Fondamentaux de l'Apprentissage Automatique

Lecturer: Yann Chevaleyre
Scribe: MAANINOU MEHDI

Lecture n°7 #
09/11/2023

Table des matières

1 Introduction **1**

2 Résolution problème linéairement séparable **1**
2.1 Formulation du problème 1
2.2 Exemple en 2D 3
2.3 Les notions de géometrie : marge et distance 3
2.4 Hyperplan cannonique 5
2.5 The Perceptron Algorithm (online) 5
2.6 Marge et borne de generalisation 7
2.7 Formulation du probleme de maximisation de marge 7
2.8 Les vecteurs support 9

3 Cas linéairement non séparable **11**
3.1 Formulation du problème 11
3.2 influence de C 13
3.3 Exemple 13
3.4 Relation entre Soft-SVM, Hinge-loss, and Hinge-loss Perceptron 14

4 Conclusion **15**

1 Introduction

Le séparateur à vaste marge (SVM) est un algorithme d'apprentissage supervisé issu d'une généralisation des classificateurs linéaires et destiné à résoudre des problèmes de classification et de régression.

Nous explorerons la séparation de données, la maximisation de la marge, et plongerons dans l'algorithme du perceptron. Nous aborderons ensuite la résolution des problèmes SVM pour les données linéairement séparables, puis nous étudierons comment étendre cette approche aux données non linéairement séparables. Enfin, nous clôturerons en examinant l'application pratique des SVM dans des contextes réels

2 Résolution problème linéairement séparable

2.1 Formulation du problème

$\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}\}_{i=1\cdots n}$: ensemble de points étiquetés.
Construire à partir de $\mathcal{D}$ une fonction $f : \mathcal{X} \to \{-1, 1\}$ ou $f : \mathcal{X} \to \mathbb{R}$ qui permet de prédire la classe -1 ou 1 d'un point $x \in \mathcal{X}$

On suppose l'espace des entrées $\mathcal{X} = \mathbb{R}^d$
Fonction de décision : $f : \mathbb{R}^d \to \mathbb{R}$ telle que si :

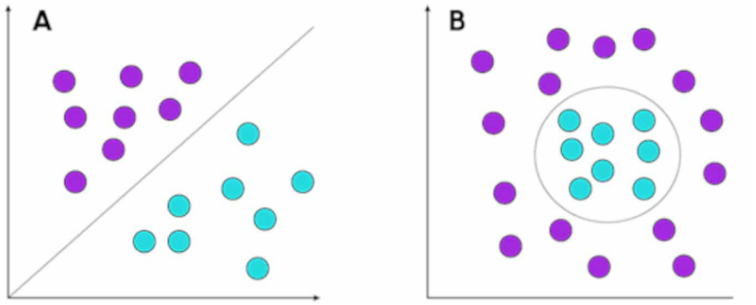
$$\begin{array}{ll} f(x) < 0 & \text{affecter } x \text{ à la classe } -1 \\ f(x) > 0 & \text{affecter } x \text{ à la classe } 1 \end{array}$$

- Fonction de décision linéaire :

$$f(x) = w^\top x + b, \quad w \in \mathbb{R}^d, b \in \mathbb{R}$$

Definition 1. ***Problème lineairement séparable****. Les points (xi, yi) sont linéairement séparables si il existe un hyperplan qui permet de discriminer correctement l'ensemble des données. Dans le cas contraire, on parle d'exemples non séparables linéairement*

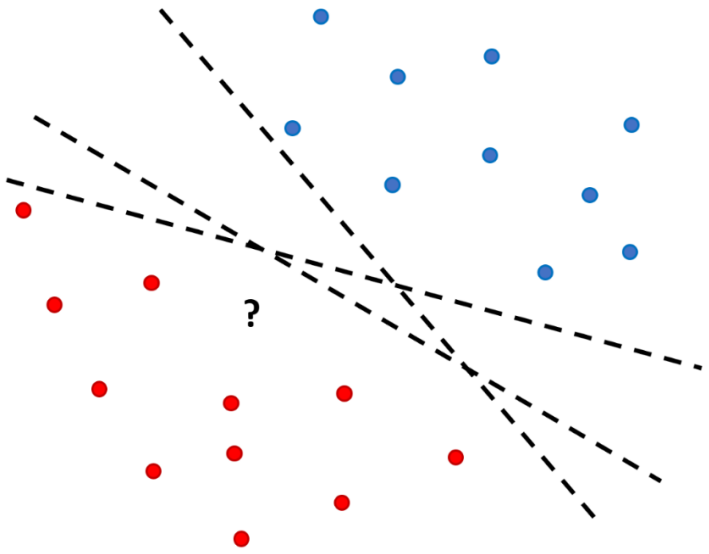
FIGURE 1 – linéairement separable et non linéairement separable



