HE VD HAUTE ÉCOLE D'INGÉNIERIE ET DE GESTION DU CANTON DE VAUD
IG

ISCD : **Apprentissage supervisé (APV)**	**TP 8**
S. Robert, F. Hêche	Décembre 2023

Introduction

Le but de ce travail pratique est de construire un algorithme de classification appelé Support Vector Machine (SVM). Dans la partie théorique, nous allons introduire la théorie nécessaire afin de comprendre cet algorithme. Ensuite dans la partie pratique, nous allons implémenter cet algorithme.

Le contexte est le suivant : nous possédons un dataset $\mathcal{D} := \{(x_1, y_1), \dots (x_N, y_N)\}$ avec $x_i \in \mathbb{R}^n$, les features et $y_i \in \{-1, 1\}$ la classe. Notons encore $C_1 := \{(x_i, y_i) \in \mathcal{D} \mid y_i = 1\}$ ainsi que $C_2 := \{(x_i, y_i) \in \mathcal{D} \mid y_i = -1\}$.

L'idée de base de SVM est la suivante : trouver un plan $w^T x + b$ qui sépare au mieux les deux classes C_1, C_2 et ensuite utiliser ce plan pour la classification en assignant chaque côté du plan à une des deux classes. Mathématiquement, notre classificateur sera de la forme h : $\mathbb{R}^n \to \{-1, 1\}$ définit par $h(x) = \text{sign}(w^T x + b)$. Evidemment, il existe potentiellement une infinité de plan qui sépare ces deux classes. Dans SVM, il a été choisi de considérer le plan qui maximise les marges.

Bien que cette idée soit très naturelle, il n'est pas évident de savoir comment déterminer ce plan $w^T x + b$. Le but de la partie théorique de travail est de comprendre comment ce plan est calculé en pratique.

Rappels

- Soient $w \in \mathbb{R}^n$, $b \in \mathbb{R}$. Un plan π de $\mathbb{R}^n$ est l'ensemble des points de $\mathbb{R}^n$ qui vérifient $w^T x + b = 0$. Par abus de notation, on note souvent cet ensemble $w^T x + b$.
- Deux ensemble de points X_1, X_2 sont (linéairement) séparables s'il existe un plan $w^T x + b$ avec $w^T x_i + b \geq 0$ pour tout $x_i \in X_0$ et $w^T x_j + b \leq 0$ pour tout $x_j \in X_1$.
- Soient X_0, X_1 deux ensembles de points séparables et $w^T x + b$ un plan séparant X_0, X_1. On appelle alors vecteurs support les points de X_0 ou X_1 les plus proches du plan $w^T x + b$. Et on appelle la distance entre les vecteur support et le plan la marge (margin).
- Soit $w^T x + b$ un plan et $x_i \in \mathbb{R}^n$. Alors la distance entre ce point et le plan est donné par $\frac{|w^T x_i + b|}{\|w\|}$

Naïve SVM

Dans le raisonnement qui suit, nous allons supposer que les classes C_1 et C_2 sont séparables.

1. Remarquer que le plan : $w^T x + b$ sépare les classes C_1, et C_2 si et seulement si

$$y_i(w^T x_i + b) \geq 0 \quad 1 \leq i \leq N \tag{1}$$

   Nous allons montrer qu'il existe un plan $w^T x + b$ qui sépare C_1 et C_2 et tel que

$$w^T x^{(1)} + b = 1 \tag{2}$$

$$w^T x^{(-1)} + b = -1 \tag{3}$$

où $x^{(1)}$, $x^{(-1)}$ sont les vecteurs support de C_1 et C_2 respectivement. Pour simplifier, nous allons supposer qu'il n'existe qu'un seul vecteur support par classe.

Pour commencer, partons d'un plan $w^T x + b$ qui sépare C_1, C_2 (un tel plan existe car on a supposé que C_1, C_2 étaient séparables).

