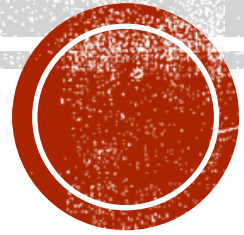
MACHINE À VECTEURS DE SUPPORTS

Hadjadj abdelhalim



Centre universitaire de Mila

Master I: Apprentissage automatique

INTRODUCTION AUX SVM

Les **Machines à Vecteurs de Support (SVM)** sont des algorithmes d'**apprentissage supervisé** utilisés pour résoudre des problèmes de **classification** et de **régression**. Elles sont particulièrement efficaces lorsque les données sont bien séparables ou lorsqu'un modèle robuste est nécessaire sur un petit ensemble de données.

MODÈLES LINÉAIRES POUR LA CLASSIFICATION

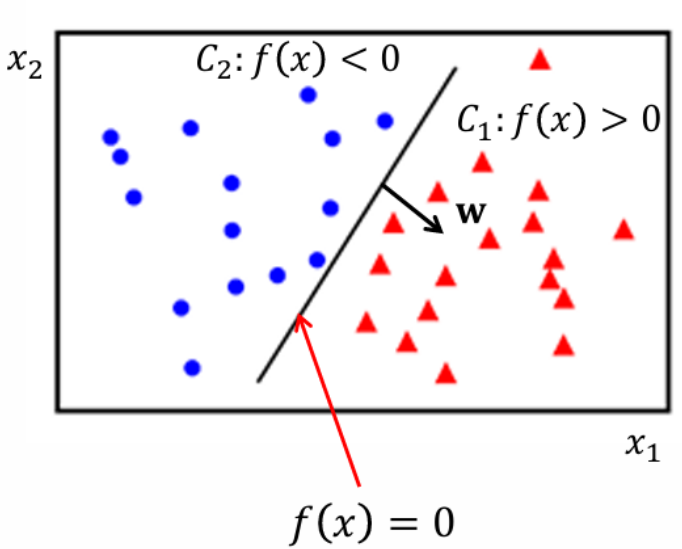
- Un classificateur linéaire cherche à trouver une frontière de décision sous forme d'un hyperplan qui divise les données en fonction de leurs classes.

- L'hyperplan est défini par l'équation :

$$f(x) = w_0 + \sum_{d=1}^{D} w_d x_d$$

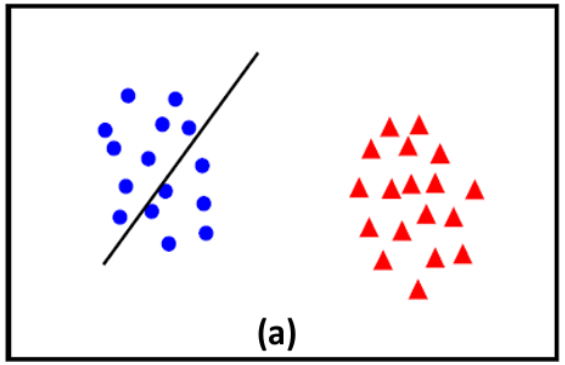
$$= w_0 + \mathbf{w}^T x$$

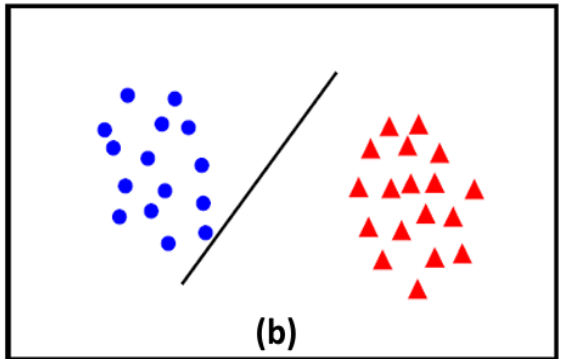
  - Où :
  - W^T est le vecteur des poids,
  - X est le vecteur d'entrée,
  - w_0 est le biais.
  - L'objectif est de trouver un hyperplan qui sépare au mieux les clas les erreurs de classification.