2.2 Exemple en 2D

Dans le cas où le problème est linéairement séparable, il existe une infinité de facon de séparer un ensemble de points étiquetés

FIGURE 2 – fonctions linéaires séparant les 2 classes



Parmi tout ces hyperplans on veux choisir celui qui maximise la marge entre les points des classes.

2.3 Les notions de géometrie : marge et distance

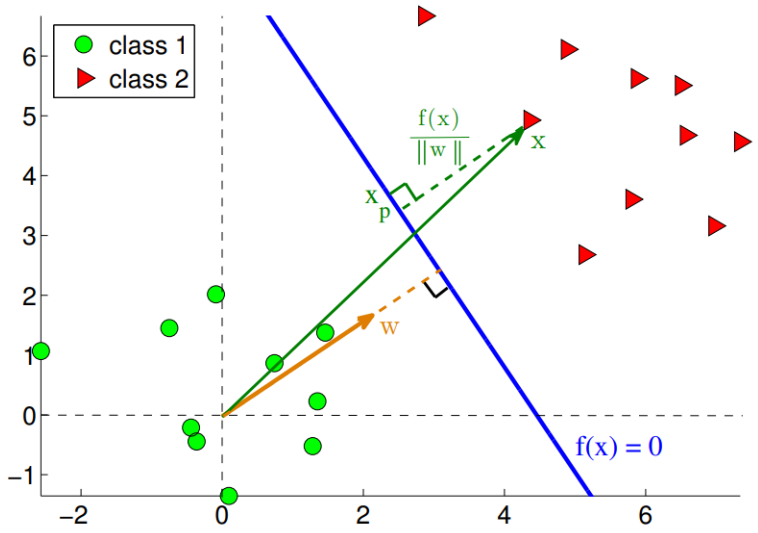
Plus précisément, la « marge » d'un problème d'apprentissage est définie comme la distance entre le plus proche exemple d'apprentissage et l'hyperplan de séparation. Pour un hyperplan H on a :

$$\text{Marge}(H) = \min_{x_i} d\left(x_i, H\right)$$

Voici comment on définit la distance d'un point par rapport à la frontière de décision : Soit $H(w,b) = \left\{z \in \mathbb{R}^d \mid f(z) = w^\top z + b = 0\right\}$ un hyperplan et soit $x \in \mathbb{R}^d$.
La distance du point x à l'hyperplane H est :

$$d(x,H) = \frac{\left|w^\top x + b\right|}{\|w\|} = \frac{|f(x)|}{|w|}$$

FIGURE 3 – Démonstration. (Distance d'un point à l'hyperplan)



Démonstration.

$$x = x_p + \frac{w}{\|w\|} \times d$$

Prenez le produit scalaire de x avec w

$$w^\top x = w^\top x_p + \underbrace{w^\top * \frac{w}{\|w\|} \times d}_{\|w\|d}$$

$$\begin{aligned} \|\omega\| d =& w^\top x - \omega^\top x_p \\ =& \left(w^\top x + b\right) - \underbrace{\left(\omega^\top x_p + b\right)}_{=0} \end{aligned}$$

