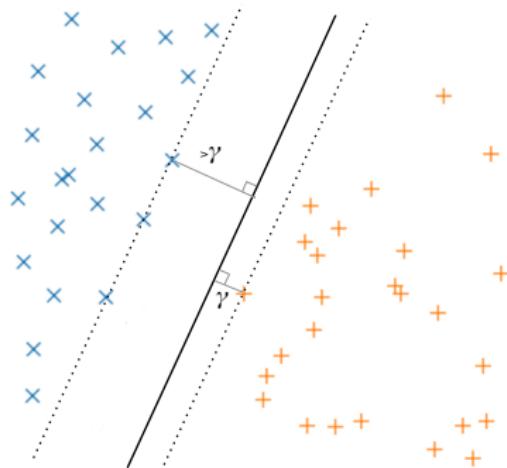
■ Marges et vecteurs de support



Définition

La marge γ d'un hyperplan séparateur est la distance de cet hyperplan à l'observation du jeu d'apprentissage la plus proche.

Objectif Chercher l'hyperplan séparateur qui est à une distance γ d'au moins une observation négative et une observation positive *i.e.* chercher l'hyperplan **qui maximise sa marge γ**



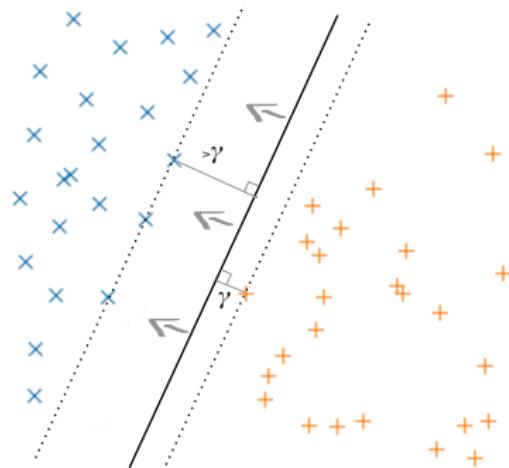
■ Marges et vecteurs de support



Définition

La marge γ d'un hyperplan séparateur est la distance de cet hyperplan à l'observation du jeu d'apprentissage la plus proche.

Objectif Chercher l'hyperplan séparateur qui est à une distance γ d'au moins une observation négative et une observation positive *i.e.* chercher l'hyperplan **qui maximise sa marge γ**



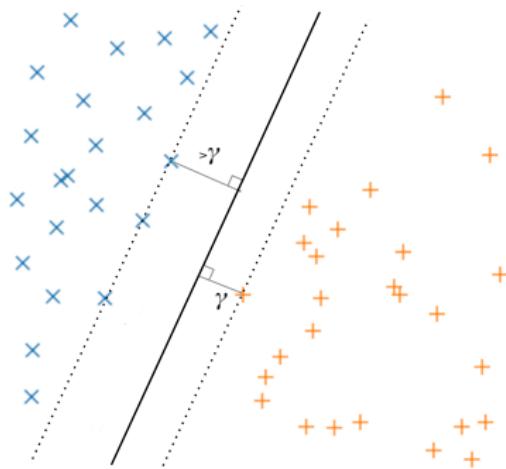
■ Marges et vecteurs de support



Définition

La marge γ d'un hyperplan séparateur est la distance de cet hyperplan à l'observation du jeu d'apprentissage la plus proche.

Objectif Chercher l'hyperplan séparateur qui est à une distance γ d'au moins une observation négative et une observation positive *i.e.* chercher l'hyperplan **qui maximise sa marge γ**



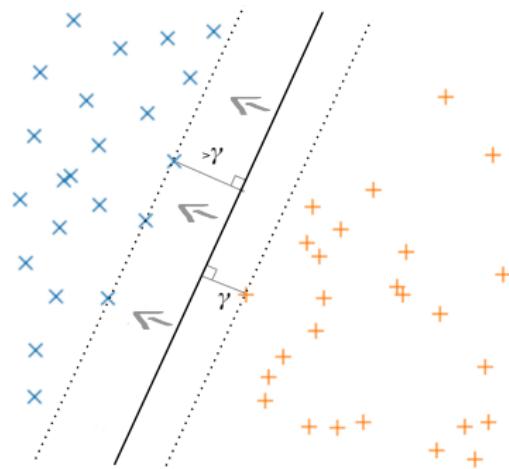
■ Marges et vecteurs de support



Définition

La marge γ d'un hyperplan séparateur est la distance de cet hyperplan à l'observation du jeu d'apprentissage la plus proche.

Objectif Chercher l'hyperplan séparateur qui est à une distance γ d'au moins une observation négative et une observation positive *i.e.* chercher l'hyperplan **qui maximise sa marge γ**



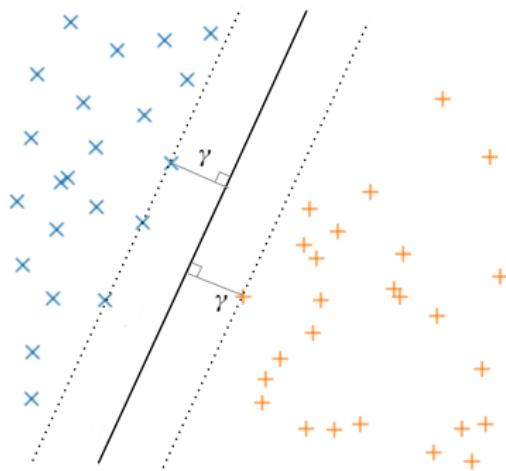
■ Marges et vecteurs de support



Définition

La marge γ d'un hyperplan séparateur est la distance de cet hyperplan à l'observation du jeu d'apprentissage la plus proche.