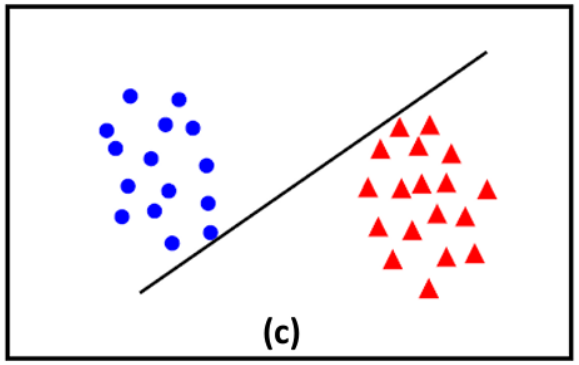
MODÈLES LINÉAIRES POUR LA CLASSIFICATION

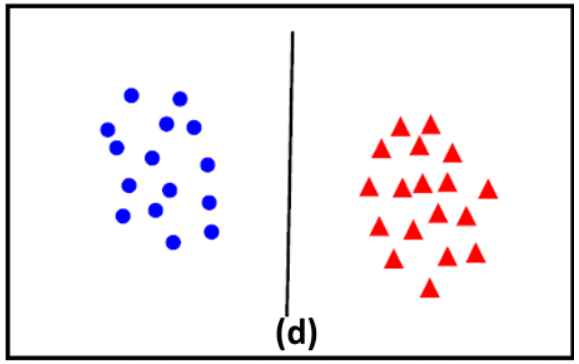
- Quelle est la meilleure frontière de décision?





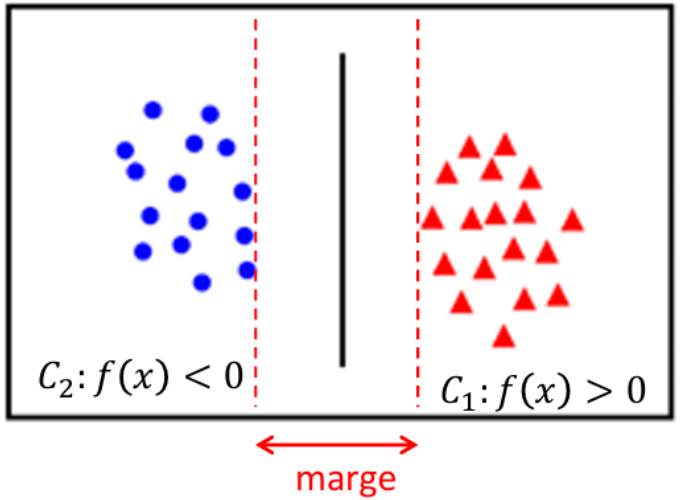
La meilleur fonction discriminante est celle qui généralise la classification et elle est stable par rapport aux nouvelles données.





MODÈLES LINÉAIRES POUR LA CLASSIFICATION

- La solution maximisant la marge d'erreur peut constituer une bonne alternative pour améliorer la généralisation.
- Par la suite, les classes C_1 et C_2 seront respectivement désignées comme classes positives et négatives.
- La variable **cible** est codée de la manière suivante : $y \in \{+1, -1\}$.



PRINCIPE DE FONCTIONNEMENT DES SVM

- Les SVM cherchent à **maximiser la marge** entre les classes en utilisant une **fonction de coût** qui pénalise les erreurs de classification.

- **Hyperplan et vecteurs de support**

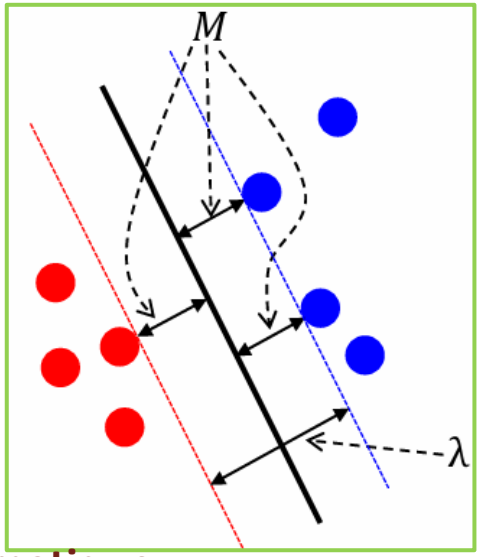
  - Un **hyperplan** est une surface de séparation dans un espace à plusieurs dimensions.