On obtient finalement $d = \frac{w^\top x + b}{\|w\|}$

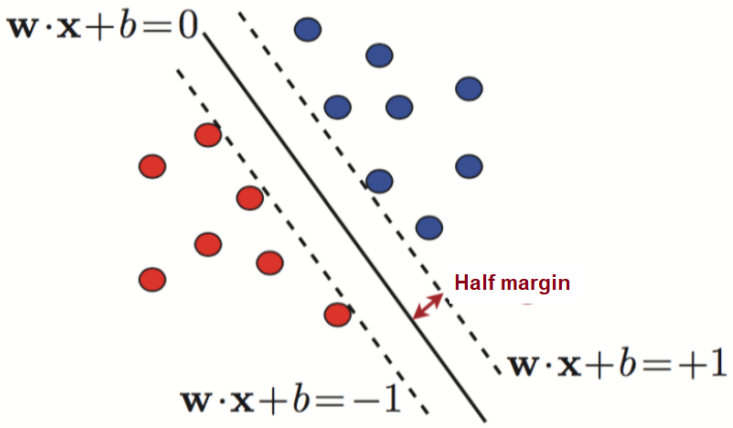
□

2.4 Hyperplan cannonique

Definition 2. ***Hyperplan cannonique****. Un hyperplan est dit canonique par rapport aux données*

$$\{x_1, \ldots, x_N\} \ si \ \min_i |w^\top x_i + b| = 1.$$

FIGURE 4 – schema d'un hyperplan cannonique



Un hyperplan canonique respecte deux conditions $y_i \cdot f(x_i) \geq 1$ ce qui signifie :

1. Tous les points sont correctement classifiés, c'est-à-dire que $y_i \cdot f(x_i) > 0$.
2. Tous les points sont à une distance d'au moins $\frac{1}{\|w\|}$ de l'hyperplan, c'est-à-dire que $|f(x_i)| = |w \cdot x_i + b| > 1$.

Donc dans le cas d'un hyperplan canonique, on a $M = \frac{2}{\|w\|}$

2.5 The Perceptron Algorithm (online)

L'algorithme du Perceptron est un modèle d'apprentissage supervisé qui vise à séparer deux classes par une ligne droite dans l'espace des caractéristiques en ajustant les poids associés aux caractéristiques. Il est principalement utilisé pour la classification binaire et est adapté aux problèmes de classification linéaire. Lorsque le problème est non linéaire il est bien moins performant que SVM.

Algorithme 1 : Perceptron algorithm for non homogenous hyperplan

Entrée : $(x_i, y_i)_{i=1}^n$ `// Ensemble de données d'apprentissage`
Nombre d'iteration max T

```
// Initialisation
```
$w_0 \longleftarrow 0$;
$b_0 \longleftarrow 0$;
$t \longleftarrow 0$;

```
// Boucle d'entrainement
```
while $(t < T)$ *ou (il existe encore des erreurs)* **do**
  `// Pour chaque point de l'ensemble des données d'apprentissage`
  for $(x_i, y_i)_{i=1}^n$ **do**
    `// Prédiction`
    Predict $f_t(x_i) = \text{sign}(\omega_t \cdot x_i + b)$;
    `// Vérification d'erreur`
    if $f_t(x_i) \neq y_i$ **then**
      $w_{t+1} \leftarrow w_t + y_i \cdot x_i$;
      $b_{t+1} \leftarrow b_t + y_i$;
    else
      $w_{t+1} = w_t$;
      $b_{t+1} = b_t$;
    $t = t + 1$;

Pour retrouver le cas homogène il suffit de ne pas prendre en compte le biais dans l'algorithme

Theorem 1. *Théorème de Novikoff.*
Soit $D = \{(x_t, y_t)\}$ un ensemble de données tel que, pour tout t, $\|x_t\| < R$ et $y_t \in \{-1, 1\}$. Supposons qu'il existe un hyperplan canonique w^* qui classe parfaitement les données, passant par l'origine, avec une marge de moitié $\rho = \frac{1}{\|w^*\|}$.

Alors, le nombre d'erreurs commises par l'algorithme du perceptron est au plus $\frac{R^2}{\rho^2}$.

Demonstration 1. .
STEP 1 : After an update (classification error)
w_{t+1} is more aligned' to ω^* :

$$\begin{aligned}\langle w_{t+1}, w^* \rangle &= \langle w_t + y_t x_t, w^* \rangle \\ &= \langle w_t, w^* \rangle + \underbrace{y_t \langle x^t, w^* \rangle}_{\geqslant 1 \ w^* \text{ is canonical}} \\ &\geqslant \langle w_t, w^* \rangle + 1\end{aligned}$$

Unrolling we get for all t :

$$\langle w_t, w^* \rangle \geqslant \underbrace{t_e}_{\text{number of mistakes}}$$