Objectif Chercher l'hyperplan séparateur qui est à une distance γ d'au moins une observation négative et une observation positive *i.e.* chercher l'hyperplan **qui maximise sa marge γ**



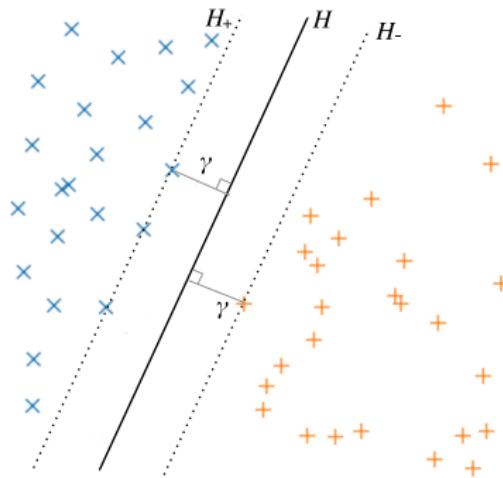
■ Marges et vecteurs de support



Définition

La marge γ d'un hyperplan séparateur est la distance de cet hyperplan à l'observation du jeu d'apprentissage la plus proche.

Objectif Chercher l'hyperplan séparateur qui est à une distance γ d'au moins une observation négative et une observation positive *i.e.* chercher l'hyperplan **qui maximise sa marge γ**



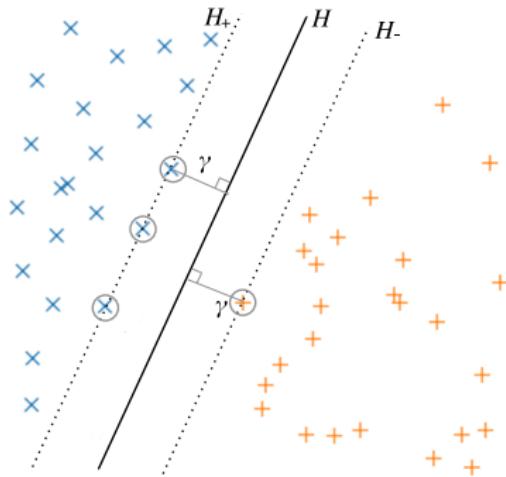
■ Marges et vecteurs de support



Définition

On appelle **vecteurs de support** les points x_i du jeu d'apprentissage situés à une distance γ de l'hyperplan séparateur H . Ils *soutiennent* les hyperplans H_+ et H_- situés à une distance γ de H

Un déplacement léger d'au moins un des vecteurs de support peut modifier l'hyperplan séparateur H qui maximise γ . À l'inverse, H ne change pas si l'on déplace légèrement une observation qui n'est pas vecteur de support

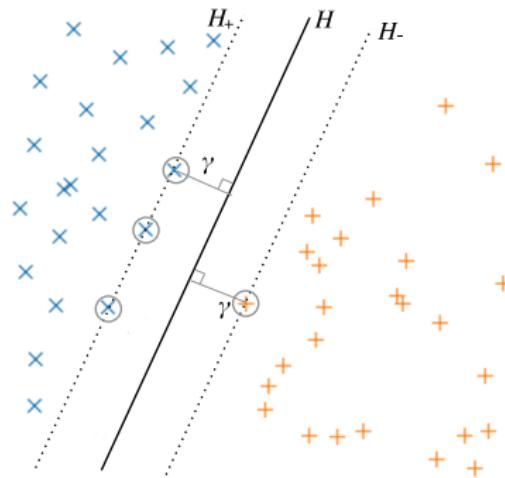


■ Marges et vecteurs de support

Définition

On appelle **vecteurs de support** les points x_i du jeu d'apprentissage situés à une distance γ de l'hyperplan séparateur H . Ils soutiennent les hyperplans H_+ et H_- situés à une distance γ de H .

Cependant, même si un des vecteurs de support est légèrement déplacé et que cela entraîne une modification de H , ce dernier sera toujours un hyperplan séparateur. La méthode est donc robuste à de légères perturbations des données d'apprentissage.

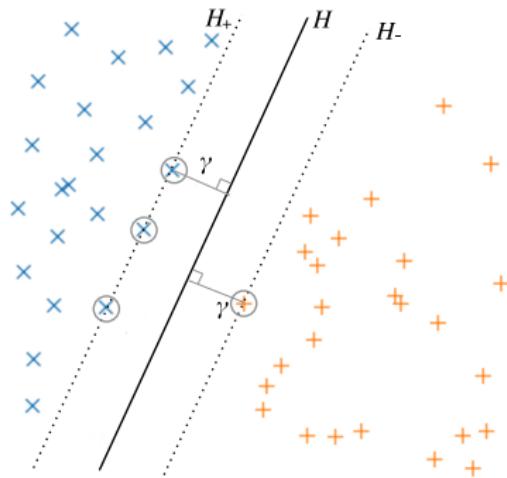


■ Marges et vecteurs de support

Définition

La zone située entre H_+ et H_- est appelée **zone d'indécision** ; elle ne contient aucune observation

Maximiser la marge permet donc de minimiser l'incertitude sur la classe à attribuer à une nouvelle observation x qui tombe près de l'hyperplan séparateur H



■ Formulation mathématique

Soit $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, 1\}$ des données d'apprentissage séparables

Question Comment trouver l'hyperplan séparateur H qui maximise la marge ?

Rappel Tout hyperplan $H \subset \mathbb{R}^d$ est donné par $H = \{x \in \mathbb{R}^d, \langle w, x \rangle + b = 0\}$ avec $(w, b) \in \mathbb{R}^d \times \mathbb{R}$

 On va reformuler le problème mathématiquement : trouver un hyperplan séparateur H qui maximise la marge revient à trouver les paramètres w et b adéquates, *i.e.* qui respectent certaines conditions.

Proposition

$\|\cdot\|_2$: dist. euclidienne

Soit un hyperplan $H \subset \mathbb{R}^d$ d'équation $\langle w, x \rangle + b = 0$ avec $(w, b) \in \mathbb{R}^d \times \mathbb{R}$. La distance entre un point $x \in \mathbb{R}^d$ et H , *i.e.* la distance entre x et son projeté orthogonal sur H est donnée par

$$d(x, H) = \frac{|\langle w, x \rangle + b|}{\|w\|_2}$$

Éléments de preuve : voir ce cours, page 293

■ Formulation primale

Proposition

SVM à marge rigide

Soit $(\hat{w}, \hat{b}) \in \mathbb{R}^d \times \mathbb{R}$ des paramètres qui vérifient

$$(\hat{w}, \hat{b}) = \arg \min_{(w,b) \in \mathbb{R}^d \times \mathbb{R}} \frac{1}{2} \|w\|_2^2 \quad \text{s.c.} \quad y_i (\langle w, x_i \rangle + b) \geq 1 \quad \forall i \in \{1, \dots, n\} \quad (1)$$