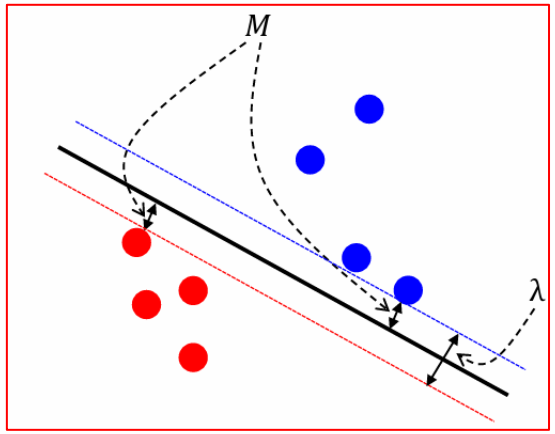
  - Les **vecteurs de support** sont les points les plus proches de l'hyperplan, qui influencent directement sa position.

  - La **marge** est la distance entre l'hyperplan et les vecteurs de support. Plus elle est grande, meilleure est la généralisation du modèle.

PRINCIPE DE FONCTIONNEMENT DES SVM

- Soit M la distance perpendiculaire entre la frontière de décision et les points les plus proches de chaque classe.
  - La marge est donnée par $\lambda=2\times M$
    - En cherche une frontière de décision qui maximise la marge

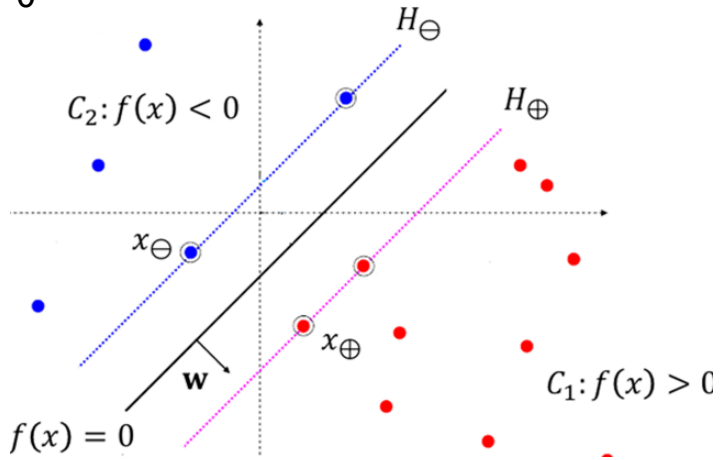




PRINCIPE DE FONCTIONNEMENT DES SVM

- On peut choisir les coefficients w et w_0 de manière à satisfaire les conditions suivantes

$$\begin{cases} \mathbf{w}^T x_\oplus + w_0 = +1. \\ \mathbf{w}^T x_\ominus + w_0 = -1. \end{cases}$$



La marge correspond à la distance entre ces deux hyperplans. Géométriquement, elle est donnée par : $\lambda = \frac{\mathbf{w}^T}{\|\mathbf{w}\|}(x_\oplus - x_\ominus) = \frac{\mathbf{w}^T(x_\oplus - x_\ominus)}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}$

où ||w|| représente la norme du vecteur w, qui mesure son *amplitude.*

Pour maximiser la marge, il faut minimiser la valeur de ||w||

PRINCIPE DE FONCTIONNEMENT DES SVM

- Démonstrations:

$$\lambda = \frac{\mathbf{w}^T}{\|\mathbf{w}\|}(x_\oplus - x_\ominus)$$

$$= \frac{1}{\|\mathbf{w}\|}(\mathbf{w}^T x_\oplus - \mathbf{w}^T x_\ominus)$$

$$= \frac{1}{\|\mathbf{w}\|}(\mathbf{w}^T x_\oplus + w_0 - \mathbf{w}^T x_\ominus - w_0)$$