Ensuite, translatons ce plan pour le placer à équidistance entre $x^{(1)}$ et $x^{(-1)}$. Pour faire cela, il suffit de considérer le plan π_1 d'équation $w^T x + b + c$ avec

$$c := -\frac{1}{2} w^T x^{(1)} - \frac{1}{2} w^T x^{(-1)} - b$$

Vérifions maintenant que π_1 se trouve à équidistance entre $x^{(-1)}$ et $x^{(1)}$. D'une part on a

$$\begin{aligned} d(\pi_1, x^{(-1)}) &= \frac{|w^T x^{(-1)} + b + c|}{\|w\|} = \frac{|w^T x^{(-1)} + b - \frac{1}{2} w^T x^{(1)} - \frac{1}{2} w^T x^{(-1)} - b|}{\|w\|} \\ &= \frac{|\frac{1}{2} w^T x^{(-1)} - \frac{1}{2} w^T x^{(1)}|}{\|w\|} = \frac{|\frac{1}{2} w^T x^{(1)} - \frac{1}{2} w^T x^{(-1)}|}{\|w\|} \end{aligned}$$

D'autre part, on obtient

$$\begin{aligned} d(\pi_1, x^{(1)}) &= \frac{|w^T x^{(1)} + b + c|}{\|w\|} = \frac{|w^T x^{(1)} + b - \frac{1}{2} w^T x^{(1)} - \frac{1}{2} w^T x^{(-1)} - b|}{\|w\|} \\ &= \frac{|\frac{1}{2} w^T x^{(1)} - \frac{1}{2} w^T x^{(-1)}|}{\|w\|} \end{aligned}$$

Ce qui montre bien que $d(\pi_1, x^{(1)}) = d(\pi_1, x^{(-1)})$.

Posons maitenant $b_1 := b + c$ et considérons le plan π_2 d'équation $\frac{1}{\|w\| k} \left(w^T x + b_1 \right)$ avec $k := d(\pi_1, x^{(-1)}) = d(\pi_1, x^{(1)})$. Montrons que le plan π_2 vérifie

$$\frac{1}{\|w\|\, k} \left(w^T x^{(1)} + b_1 \right) = 1 \tag{4}$$

$$\frac{1}{\|w\|\, k} \left(w^T x^{(-1)} + b_1 \right) = -1 \tag{5}$$

Pour commencer, remarquons que comme $(x^{(1)}, 1) \in C_1$ alors $w^T x^{(1)} + b_1 \geq 0$. Par conséquent, on obtient

$$d(\pi_1, x^{(1)}) = \frac{|w^T x^{(1)} + b_1|}{\|w\|} = \frac{w^T x^{(1)} + b_1}{\|w\|}$$

Et donc

$$\begin{aligned} \frac{1}{\|w\|\, k} \left(w^T x^{(1)} + b_1 \right) &= \frac{1}{\|w\|\, d(\pi_1, x^{(1)})} \left(w^T x^{(1)} + b_1 \right) \\ &= \frac{1}{d(\pi_1, x^{(1)})} d(\pi_1, x^{(1)}) = 1 \end{aligned}$$

Ce qui montre (4).
Ensuite, remarquons que comme $(x^{(-1)}, -1) \in C_2$ alors $w^T x^{(-1)} + b_1 < 0$. Par conséquent, on obtient

$$d(\pi_1, x^{(-1)}) = \frac{|w^T x^{(-1)} + b_1|}{\|w\|} = -\frac{w^T x^{(-1)} + b_1}{\|w\|}$$

On en déduit alors

$$\frac{1}{\|w\|\,k}\left(w^T x^{(-1)} + b_1\right) = \frac{1}{\|w\|\,d(\pi_1, x^{(-1)})}\left(w^T x^{(-1)} + b_1\right)$$
$$= -\frac{1}{d(\pi_1, x^{(-1)})} d(\pi_1, x^{(-1)}) = -1$$

Ce qui montre (5) et par conséquent le plan π_2 est bien le plan cherché.