STEP 2 : After an update (classification error)

$$\begin{aligned}\|\omega_{t+1}\|^2 &= \langle \omega_t + y_t x_t, \omega_t + y_t x_t \rangle \\ &= \|w_t\|^2 + \underbrace{2y_t \langle w_t, x_t \rangle}_{\leqslant 0 \text{ (erreur de classification à cette étape)}} + \|y_t x_t\|^2 \\ &\leqslant \|w_t\|^2 + R^2\end{aligned}$$

Unrolling we get for all t :

$$\|\omega_t\|^2 \leqslant t_e R^2$$

STEP 3 :

$$t_e \leqslant \langle w_t, w^* \rangle \leqslant \underbrace{\|w_t\| \cdot \|w^*\|}_{\text{Cauchy-Schwartz}} \leqslant \underbrace{\sqrt{t_e} \cdot R \cdot \|w^*\|}_{\text{step 2}} = \sqrt{t_l}\frac{R}{\rho}$$

donc

$t_l \leqslant \frac{R^2}{\rho^2}$

Ici, on peut voir l'importance d'avoir une marge importante car elle réduit le nombre d'erreurs.

□

2.6 Marge et borne de generalisation

Risque sur une classe de fonction $\mathcal{H}$. Avec une probabilité $1 - \delta$

$R(\hat{h}) \leq R_{\text{emp}}(\hat{h}) + C\sqrt{\frac{D(\log(2N/D)+1)+\log(4\delta)}{N}}$

Cette inégalité est intéressante car elle fournit une garantie théorique sur la performance du modèle ($R(\hat{h})$) basée sur son erreur empirique ($R_{\text{emp}}(\hat{h})$) et d'autres facteurs liés à la complexité de l'espace des hypothèses (D) et à la taille de l'ensemble d'entraînement (N). Elle illustre comment la qualité de l'apprentissage sur les données d'entraînement peut être liée à la capacité de généralisation sur de nouvelles données. L'inclusion de la VC-dimension suggère également que des classes d'hypothèses plus complexes peuvent nécessiter des ensembles d'entraînement plus importants pour obtenir une bonne généralisation.

VC dim de la classe des fonctions linéaires à marge ρ :
Soit $\mathcal{H}$ la classe de fonction $f(x) = w^T x + b$ à une marge ρ des exemples d'apprentissage alors

$D \leq 1 + \min\left(d, \frac{R^2}{\rho^2}\right)$,

où R est le rayon d'une boule contenant les données d'apprentissage.
Le terme d représente la dimension intrinsèque de l'espace dans lequel les exemples d'apprentissage résident. ex :($\mathbb{R}^d$)

2.7 Formulation du probleme de maximisation de marge

Un ensemble d'apprentissage $D = \{(x_i, y_i)\}$ où $y_i \in \{-1, 1\}$. L'objectif est de trouver un hyperplan de séparation de la forme $w^T x + b = 0$ qui maximise la marge en discriminant

correctement les points de D.
On sait que la marge

$$M = \frac{2}{\|w\|}$$

On arrive donc au problème d'optimisation suivant (appelé problème primal) :

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{tel que } y_i(w \cdot x_i + b) \geq 1, \quad i = 1, \ldots, n$$

On a choisis de mettre le problème sous cette forme car cela permet de faciliter l'optimisaton. Nous allons résoudre ce problème à l'aide De la méthode des multiplicateurs de Lagrange.

Formulation du Problème Primal :

Le problème primal du SVM consiste à trouver les poids w et le biais b qui définissent le plan de séparation optimal. Cela peut être formulé comme une maximisation de la marge tout en minimisant la norme des poids w. Le problème primal est le suivant :

$$\text{Maximiser : } L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(w^T x_i + b) - 1]$$

où α_i sont les multiplicateurs de Lagrange.

Dérivation des Conditions du Dual :

Pour optimiser le lagrangien, on prend les dérivées partielles par rapport à w et b et on les égalise à zéro.

1. En annulant la dérivée partielle par rapport à b, on obtient :
   $$\sum_{i=1}^{n} \alpha_i y_i = 0.$$
   Cela provient des conditions de Karush-Kuhn-Tucker (KKT).
2. En annulant la dérivée partielle par rapport à w, on obtient :
   $$w = \sum_{i=1}^{n} \alpha_i y_i x_i.$$
   Cela exprime les poids w en fonction des multiplicateurs α_i.