$$= \frac{1}{\|\mathbf{w}\|}(1+1) = \frac{1}{\|\mathbf{w}\|}(1+1)$$

Donc $\lambda = \frac{2}{\|\mathbf{w}\|}$ **et** $M = \frac{1}{\|\mathbf{w}\|}$

PRINCIPE DE FONCTIONNEMENT DES SVM

- En ayant un ensemble d'apprentissage $\mathcal{D} = \{(x^{(i)}, y^{(i)}), i = 1, \dots, N\}$, l'algorithme de SVM procède alors comme suit:

$$argmax_{\mathbf{w}} \frac{2}{\|\mathbf{w}\|} \quad \text{sujet à:}$$

$$\begin{cases} \mathbf{w}^T x^{(i)} + w_0 > 1 \ \text{ si } \ y^{(i)} = +1. \\ \mathbf{w}^T x^{(i)} + w_0 < -1 \ \text{ si } \ y^{(i)} = -1. \end{cases} \qquad \forall i = 1, \cdots, N.$$

- Ce qui peut être réécrit, de manière équivalente, comme suit:

$$argmin_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 \ \text{sujet à:} \ (\mathbf{w}^T x^{(i)} + w_0) y^{(i)} > 1, \forall i = 1, .., N$$

PRINCIPE DE FONCTIONNEMENT DES SVM

- Le problème d'optimisation devient :

$$\arg\min_{\mathbf{w}} \quad \frac{1}{2}||\mathbf{w}||^2$$

- sous la contrainte :

$$y^{(i)}(\mathbf{w}^T x^{(i)} + w_0) \geq 1, \quad \forall i$$

- Pour déterminer w et w_0, nous devons intégrer ces contraintes dans la fonction objectif à l'aide des multiplicateurs de Lagrange.
- Les multiplicateurs de Lagrange ajoutent un terme pour chaque contrainte, garantissant que l'optimum de la nouvelle fonction coïncide avec celui du problème initial.

PRINCIPE DE FONCTIONNEMENT DES SVM

- La fonction objectif à optimiser, intégrant les contraintes, est définie comme suit :

$$Q(\mathbf{w}, w_0, \alpha) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{N} \alpha_i \left[y^{(i)}(\mathbf{w}^T x^{(i)} + w_0) - 1 \right]$$

- où les $\alpha_i \geq 0$ sont les multiplicateurs de Lagrange.
- L'optimisation consiste à minimiser Q par rapport à w et w_0, tout en la maximisant par rapport aux α_i.
- En annulant les dérivées partielles, nous obtenons les conditions suivantes :

$$\frac{\partial Q}{\partial \mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)}$$

$$\frac{\partial Q}{\partial w_0} = 0 \quad \Rightarrow \quad \sum_{i=1}^{N} \alpha_i y^{(i)} = 0$$

PRINCIPE DE FONCTIONNEMENT DES SVM

- En remplaçant la valeur de **w** dans la fonction objective Q, on obtient

$$Q = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \mathbf{w}^T\left(\sum_{i=1}^{N}\alpha_i y^{(i)}x^{(i)}\right) - w_0\left(\sum_{i=1}^{N}\alpha_i y^{(i)}\right) + \sum_{i=1}^{N}\alpha_i \quad = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\, y^{(i)}y^{(j)}x^{(i)^T}x^{(j)}$$

($\mathbf{w}$ remplace $\sum_{i=1}^{N}\alpha_i y^{(i)}x^{(i)}$; 0 remplace $\sum_{i=1}^{N}\alpha_i y^{(i)}$)

Cette fonction sera maximisée en fonction des α_i avec la contrainte:

$$\sum_{i=1}^{N}\alpha_i y^{(i)} = 0 \text{ et } \alpha_i >= 0, \forall i = 1,..,N$$

La fonction Q est maximisée en utilisant l'optimisation quadratique.

- L' appellation quadratique est due au terme $\alpha_i\alpha_j$.

• Il est à noter qu'il n'existe pas de solution analytique pour ce problème, mais il peut être simplement résolut avec des méthodes numériques.