2. Considérons maintenant le plan π d'équation $w^T x + b$ qui maximise le problème suivant.

$$\max_{w,b} \frac{2}{\|w\|} \tag{6}$$

$$\text{tels que } y_i(w^T x_i + b) \geq 1 \quad 1 \leq i \leq N \tag{7}$$

Nous allons montrer que

(a) π sépare C_1 et C_2.

(b) π se trouve à équidistance entre $x^{(1)}$ et $x^{(-1)}$.

(c) π maximise les marges.

Montrons les trois points qui nous sont demandés les uns après les autres

(a) Pour commencer, montrons que π sépare C_1 et C_2. Comme $y_i(w^T x_i + b) \geq 1\ \forall i$ alors en particulier $y_i(w^T x_i + b) \geq 0\ \forall i$ et donc par le point 1 de cet exercice, π sépare C_1 et C_2.

(b) Ensuite, montrons que π se trouve à équidistance entre $x^{(1)}$ et $x^{(-1)}$. Ce point étant un peu plus complexe, nous allons procéder par étape.

i. Pour commencer, montrons qu'il existe $(x_i, y_i) \in C_1$ et $(x_j, y_j) \in C_2$ tel que

$$y_i(w^T x_i + b) = 1 \tag{8}$$

$$y_j(w^T x_j + b) = 1 \tag{9}$$

Pour montrer cela, nous allons utiliser une preuve par l'absurde. Par l'absurde supposons donc que $y_k(w^T x_k + b) > 1\ \forall k$. Par conséquent, il existe $\alpha > 0$ avec $0 < \alpha < 1$ tel que $y_k(\alpha w^T x_k + \alpha b) \geq 1$. Par conséquent, le plan $\alpha w^T x + \alpha b$ satisfait la contrainte (7). Or de plus, comme $\alpha < 1$, remarquons que $\frac{2}{\|\alpha w\|} > \frac{2}{\|w\|}$, ce qui contredit le fait que le plan π maximise le problème d'optimisation.

L'argument présenté ci-dessus montre donc qu'il existe $(x_i, y_i) \in \mathcal{D}$ avec $y_i(w^T x_i + b) = 1$. Sans perte de généralité, nous allons supposer que $(x_i, y_i) \in C_1$.

Montrons maintenant qu'il existe $(x_j, y_j) \in C_2$ avec $y_j(w^T x_j + b) = 1$. Comme avant, nous allons faire une preuve par l'absurde. Par l'absurde, supposons donc que $y_k(w^T x_k + b) > 1\ \forall (x_k, y_k) \in C_2$. Par conséquent, il existe $c > 0$ avec $y_k(w^T x_k + b + c) = y_k(w^T x_k + b) - c > 0$. De plus, comme $c > 0$ et que $y_l = 1$ si $(x_l, y_l) \in C_1$ alors $w^T x + b + c$ vérifie également $y_l(w^T x_l + b + c) > 1\ \forall (x_l, y_l) \in C_1$. Par conséquent, le plan $w^T x + b + c$ vérifie $y_i(w^T x_i + b + c) > 1\ \forall (x_i, y_i) \in \mathcal{D}$. Ensuite, nous allons procéder comme avant. En effet, il existe $\alpha > 0$ avec $0 < \alpha < 1$ tel que $y_i(\alpha w^T x_i + \alpha b + \alpha c) \geq 1\ \forall (x_i, y_i) \in \mathcal{D}$. Et par conséquent, le plan $\alpha w^T x_i + \alpha b + \alpha c$ vérifie la contrainte (7). Et de plus, remarquons que $\frac{2}{\|\alpha w\|} > \frac{2}{\|w\|}$, ce qui contredit le fait que le plan π maximise le

problème d'optimisation présenté ci-dessus.

Ces deux arguments montrent donc qu'il existe $(x_i, y_i) \in C_1$ ainsi que $(x_j, y_j) \in C_2$ qui vérifient (8) et (9).

ii. Montrons maintenant que $x_i = x^{(1)}$ et $x_j = x^{(-1)}$. Pour faire cela, il suffit de vérifier que si $(x_k, y_k) \in C_1$ alors $d(\pi, x_k) \geq d(\pi, x_i)$ et que si $(x_l, y_l) \in C_2$ alors $d(\pi, y_l) \geq d(\pi, x_j)$.