Substitution dans le Lagrangien :

Après avoir obtenu les expressions pour w et b en annulant les dérivées partielles du lagrangien, nous avons :

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \quad \text{et} \quad \sum_{i=1}^{n} \alpha_i y_i = 0.$$

Maintenant, nous pouvons substituer ces expressions dans le lagrangien $L(w, b, \alpha)$ que nous avions initialement :

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(w^T x_i + b) - 1]$$

Substituant w dans le lagrangien :

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{n}\alpha_i$$

En simplifiant davantage le lagrangien en utilisant les expressions pour w et b, nous obtenons :

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{n}\alpha_i$$

Le problème dual peut donc être formulé comme suit :

$$\max_{\alpha_i} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

Sous les contraintes : $\alpha_i \geq 0$ pour $i = 1, \ldots, n$, et $\sum_{i=1}^{n}\alpha_i y_i = 0$.

Contraintes de complémentarité : $\alpha_i h_i(w) = 0 \quad$ avec $\quad h_i(w) = 1 - y_i(w^T x_i + b)$.

2.8 Les vecteurs support

On obtient deux types de paramètres α :

1. Pour un point x_j :

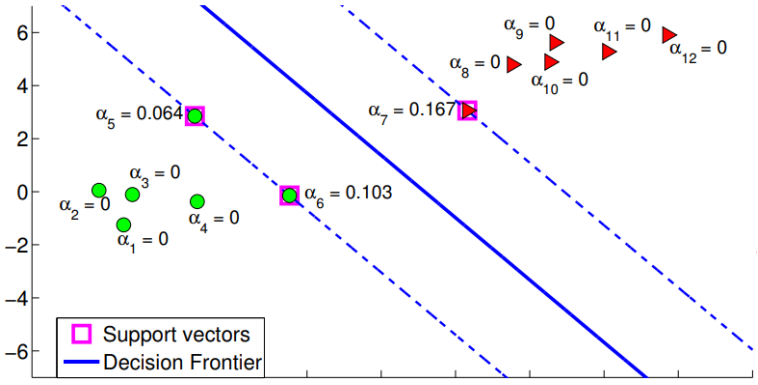
$$\alpha_j = 0 \quad \text{si} \quad y_j(w^T x_j + b) > 1.$$

2. Pour un point x_i :

$$\alpha_i \geq 0 \quad \text{si} \quad y_i(w^T x_i + b) = 1.$$

On comprend donc que seuls les vecteurs supports contribuent à la définition de w, ce qui signifie que les autres exemples qui sont correctement classés avec une marge suffisante (pour lesquels $y_i(w^T x_i + b) > 1$) n'influencent pas la détermination de w.

FIGURE 5 – expression de w avec les vecteurs support



Par exemple, dans le schéma précédent, on obtient :

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$w = 0.16 \cdot x_7 - 0.064 \cdot x_5 - 0.103 \cdot x_6$$

3 Cas linéairement non séparable

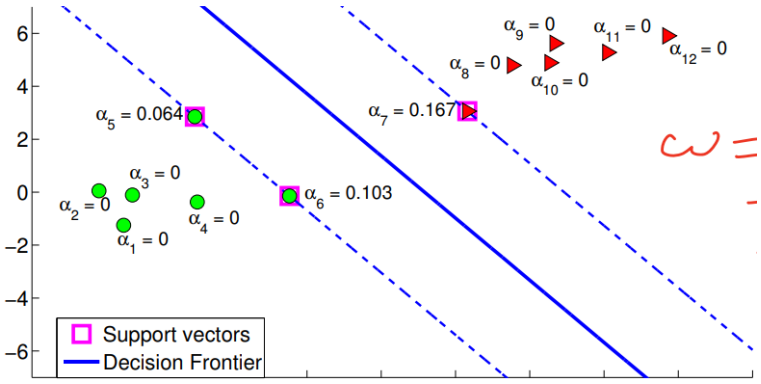
3.1 Formulation du problème

Souvent il arrive que les deux classes se retrouvent mélangées autour de l'hyperplan de séparation. Pour gérer ce type de problème, on utilise une technique dite de marge souple, qui tolère les mauvais classements :

La marge souple est introduite en ajoutant des variables de relâchement ξ_i.