PRINCIPE DE FONCTIONNEMENT DES SVM

- Une fois les valeurs optimale des α_i obtenues, la plupart d'elles vont s'annuler $\alpha_i \rightarrow 0$ et une petite portion seront supérieure à 0, $\alpha_i > 0$.

- L'ensemble des données $x^{(i)}$ pour lesquelles $\alpha_i > 0$ sont appelées les vecteurs de supports.
- Le vecteur **w** est réécrit comme une somme pondérée des vecteurs de supports. Ces vecteurs se trouvent sur la marge et donc satisfont l'équation:

- $$\left(\mathbf{w}^T x^{(i)} + w_0\right) y^{(i)} = 1 \qquad \text{on} \qquad \mathbf{w} = \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)}$$

PRINCIPE DE FONCTIONNEMENT DES SVM

- La valeur de w_0 peut être alors calculée directement:

$$w_0 = y^{(i)} - \mathbf{w}^T x^{(i)}$$

☞ On peut aussi utiliser **la moyenne sur tous les vecteurs de supports** pour avoir une valeur robuste de w_0.

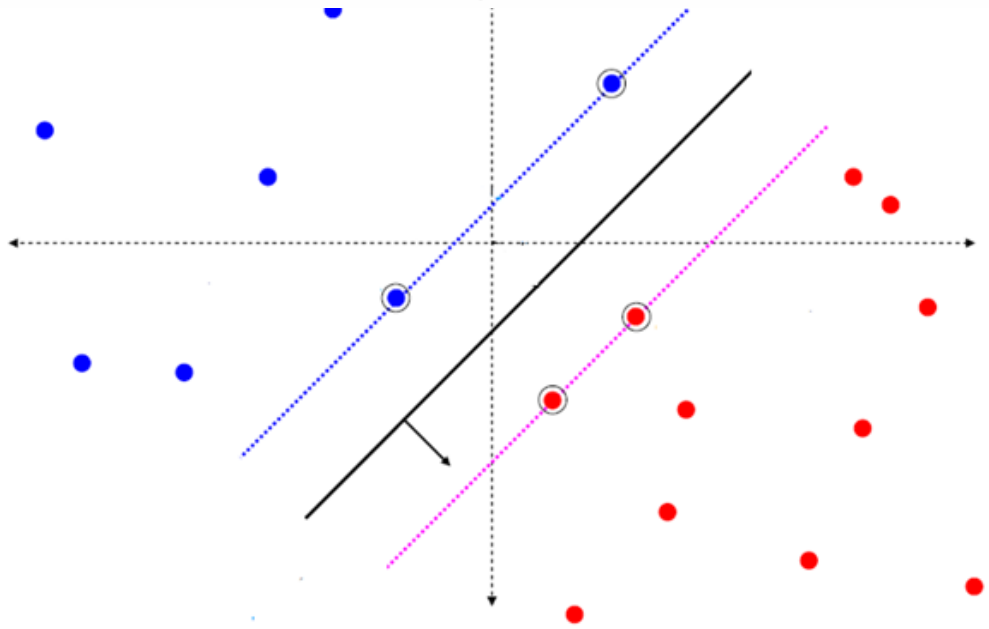
- La plupart des α_i sont nuls et on aura pour leurs données $\left(\mathbf{w}^T x^{(i)} + w_0\right) y^{(i)} > 1$.

- Ces données sont situées **loin de la marge** et elles n'ont aucun effet (information) sur la solution.