Alors $H = \{x \in \mathbb{R}^d, \langle \hat{w}, x \rangle + \hat{b} = 0\}$ est un hyperplan séparateur qui maximise la marge. En particulier, il existe $i, j \in \{1, \dots, n\}$ tels que $x_i \in H_+ = \{x \in \mathbb{R}^d, \langle \hat{w}, x \rangle + \hat{b} = 1\}$ et $x_j \in H_- = \{x \in \mathbb{R}^d, \langle \hat{w}, x \rangle + \hat{b} = -1\}$

- Le problème défini par (1) est un **problème d'optimisation convexe**
- Le problème défini par (1) est un **problème d'optimisation quadratique** : de nombreuses solutions ont été proposées pour résoudre ce type de problème
- Tout hyperplan d'équation $\langle w, x \rangle + b = 0$ est invariant à une multiplication près des coefficients w et b par $k \in \mathbb{R}^*$. Les solutions données par les équations (??) et (1) définissent donc le même hyperplan H

■ Formulation primale

Proposition

SVM à marge rigide

Soit $(\hat{w}, \hat{b}) \in \mathbb{R}^d \times \mathbb{R}$ des paramètres qui vérifient

$$(\hat{w}, \hat{b}) = \arg \min_{(w,b) \in \mathbb{R}^d \times \mathbb{R}} \frac{1}{2} \|w\|_2^2 \quad \text{s.c.} \quad y_i (\langle w, x_i \rangle + b) \geq 1 \quad \forall i \in \{1, \dots, n\} \quad (1)$$

Alors $H = \{x \in \mathbb{R}^d, \langle \hat{w}, x \rangle + \hat{b} = 0\}$ est un hyperplan séparateur qui maximise la marge. En particulier, il existe $i, j \in \{1, \dots, n\}$ tels que $x_i \in H_+ = \{x \in \mathbb{R}^d, \langle \hat{w}, x \rangle + \hat{b} = 1\}$ et $x_j \in H_- = \{x \in \mathbb{R}^d, \langle \hat{w}, x \rangle + \hat{b} = -1\}$

Question Quelle étiquette attribuer à un nouveau point $x \in \mathbb{R}^d$?

■ Formulation primale

Proposition

SVM à marge rigide

Soit $(\hat{w}, \hat{b}) \in \mathbb{R}^d \times \mathbb{R}$ des paramètres qui vérifient

$$(\hat{w}, \hat{b}) = \arg \min_{(w,b) \in \mathbb{R}^d \times \mathbb{R}} \frac{1}{2} \|w\|_2^2 \quad \text{s.c.} \quad y_i (\langle w, x_i \rangle + b) \geq 1 \quad \forall i \in \{1, \dots, n\} \quad (1)$$

Alors $H = \{x \in \mathbb{R}^d, \langle \hat{w}, x \rangle + \hat{b} = 0\}$ est un hyperplan séparateur qui maximise la marge. En particulier, il existe $i, j \in \{1, \dots, n\}$ tels que $x_i \in H_+ = \{x \in \mathbb{R}^d, \langle \hat{w}, x \rangle + \hat{b} = 1\}$ et $x_j \in H_- = \{x \in \mathbb{R}^d, \langle \hat{w}, x \rangle + \hat{b} = -1\}$

Question Quelle étiquette attribuer à un nouveau point $x \in \mathbb{R}^d$?

Classifieur SVM à marge rigide :

$$\text{sign} \left[\langle \hat{w}, x \rangle + \hat{b} \right]$$

avec (\hat{w}, \hat{b}) la solution donnée par l'équation (??) ou l'équation (1)

■ Formulation duale

Proposition

SVM à marge rigide

Soit $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_n) \in \mathbb{R}^n$ qui vérifie

$$\begin{aligned}\hat{\alpha} &= \arg \max_{\alpha=(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{\ell=\ell}^n \alpha_i \alpha_\ell y_i y_\ell \langle x_i, x_\ell \rangle \\ &\text{s.c. } \sum_{i=1}^n \alpha_i y_i = 0 \text{ et } \alpha_i \geq 0 \quad \forall i \in \{1, \dots, n\}\end{aligned}\tag{2}$$

Alors le couple (\hat{w}, \hat{b}) défini par

$$\hat{w} = \sum_{i=1}^n \hat{\alpha}_i y_i x_i \quad \text{et} \quad \hat{b} = 1 - \min_{\substack{i \in \{1, \dots, n\} \\ y_i=1}} \langle \hat{w}, x_i \rangle$$

est identique à la solution donnée par l'équation (1)

Preuve : section 10.1.3 d'*Introduction au Machine Learning* de Chloé-Agathe Azencott

Classifieur SVM à marge rigide $\text{sign} \left[\sum_{i=1}^n \hat{\alpha}_i y_i \langle x_i, x \rangle + \hat{b} \right]$

■ Formulation duale



Le problème défini par (2) est un **problème d'optimisation convexe et quadratique**. Il est appelé *problème d'optimisation dual* et est obtenu à partir du *problème d'optimisation primal* défini par (1) de la manière suivante :

Le problème primal (P) induit par (1) peut se réécrire à l'aide de son Lagrangien \mathcal{L} :

$$(P) : \min_{(w,b) \in \mathbb{R}^d \times \mathbb{R}} \max_{\substack{\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n \\ \alpha_1 \geq 0, \dots, \alpha_n \geq 0}} \mathcal{L}(w, b, \alpha)$$

$$\text{où } \mathcal{L}(w, \alpha) = \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^n \alpha_i (1 - y_i \langle w, x_i \rangle)$$

En inversant le min et le max dans (P) on obtient le problème dual (Q) suivant :

$$(Q) : \max_{\substack{\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n \\ \alpha_1 \geq 0, \dots, \alpha_n \geq 0}} \min_{(w,b) \in \mathbb{R}^d \times \mathbb{R}} \mathcal{L}(w, b, \alpha)$$

De manière générale, n'importe quelles solutions p^* de (P) et q^* de (Q) vérifient toujours $q^* \leq p^*$ (dualité faible) mais puisque (P) est un problème d'optimisation convexe avec des contraintes affines alors on a $p^* = q^*$ (condition de Slater), i.e., le saut de dualité $p^* - q^*$ est nulle

■ Formulation duale



- **Interprétation géométrique** : les solutions $\hat{\alpha}$ et (\hat{w}, \hat{b}) vérifient les *conditions d'optimalité de Karush-Kuhn-Tucker* dont la *condition d'écart complémentaire* qui stipule que, pour tout $i \in \llbracket 1, n \rrbracket$, $\hat{\alpha}_i \left[y_i \left(\langle \hat{w}, x_i \rangle + \hat{b} \right) - 1 \right] = 0$. Il y a alors deux possibilités, lesquelles ?
- **Complexité algorithmique** : la formulation primaire est un problème d'optimisation en $d + 1$ dimensions tandis que la formulation duale est un problème en n dimensions. Avec peu de données et beaucoup de variables explicatives, on préférera la formulation duale ; dans le cas inverse, on préférera résoudre le problème primal.