Soit donc $(x_k, y_k) \in C_1$. Remarquons alors que $y_k(w^T x_k + b) \geq 1$ mais comme $y_k = 1$, alors $(w^T x_k + b) \geq 1$. De plus, remarquons que comme $(x_i, y_i) \in C_1$ et que $y_i(w^T x_i + b) = 1$ alors $w^T x_i + b = 1$. Ces deux observations impliquent que

$$d(\pi, x_k) = \frac{|w^T x_k + b|}{\|w\|} = \frac{w^T x_k + b}{\|w\|} \geq \frac{1}{\|w\|} = \frac{w^T x_i + b}{\|w\|} = d(\pi, x_i)$$

Ce qui montre bien que $d(\pi, x_k) \geq d(\pi, x_i)$

Soit maintenant $(x_l, y_l) \in C_2$. Comme $y_l(w^T x_l + b) \geq 1$ et que $y_l = -1$ alors $-w^T x_l - b \geq 1$. De plus, comme $(x_j, y_j) \in C_2$ et que $y_j(w^T x_j + b) = 1$, alors $-w^T x_j - b = 1$. On en déduit alors

$$d(\pi, x_l) = \frac{|w^T x_l + b|}{\|w\|} = \frac{-w^T x_l - b}{\|w\|} \geq \frac{1}{\|w\|} = \frac{-w^T x_j - b}{\|w\|} = d(\pi, x_j)$$

Ce qui montre bien que $d(\pi, x_l) \geq d(\pi, x_j)$.

(c) Pour finir, nous allons montrer que le plan π maximise les marges. Pour commencer, remarquons que le point 2 de cet exercice, alors $w^T x^{(1)} + b = 1$ et $w^T x^{(-1)} + b = -1$. On en déduit alors

$$\begin{aligned} d(\pi, x^{(1)}) + d(\pi, x^{(-1)}) &= \frac{|w^T x^{(1)} + b|}{\|w\|} + \frac{|w^T x^{(-1)} + b|}{\|w\|} \\ &= \frac{1}{\|w\|} + \frac{1}{\|w\|} = \frac{2}{\|w\|} \end{aligned}$$

Ce qui montre que la fonction objective de notre problème d'optimisation correspond bien à la somme des deux marges.

Exercice 1

1. Pour commencer supposons que les deux classes C_1, C_2 sont séparables. Le raisonnement mathématique présenté ci-dessus explique pourquoi trouver le plan $w^T x + b$ qui maximise les marges est équivalent à trouver le plan qui maximise le problème suivant

$$\max_{w,b} \frac{2}{\|w\|} \tag{10}$$

$$\text{tels que } y_i(w^T x_i + b) \geq 1 \quad 1 \leq i \leq N \tag{11}$$

En pratique, il est plus commun de considérer le problème suivant

$$\min_{w,b} \frac{\|w\|^2}{2} \tag{12}$$

$$\text{tels que } y_i(w^T x_i + b) \geq 1 \quad 1 \leq i \leq N \tag{13}$$

Expliquer pourquoi ces deux problèmes sont équivalents.

2. En pratique les classes C_1 et C_2 ne sont que rarement séparables. Pour trouver un plan pertinent dans ce cas (et minimiser les risques d'overfitting), nous allons introduire de nouvelles variables $\xi_i \geq 0$ ($1 \leq i \leq N$) ainsi qu'un hyperparamètre C. Expliquer pourquoi dans cette situation, il pourrait être pertinent de considérer le problème d'optimisation suivant.

$$\min_{w,b,\xi_i} \frac{\|w\|^2}{2} + C \sum_{i=1}^{N} \xi_i \tag{14}$$

$$\text{tels que } y_i(w^T x_i + b) \geq 1 - \xi_i \quad 1 \leq i \leq N \tag{15}$$

Expliquer également le rôle des ξ_i et de C.