PRINCIPE DE FONCTIONNEMENT DES SVM

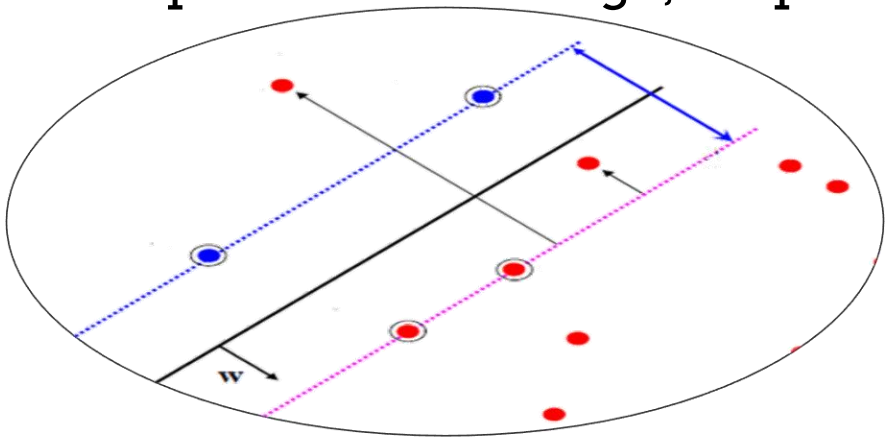
Pour classer une nouvelle donnée x, il faudra juste calculer:

$$f(x) = \mathbf{w}^T x + w_0 \begin{cases} \geq 0? \text{ classe } C_1 \\ < 0? \text{ classe } C_2 \end{cases}$$