■ Formulation duale



- **Interprétation géométrique** : les solutions $\hat{\alpha}$ et (\hat{w}, \hat{b}) vérifient les *conditions d'optimalité de Karush-Kuhn-Tucker* dont la *condition d'écart complémentaire* qui stipule que, pour tout $i \in \llbracket 1, n \rrbracket$, $\hat{\alpha}_i \left[y_i \left(\langle \hat{w}, x_i \rangle + \hat{b} \right) - 1 \right] = 0$. Il y a alors deux possibilités, lesquelles ?

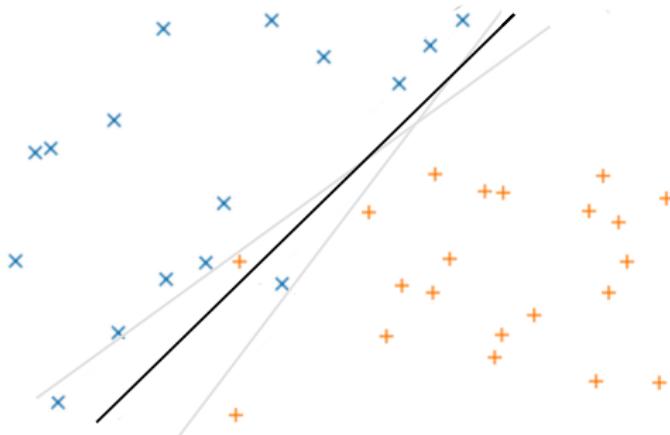
- Si x_i est à l'*extérieur* des hyperplans H_+ et H_- , i.e. $y_i \left(\langle \hat{w}, x_i \rangle + \hat{b} \right) > 1$ alors $\hat{\alpha}_i = 0$
- Si $\hat{\alpha}_i > 0$ alors $y_i \left(\langle \hat{w}, x_i \rangle + \hat{b} \right) = 1$, i.e. $x_i \in H_+$ ou $x_i \in H_-$; x_i est donc un vecteur de support

La formulation duale permet d'identifier des vecteurs de support (pas forcément tous). Ce sont seulement ces vecteurs de support qui sont utilisés pour construire le classifieur à marge rigide

- **Complexité algorithmique** : la formulation primaire est un problème d'optimisation en $d + 1$ dimensions tandis que la formulation duale est un problème en n dimensions. Avec peu de données et beaucoup de variables explicatives, on préférera la formulation duale ; dans le cas inverse, on préférera résoudre le problème primal.

LE CAS LINÉAIREMENT NON SÉPARABLE : SVM À MARGE SOUPLE

■ Données non séparables linéairement

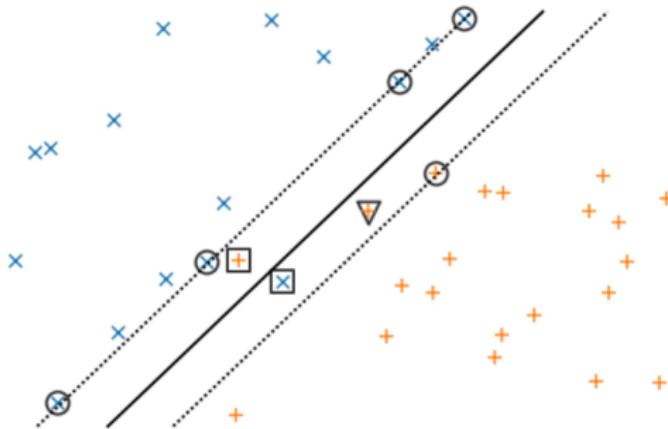


Tirée de l'ouvrage *Introduction au Machine Learning* de Chloé-Agathe Azencott



Données non séparables linéairement, mais seulement à quelques observations près (bruit ?)

■ Données non séparables linéairement



Tirée de l'ouvrage *Introduction au Machine Learning* de Chloé-Agathe Azencott



Données non séparables linéairement, mais seulement à quelques observations près (bruit ?)



On va quand même séparer linéairement les données en maximisant la marge mais en autorisant maintenant quelques erreurs de classification

■ Variable d'ajustement ξ_i

Contrainte de séparabilité pour les SVM à marge rigide de l'équation (1), i.e. tous les points du jeu d'entraînement sont bien classés :

$$\forall i \in \{1, \dots, n\} \quad y_i (\langle w, x_i \rangle + b) \geq 1 \quad (2)$$

Contrainte de séparabilité relachée pour SVM à marge souple, i.e. un point du jeu d'entraînement peut maintenant être dans la zone d'indécision :

$$\forall i \in \{1, \dots, n\} \quad y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i$$

où $\xi_i \in \mathbb{R}^*$, appelée *variable d'ajustement* (*slack variable* en anglais), mesure à quel point la contrainte (2) a été enfreinte, i.e. à quel point (x_i, y_i) échoue à être bien séparé. Trois cas, lequels ?