Ensuite, on pénalise ces "erreurs" $\sum_{i=1}^{n} \xi_i$ dans le problème de SVM.

FIGURE 6 – Variable de relâchement dans le cas non linéairement séparable



Nous avons 2 situations possibles :

1. Pas d'erreur : $y_i(w^T x_i + b) \geq 1 \implies \xi_i = 0$,
2. Erreur : $y_i(w^T x_i + b) < 1 \implies \xi_i = \max(0, 1 - y_i(w^T x_i + b)) > 0$.

La fonction coût charnière associée est définie comme :

$$\xi_i = \max(0, 1 - y_i(w^T x_i + b))$$

Le problème d'optimisation dans le cas des données non-séparables est donc :

$$\begin{cases} \min_{w,b} \left(\frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i\right) \\ \text{tel que } y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, n \\ \text{et } \xi_i \geq 0, \quad i = 1, \ldots, n \end{cases}$$

avec $C > 0$: paramètre de régularisation (compromis entre erreur et marge à fixer par l'utilisateur)

Formulation du Problème d'optimisation

Pour formuler le problème d'optimisation avec variables d'écart en utilisant la méthode des multiplicateurs de Lagrange, introduisons les multiplicateurs de Lagrange α_i pour les contraintes d'inégalité et β_i pour les contraintes de non-négativité des variables d'écart ξ_i. Le lagrangien associé à ce problème est donné par :

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^{n}\nu_i\xi_i$$

Dérivation des conditions du dual

Le problème dual est obtenu en cherchant à maximiser le lagrangien par rapport aux multiplicateurs de Lagrange α_i et β_i, tout en minimisant par rapport à w, b, et ξ_i. Pour cela, on annule les dérivées partielles par rapport à w, b, et ξ_i, et on impose les conditions de stationnarité.

Les dérivées partielles par rapport à w, b, et ξ_i sont respectivement données par :

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n}\alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \quad \text{pour } i = 1, \ldots, n$$

En imposant ces conditions de stationnarité dans le lagrangien, on obtient les relations suivantes :

$$w = \sum_{i=1}^{n}\alpha_i y_i x_i$$

$$\sum_{i=1}^{n}\alpha_i y_i = 0$$

$$C - \alpha_i - \beta_i = 0 \quad \text{pour } i = 1, \ldots, n$$

Ces relations expriment les paramètres optimaux w, b, et ξ_i en termes des multiplicateurs de Lagrange α_i et β_i.

Substitution dans le lagrangien

Par la même procédure qu'avant, on obtient le problème dual :

$$\text{Maximiser :} \quad \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

$$\text{Sous les contraintes :} \quad 0 \leq \alpha_i \leq C \quad \text{pour } i = 1, \ldots, n \quad \text{(contraintes d'admissibilité)}$$

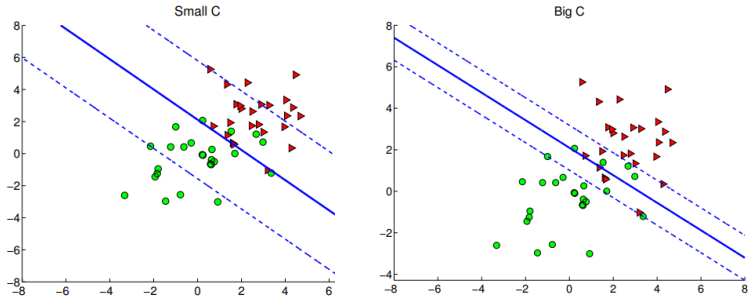
$$\sum_{i=1}^{n}\alpha_i y_i = 0 \quad \text{(contrainte de stationnarité)}$$

Theoreme. [Solution d'un SVM linéaire : cas non séparable]
Soit un problème de SVM linéaire non-séparable de fonction de décision
$f(x) = w^T x + b$. Le vecteur w est défini par $w = \sum_{i=1}^{n} \alpha_i y_i x_i$
où les coefficients α_i sont solutions du problème dual ci-dessus.

3.2 influence de C

Le choix approprié de C dépend du problème spécifique et du compromis souhaité entre la classification précise des données d'entraînement et la capacité à généraliser à de nouvelles données.