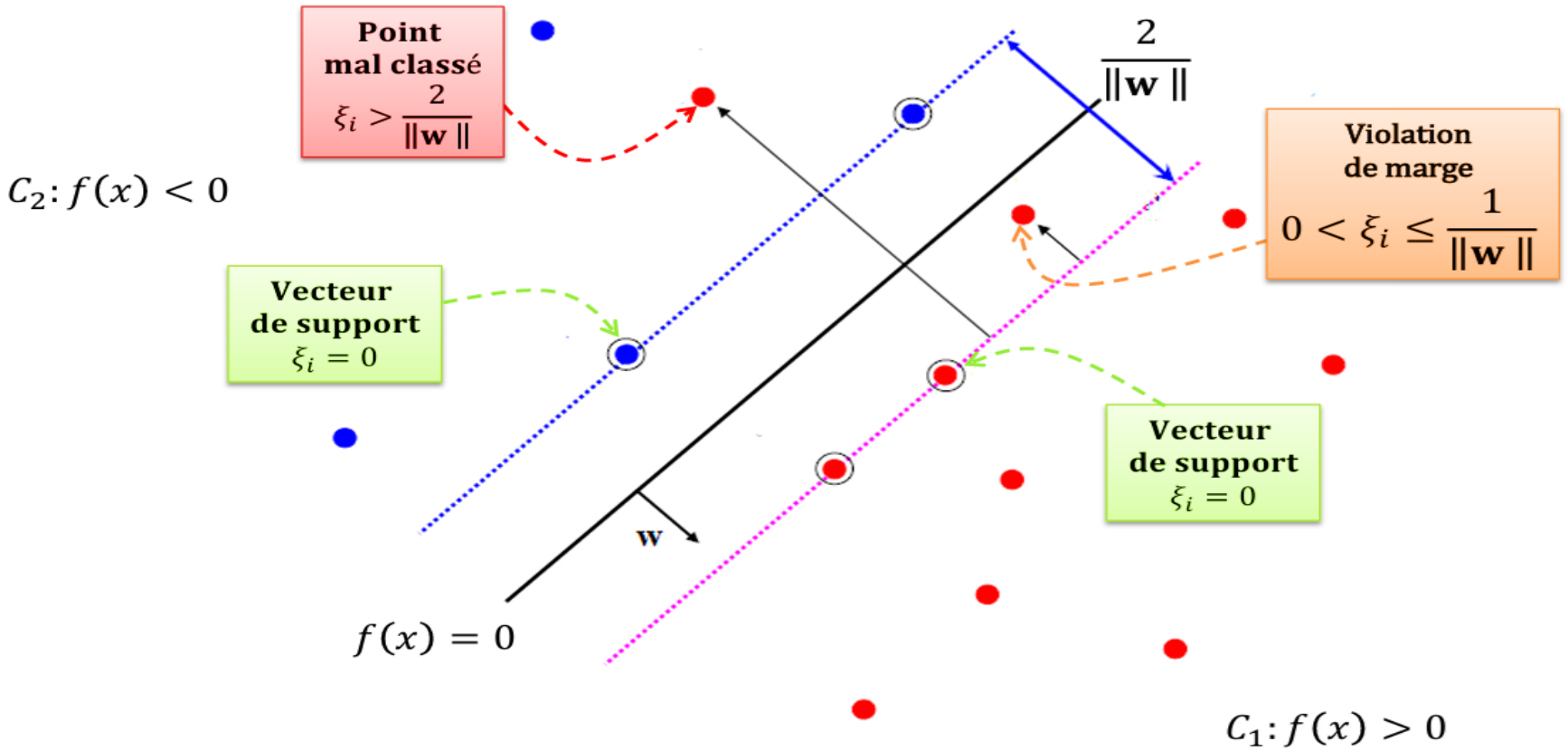


MACHINE À SUPPORTS DE VECTEURS CAS NON SÉPARABLE DE SVM

- Si les classes C1 et C2 ne sont pas séparables de manière linéaire, l'algorithme SVM classique ne pourra pas être appliqué directement.
- Il est possible d'introduire une certaine tolérance aux erreurs de classification afin de permettre une meilleure généralisation du modèle.
- Pour cela, une variable $\xi i \geq 0$ est associée à chaque donnée $\mathbf{x}^{(i)}$, indiquant l'écart de cette donnée par rapport à la marge :
  - Si $0 < \xi_i \leq 1$, alors le point se situe entre l'hyperplan de séparation et la marge, ce qui constitue une violation de la marge.
  - Si $\xi_i > 1$, cela signifie que le point est mal classé.



MACHINE À SUPPORTS DE VECTEURS CAS NON SÉPARABLE DE SVM



MACHINE À SUPPORTS DE VECTEURS CAS NON SÉPARABLE DE SVM

- Il est nécessaire de modifier l'équation des contraintes afin de permettre à certains points de violer la marge ou d'être mal classés. La nouvelle contrainte s'écrit :

$$\left(\mathbf{w}^T x^{(i)} + w_0\right) y^{(i)} \geq 1 - \xi_i$$

- L'erreur de classification globale est définie comme la somme des variables de relaxation : $\sum_{i=1}^{N} \xi_i$
- En intégrant cette erreur, on obtient une fonction objectif combinée qui vise à la fois à maximiser la marge et à minimiser les erreurs de classification.

$$argmin_{\mathbf{w}} \left(\frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^{N} \xi_i \right)$$

$$\text{Sujet à : } \left(\mathbf{w}^T x^{(i)} + w_0\right) y^{(i)} \geq 1 - \xi_i, \forall i = 1, \ldots, N.$$

MACHINE À SUPPORTS DE VECTEURS CAS NON SÉPARABLE DE SVM

- La fonction globale à optimiser est donnée par:

$$Q = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$
$$-\sum_{i=1}^{N}\alpha_i\left[\left(\mathbf{w}^T x^{(i)} + w_0\right)y^{(i)} - 1 + \xi_i\right] - \sum_{i=1}^{N}\tau_i\xi_i$$

- Où τ_i sont les multiplicateurs de Lagrange qui permettent de garder les ξ_i positifs et C est une constante.
- En prenant les dérivées, comme précédemment, on obtient:

$$\begin{cases} \frac{\partial Q}{\partial \mathbf{w}} = 0. \\ \frac{\partial Q}{\partial w_0} = 0. \\ \frac{\partial Q}{\partial \xi_i} = 0. \end{cases} \Rightarrow \begin{cases} \mathbf{w} = \sum_{i=1}^{N}\alpha_i x^{(i)} y^{(i)} \\ \sum_{i=1}^{N}\alpha_i y^{(i)} = 0. \\ C - \alpha_i - \tau_i = 0 \end{cases}$$

MACHINE À SUPPORTS DE VECTEURS CAS NON SÉPARABLE DE SVM

- Puisqu'on a $\tau_i \geq 0$, alors $0 \leq \alpha_i \leq C$. On aura alors à maximiser sur les α_i :

$$Q = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\, y^{(i)}y^{(j)}x^{(i)^T}x^{(j)} + \sum_{i=1}^{N}\alpha_i$$

- Sujet à : $y^{(i)}\alpha_i = 0$ N $i=1$ et $0 \leq \alpha i \leq C, \forall i = 1,..,N$
- De la même manière que pour le cas séparable, les données bien classées (loin de la marge) auront leur $\alpha_i = 0$.
- Les vecteur à support auront $\alpha_i > 0$ et ils définissent le **w**.
- Les vecteurs à support ayant $\alpha_i < C$ seront sur la marge et auront $\xi_i = 0$. Ils satisfont $(\mathrm{W}^T x(i) + w_0)y^{(i)} = 1$. On peut les utiliser pour calculer w_0.
- Les vecteurs qui seront à l'intérieur de la marge ou mal classés auront $\alpha_i = C$.