■ Variable d'ajustement ξ_i

Contrainte de séparabilité pour les SVM à marge rigide de l'équation (1), i.e. tous les points du jeu d'entraînement sont bien classés :

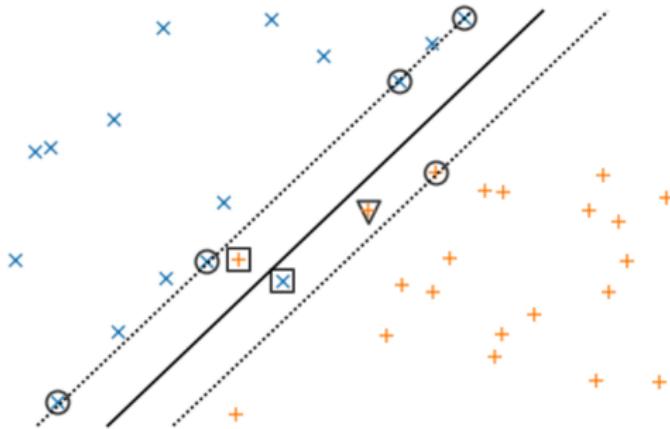
$$\forall i \in \{1, \dots, n\} \quad y_i (\langle w, x_i \rangle + b) \geq 1 \quad (2)$$

Contrainte de séparabilité relachée pour SVM à marge souple, i.e. un point du jeu d'entraînement peut maintenant être dans la zone d'indécision :

$$\forall i \in \{1, \dots, n\} \quad y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i$$

où $\xi_i \in \mathbb{R}^*$, appelée *variable d'ajustement* (*slack variable* en anglais), mesure à quel point la contrainte (2) a été enfreinte, i.e. à quel point (x_i, y_i) échoue à être bien séparé. Trois cas, lequels ?

- $\xi_i = 0$ $y_i (\langle w, x_i \rangle + b) \geq 1$, i.e. (x_i, y_i) est bien classé et est en dehors de la zone d'indécision
- $0 < \xi_i \leq 1$ $0 \leq y_i (\langle w, x_i \rangle + b) < 1$, i.e. (x_i, y_i) est bien classé mais est dans la zone d'indécision
- $\xi_i > 1$ $y_i (\langle w, x_i \rangle + b) < 0$, i.e. (x_i, y_i) est mal classé (et peut être dans la zone d'indécision ou non)



- On souhaite minimiser le nombre d'*outliers*, i.e. les observations qui sont dans la zone d'indécision ou qui sont mal classées pour faire le moins d'erreurs de classification possible. Cela revient à minimiser $\sum_{i=1}^n \xi_i$
- On souhaite toujours minimiser $\frac{1}{2} \|w\|_2^2$ pour maximiser la marge



Lorsque $\sum_{i=1}^n \xi_i$ diminue, la marge diminue et inversement. Il faut donc faire un compromis entre maximiser la marge et minimiser le nombre d'outliers

■ Formulation primale



Problème d'optimisation primale

SVM à marge souple

On cherche $(\hat{w}, \hat{b}, \hat{\xi}) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+$ tels que

$$\begin{aligned} (\hat{w}, \hat{b}, \hat{\xi}) &= \arg \min_{(w, b, \xi) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+} \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^n \xi_i \\ \text{s.c. } &y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i \text{ et } \xi_i \geq 0 \quad \forall i \in \{1, \dots, n\} \end{aligned} \tag{3}$$

Le paramètre $C \in \mathbb{R}_+$ permet d'indiquer une préférence entre maximiser la marge et minimiser le nombre d'outliers :

- Si C est large alors on préfère minimiser le nombre d'outliers
- Si C est petit, on accorde moins d'importance au fait qu'il y ait des outliers

Le paramètre C est un hyperparamètre que l'utilisateur doit choisir. Ce choix peut être guidé en comparant, pour différentes valeurs de C , les performances des modèles sur les données de validation

■ Formulation primale

Problème d'optimisation primale

SVM à marge souple

On cherche $(\hat{w}, \hat{b}, \hat{\xi}) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+$ tels que

$$\begin{aligned} (\hat{w}, \hat{b}, \hat{\xi}) &= \arg \min_{(w, b, \xi) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+} \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^n \xi_i \\ \text{s.c. } &y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i \text{ et } \xi_i \geq 0 \quad \forall i \in \{1, \dots, n\} \end{aligned} \tag{3}$$

Quand $C > 0$, le problème de minimisation (P) induit par (3) peut se réécrire

$$(P) : \min_{(w, b) \in \mathbb{R}^d \times \mathbb{R}} \frac{1}{2nC} \|w\|_2^2 + \frac{1}{n} \sum_{i=1}^n \ell^{\text{hinge}}((w, b), (x_i, y_i))$$

où $\ell^{\text{hinge}}((w, b), (x_i, y_i)) = \max(0, 1 - y_i (\langle w, x_i \rangle + b))$ est la fonction de coût *hinge*

Le problème (P) peut donc être vu comme un problème de minimisation du risque empirique régularisé. La fonction de régularisation $w \mapsto \frac{1}{2nC} \|w\|_2^2$ permet d'éviter le sur-apprentissage

■ Formulation primale



Ici les données sont séparables et un algorithme de SVM à marge rigide donnerait l'hyperplan séparateur en rouge. Pourtant, cette frontière de décision risque de ne pas bien fonctionner avec de nouvelles données. Même dans le cas de données séparables, il peut être préférable de recourir à un algorithme de SVM à marge souple pour obtenir l'hyperplan vert afin d'éviter l'écueil du sur-apprentissage

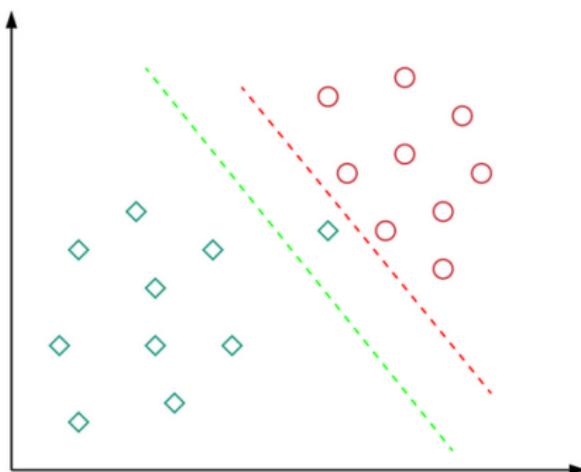


Image tirée de l'article *Support Vector Machines – Soft Margin Formulation and Kernel Trick* publié dans Towards Data Science