FIGURE 7 – Variable de relâchement dans le cas non linéairement séparable

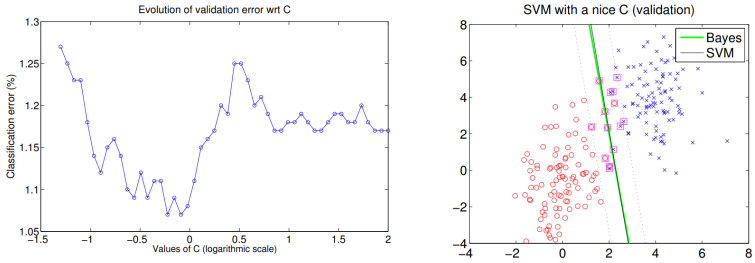


Le choix de C influence la solution :
— C petit, marge grande
— C grand, marge petite.

3.3 Exemple

— Les valeurs de C sont choisies sur une échelle logarithmique.
— Pour chaque C, on apprend un SVM et on calcule son erreur de validation.
— Le minimum de la courbe d'erreur correspond à la "meilleure" valeur C^*.
— Le SVM correspondant est illustré sur la figure de droite.

FIGURE 8 – recherche de la meilleur valeur de C



3.4 Relation entre Soft-SVM, Hinge-loss, and Hinge-loss Perceptron

The soft-SVM problem (SVM with slack variables) is given by :

$$\min_{\omega} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}\xi_i$$

constraints n ξ : $\begin{array}{l} y_i(\langle\omega, x_i\rangle + b) \geq 1 - \xi_i, \\ \xi_i \geq 0. \end{array}$

The constraint on ξ_i can be further simplified as follows :

$$\begin{aligned} \xi_i &\geq 1 - y_i(\langle\omega, x_i\rangle + b) \\ \xi_i &\geq \max(0, 1 - y_i(\langle\omega, x_i\rangle + b)) \\ \xi_i &\geq \max\left(0, 1 - y_i s_i\right), \end{aligned}$$

where $s_i = \langle\omega, x_i\rangle + b$.

Consider the optimization sub-problem :

$$\begin{aligned} &\min_{\boldsymbol{\xi}} \sum_i \xi_i \\ &\text{s.t. } \xi_i \geq \max(0, 1 - y_i s_i), \end{aligned}$$

The solution to this sub-problem is :

$$\xi_i = \max(0, 1 - y_i s_i),$$

and this is also the solution for ξ_i on x_i in the original problem.

The soft-SVM problem becomes :

$$\min_{\omega} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}\max(0, 1 - y_i(\langle\omega, x_i\rangle + b))$$

Which reduces to :

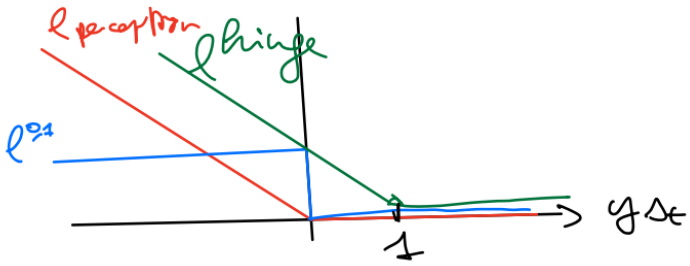
$$\min_{\omega} \frac{1}{2C}\|\omega\|^2 + \sum_{i=1}^{N}\ell_{\text{hinge}}(\langle\omega, x_i\rangle + b, y_i)$$

where $\ell_{\text{hinge}}(s_i, y_i) = \max(0, 1 - y_i s_i)$.

Perceptron loss :

$$(\ell_{\text{perceptron}}(s_i, y_i) = \max(0, -y_i s_i)$$

FIGURE 9 – Comparaison des différentes loss



4 Conclusion

En conclusion, la construction d'un hyperplan optimal, visant à maximiser la marge, se révèle théoriquement solide en minimisant l'erreur de généralisation. L'extension à des cas non linéaires par l'utilisation de noyaux et la généralisation à plusieurs classes confèrent à cet algorithme une polyvalence remarquable. Son adoption répandue en pratique atteste de son efficacité, faisant de la construction d'hyperplan un choix privilégié pour la classification dans divers contextes réels.