■ Formulation primaile

Problème d'optimisation primaile

SVM à marge souple

On cherche $(\hat{w}, \hat{b}, \hat{\xi}) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+$ tels que

$$\begin{aligned} (\hat{w}, \hat{b}, \hat{\xi}) &= \arg \min_{(w, b, \xi) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+} \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^n \xi_i \\ \text{s.c. } y_i (\langle w, x_i \rangle + b) &\geq 1 - \xi_i \text{ et } \xi_i \geq 0 \quad \forall i \in \{1, \dots, n\} \end{aligned} \tag{3}$$

Classifieur SVM à marge souple :

$$\text{sign} \left[\langle \hat{w}, x \rangle + \hat{b} \right]$$

avec (\hat{w}, \hat{b}) donnés par l'équation (3)

- Le problème d'optimisation (P) induit par (3) est un **problème d'optimisation convexe et quadratique**. À partir de ce problème d'optimisation primal, on peut définir un problème d'optimisation dual équivalent

■ Formulation duale

Proposition

SVM à marge souple

Soit $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_n) \in \mathbb{R}^n$ qui vérifie

$$\begin{aligned}\hat{\alpha} &= \arg \max_{\alpha=(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{\ell=1}^n \alpha_i \alpha_\ell y_i y_\ell \langle x_i, x_\ell \rangle \\ &\text{s.c. } \sum_{i=1}^n \alpha_i y_i = 0 \text{ et } 0 \leq \alpha_i \leq C \quad \forall i \in \{1, \dots, n\}\end{aligned}\tag{4}$$

Alors le couple (\hat{w}, \hat{b}) défini par

$$\hat{w} = \sum_{i=1}^n \hat{\alpha}_i y_i x_i \quad \text{et} \quad \hat{b} = y_i - \sum_{j=1}^n \hat{\alpha}_j y_j \langle x_j, x_i \rangle \quad \text{pour } (x_i, y_i) \text{ t.q. } 0 < \hat{\alpha}_i < C$$

est identique à la solution donnée par l'équation (3)

Preuve : section 10.2.2 d'*Introduction au Machine Learning* de Chloé-Agathe Azencott

Il est possible (mais rare) qu'il n'existe pas $i \in \{1, \dots, n\}$ tel que $0 < \hat{\alpha}_i < C$, dans ce cas, il est possible de prendre une valeur au hasard dans un certain intervalle pour \hat{b} . Il est cependant plutôt conseillé de changer C pour avoir $0 < \hat{\alpha}_i < C$

■ Formulation duale

Proposition

SVM à marge souple

Soit $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_n) \in \mathbb{R}^n$ qui vérifie

$$\begin{aligned}\hat{\alpha} &= \arg \max_{\alpha=(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{\ell=1}^n \alpha_i \alpha_\ell y_i y_\ell \langle x_i, x_\ell \rangle \\ &\text{s.c. } \sum_{i=1}^n \alpha_i y_i = 0 \text{ et } 0 \leq \alpha_i \leq C \quad \forall i \in \{1, \dots, n\}\end{aligned}\tag{4}$$

Alors le couple (\hat{w}, \hat{b}) défini par

$$\hat{w} = \sum_{i=1}^n \hat{\alpha}_i y_i x_i \quad \text{et} \quad \hat{b} = y_i - \sum_{j=1}^n \hat{\alpha}_j y_j \langle x_j, x_i \rangle \quad \text{pour } (x_i, y_i) \text{ t.q. } 0 < \hat{\alpha}_i < C$$

est identique à la solution donnée par l'équation (3)

Preuve : section 10.2.2 d'*Introduction au Machine Learning* de Chloé-Agathe Azencott

Classifieur SVM à marge souple $\text{sign} \left[\sum_{i=1}^n \alpha_i y_i \langle x_i, x \rangle + \hat{b} \right]$

■ Formulation duale

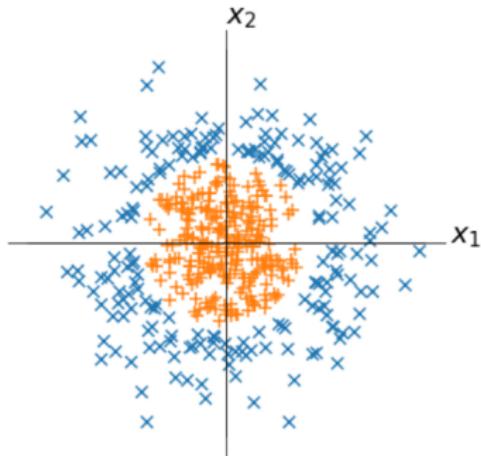


- Le problème défini par (4) est un **problème d'optimisation convexe et quadratique**.
- **Interprétation géométrique** : les solutions $\hat{\alpha}$ et (\hat{w}, \hat{b}) vérifient les *conditions d'optimalité de Karush-Kuhn-Tucker* dont la *condition d'écart complémentaire* qui stipule que, pour tout $i \in \llbracket 1, n \rrbracket$, $\hat{\alpha}_i \left[y_i \left(\langle \hat{w}, x_i \rangle + \hat{b} \right) - 1 + \xi_i \right] = 0$.
 - Si x_i est *au dessus* de H_+ pour $y_i = 1$ et en dessous de H_- pour $y_i = -1$, i.e. $y_i \left(\langle \hat{w}, x_i \rangle + \hat{b} \right) > 1$ alors $\xi_i = 0$ et donc $\hat{\alpha}_i = 0$
 - Si $\hat{\alpha}_i > 0$ alors
 - Si $\xi_i = 0$ alors $y_i \left(\langle \hat{w}, x_i \rangle + \hat{b} \right) = 1$, i.e. $x_i \in H_+$ si $y_i = 1$ et $x_i \in H_-$ si $y_i = -1$; x_i est donc un vecteur de support
 - Si $\xi_i \neq 0$ alors x_i est un outlier.

La formulation duale permet d'identifier des vecteurs de support (pas forcément tous) et des outliers (pas forcément tous). Ce sont seulement ces vecteurs de support et outliers qui sont utilisés pour construire le classifieur à marge souple

LE CAS LINÉAIREMENT NON SÉPARABLE : SVM À NOYAU

■ Problématique

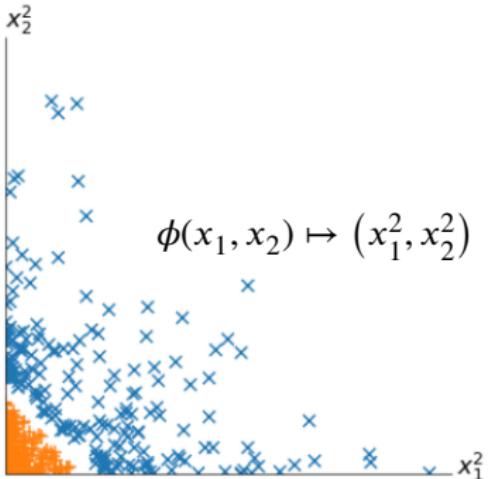
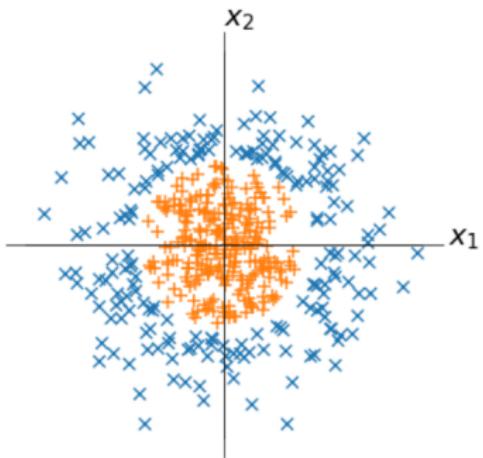


Tirée de l'ouvrage *Introduction au Machine Learning* de Chloé-Agathe Azencott



Les données ne sont pas séparables linéairement (même à quelques observations près)

■ Problématique



Tirée de l'ouvrage *Introduction au Machine Learning* de Chloé-Agathe Azencott



Les données ne sont pas séparables linéairement (même à quelques observations près)



On va transformer les données par une application $\phi : \mathcal{X} \rightarrow \mathcal{H}$ afin de les séparer linéairement dans l'espace de redescription \mathcal{H} en utilisant une SVM à marge souple

■ Formulation duale

Problème duale

SVM à marge souple sur données transformées

On cherche $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_n) \in \mathbb{R}^n$ qui vérifie

$$\begin{aligned}\hat{\alpha} &= \arg \max_{\alpha=(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{\ell=1}^n \alpha_i \alpha_{\ell} y_i y_{\ell} \langle \phi(x_i), \phi(x_{\ell}) \rangle \\ &\text{s.c. } \sum_{i=1}^n \alpha_i y_i = 0 \text{ et } 0 \leq \alpha_i \leq C \quad \forall i \in \{1, \dots, n\}\end{aligned}\tag{4}$$

avec \hat{w} défini comme précédemment (voir problème dual, SVM marge souple) et \hat{b} défini par

$$\hat{b} = y_i - \sum_{j=1}^n \hat{\alpha}_j y_j \langle \phi(x_j), \phi(x_i) \rangle \quad \text{pour pour } i \in \{1, \dots, n\} \text{ t.q. } 0 < \hat{\alpha}_i < C$$

Classifieur SVM à marge souple $f(x) = \text{sign} \left[\sum_{i=1}^n \alpha_i y_i \langle \phi(x_i), \phi(x) \rangle + \hat{b} \right]$

■ L'astuce du noyau

L'application $\phi : \mathcal{X} \rightarrow \mathcal{H}$ intervient dans le problème d'optimisation induit par (4) et dans l'expression du classifieur à marge souple uniquement via les produits scalaires

$$\langle \phi(x_j), \phi(x_i) \rangle \quad i, j \in \{1, \dots, n\}$$

 La dimension de \mathcal{H} peut être grande et donc le calcul de $\langle \phi(x_j), \phi(x_i) \rangle$ coûteux

 Essayer de calculer $\langle \phi(x_j), \phi(x_i) \rangle$ de manière plus efficace

Ex. $\phi : x = (x_1, \dots, x_d) \in \mathbb{R}^d \mapsto (1, x_1, \dots, x_d, x_1 x_1, x_1 x_2, \dots, x_d x_d) \in \mathbb{R}^{1+d+d^2}$

$$\begin{aligned}\langle \phi(x), \phi(\tilde{x}) \rangle &= 1 + \sum_{i=1}^d x_i \tilde{x}_i + \sum_{i=1}^d \sum_{j=1}^d x_i \tilde{x}_i x_j \tilde{x}_j \quad \text{complexité algo. } O(1 + d + d^2) \\ &= 1 + \langle x, \tilde{x} \rangle + \langle x, \tilde{x} \rangle^2 \quad \text{complexité algo. } O(d) \\ &= K(x, \tilde{x})\end{aligned}$$

■ L'astuce du noyau

Soit la fonction K , appelée noyau, définie par

$$K : (x, \tilde{x}) \in \mathcal{X} \times \mathcal{X} \mapsto \langle \phi(x), \phi(\tilde{x}) \rangle \in \mathbb{R}$$

Astuce du noyau

$$i, j \in \{1, \dots, n\}$$

Au lieu de calculer $\phi(x_i)$ et $\phi(x_j)$ puis de calculer $\langle \phi(x_i), \phi(x_j) \rangle$, on calcule directement $K(x_i, x_j)$

Remarque L'astuce du noyau intervient dans le cadre du problème d'optimisation duale

- Il est souvent plus efficace en terme de temps de calcul de calculer $K(x_i, x_j)$ plutôt que $\phi(x_i), \phi(x_j)$ puis $\langle \phi(x_i), \phi(x_j) \rangle$
- Il n'y a pas besoin de connaître ϕ explicitement, il nous suffit juste de connaître le noyau K . Aussi, K peut être choisi de manière arbitraire (hyperparamètre), du moment que l'existence de ϕ est garantie.

Question Soit $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. Qu'est-ce qui garantit qu'il existe $\phi : \mathcal{X} \rightarrow \mathcal{H}$ telle que $K(x, \tilde{x}) = \langle \phi(x), \phi(\tilde{x}) \rangle$?

■ Propriétés du noyau K

Définition

Soit une fonction $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$

- K est dite **symétrique** si $K(x, \tilde{x}) = K(\tilde{x}, x)$ pour tout $x, \tilde{x} \in \mathcal{X}$
- K est dite **semi-définie positive** si

$$\forall m \in \mathbb{N}, \forall x_1, \dots, x_m \in \mathcal{X}, \forall a_1, \dots, a_m \in \mathbb{R} \quad \sum_{i=1}^m \sum_{j=i}^m a_i a_j K(x_i, x_j) \geq 0$$

Théorème

Moore-Aronszajn

Soit $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ une fonction symétrique et semi-définie positive alors il existe un espace (de Hilbert) \mathcal{H} et une application $\phi : \mathcal{X} \rightarrow \mathcal{H}$ telle que

$$\forall x, \tilde{x} \in \mathcal{X} \quad K(x, \tilde{x}) = \langle \phi(x), \phi(\tilde{x}) \rangle$$



On choisit donc le noyau K tel qu'il soit symétrique et semi-défini positif

Exercice Écrire le noyau $K : (x, \tilde{x}) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto (\langle x, \tilde{x} \rangle + c)^2, c \in \mathbb{R}_+$ sous forme d'un produit scalaire $\langle \phi(x), \phi(\tilde{x}) \rangle$

■ Propriétés du noyau K

Définition

Soit une fonction $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$

- K est dite **symétrique** si $K(x, \tilde{x}) = K(\tilde{x}, x)$ pour tout $x, \tilde{x} \in \mathcal{X}$
- K est dite **semi-définie positive** si

$$\forall m \in \mathbb{N}, \forall x_1, \dots, x_m \in \mathcal{X}, \forall a_1, \dots, a_m \in \mathbb{R} \quad \sum_{i=1}^m \sum_{j=i}^m a_i a_j K(x_i, x_j) \geq 0$$

Théorème

Moore-Aronszajn

Soit $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ une fonction symétrique et semi-définie positive alors il existe un espace (de Hilbert) \mathcal{H} et une application $\phi : \mathcal{X} \rightarrow \mathcal{H}$ telle que

$$\forall x, \tilde{x} \in \mathcal{X} \quad K(x, \tilde{x}) = \langle \phi(x), \phi(\tilde{x}) \rangle$$



On choisit donc le noyau K tel qu'il soit symétrique et semi-défini positif

Remarque Soit un noyau K et des observations x_1, \dots, x_n . La matrice $M \in \mathbb{R}^{n \times n}$ définie par $M_{i,j} = K(x_i, x_j)$ est appelée *matrice de Gram*

■ Exemples de noyau



Soit $x, \tilde{x} \in \mathbb{R}^d \times \mathbb{R}^d$

Noyau linéaire $K(x, \tilde{x}) = \langle x, \tilde{x} \rangle$

Noyau cosinus $K(x, \tilde{x}) = \frac{\langle x, \tilde{x} \rangle}{\|x\|_2 \|\tilde{x}\|_2}$

Noyau quadratique $K(x, \tilde{x}) = (\langle x, \tilde{x} \rangle + c)^2 \quad c \in \mathbb{R}_+$

Noyau polynomial $K(x, \tilde{x}) = (\langle x, \tilde{x} \rangle + c)^d \quad c \in \mathbb{R}_+, d \in \mathbb{N}$

Noyau gaussien $K(x, \tilde{x}) = \exp \left\{ -\frac{1}{2} (x - \tilde{x})^T \Sigma (x - \tilde{x}) \right\}$
 $\Sigma \in \mathbb{R}^{d \times d}$ matrice semi-définie positive

Noyau gaussien $K(x, \tilde{x}) = \exp \left\{ -\frac{\|x - \tilde{x}\|_2^2}{2\sigma^2} \right\} \quad \sigma \in \mathbb{R}_+^*$

■ Formulation duale avec le noyau

Problème duale

SVM à marge souple et à noyau

On cherche $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_n) \in \mathbb{R}^n$ qui vérifie

$$\begin{aligned}\hat{\alpha} &= \arg \max_{\alpha=(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{\ell=1}^n \alpha_i \alpha_\ell y_i y_\ell K(\mathbf{x}_i, \mathbf{x}_\ell) \\ &\text{s.c. } \sum_{i=1}^n \alpha_i y_i = 0 \text{ et } 0 \leq \alpha_i \leq C \quad \forall i \in \{1, \dots, n\}\end{aligned}\tag{5}$$

avec \hat{w} défini comme précédemment (voir problème dual, SVM marge souple) et \hat{b} défini par

$$\hat{b} = y_i - \sum_{j=1}^n \hat{\alpha}_j y_j K(\mathbf{x}_j, \mathbf{x}_i) \quad \text{pour } i \in \{1, \dots, n\} \text{ t.q. } 0 < \hat{\alpha}_i < C$$

Le paramètre C et le noyau K sont des hyperparamètres que l'utilisateur doit choisir. Ce choix peut être guidé en comparant, pour différents noyaux K et constantes C , les performances des modèles sur les données de validation

■ Formulation duale avec le noyau

Problème duale

SVM à marge souple et à noyau

On cherche $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_n) \in \mathbb{R}^n$ qui vérifie

$$\begin{aligned}\hat{\alpha} &= \arg \max_{\alpha=(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{\ell=1}^n \alpha_i \alpha_\ell y_i y_\ell K(\mathbf{x}_i, \mathbf{x}_\ell) \\ &\text{s.c. } \sum_{i=1}^n \alpha_i y_i = 0 \text{ et } 0 \leq \alpha_i \leq C \quad \forall i \in \{1, \dots, n\}\end{aligned}\tag{5}$$

avec \hat{w} défini comme précédemment (voir problème dual, SVM marge souple) et \hat{b} défini par

$$\hat{b} = y_i - \sum_{j=1}^n \hat{\alpha}_j y_j K(\mathbf{x}_j, \mathbf{x}_i) \quad \text{pour } i \in \{1, \dots, n\} \text{ t.q. } 0 < \hat{\alpha}_i < C$$

Classifieur SVM à marge souple $f(x) = \text{sign} \left[\sum_{i=1}^n \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + \hat{b} \right]$

■ Références



-  C.A. Azencot *Introduction au Machine Learning*, *Dunod*,
ISBN : 978-2-10-084143-1
-  S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, *Cambridge University Press*, DOI : [10.1017/CBO9781107298019](https://doi.org/10.1017/CBO9781107298019), 2014
-  T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*,
Springer New York, DOI : [10.1007/978-0-387-84858-7](https://doi.org/10.1007/978-0-387-84858-7), 2009