MACHINE À SUPPORTS DE VECTEURS CAS NON SÉPARABLE DE SVM

- Lorsque la constance C est petite,

  la contrainte est ignorée ⇒ grande marge.

- Lorsque la constance C est grande,

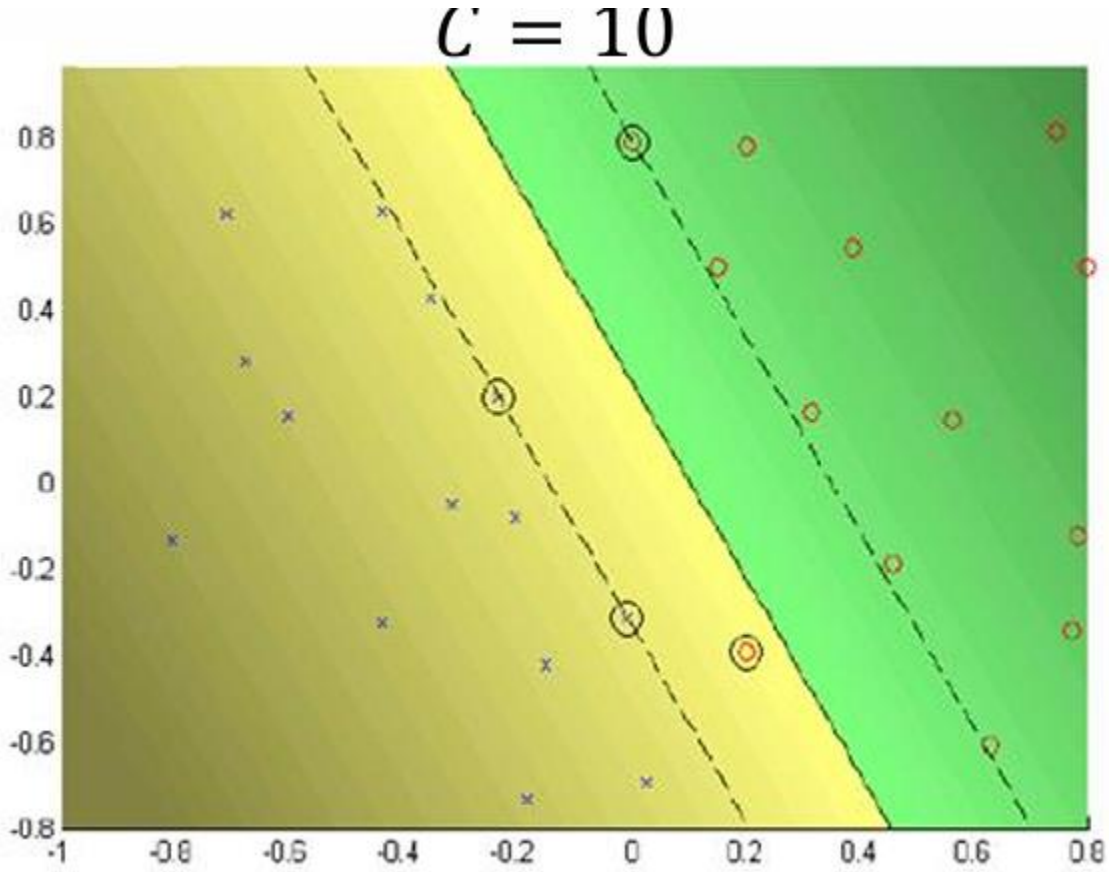
  la contrainte n'est pas ignorée ⇒ petite marge.

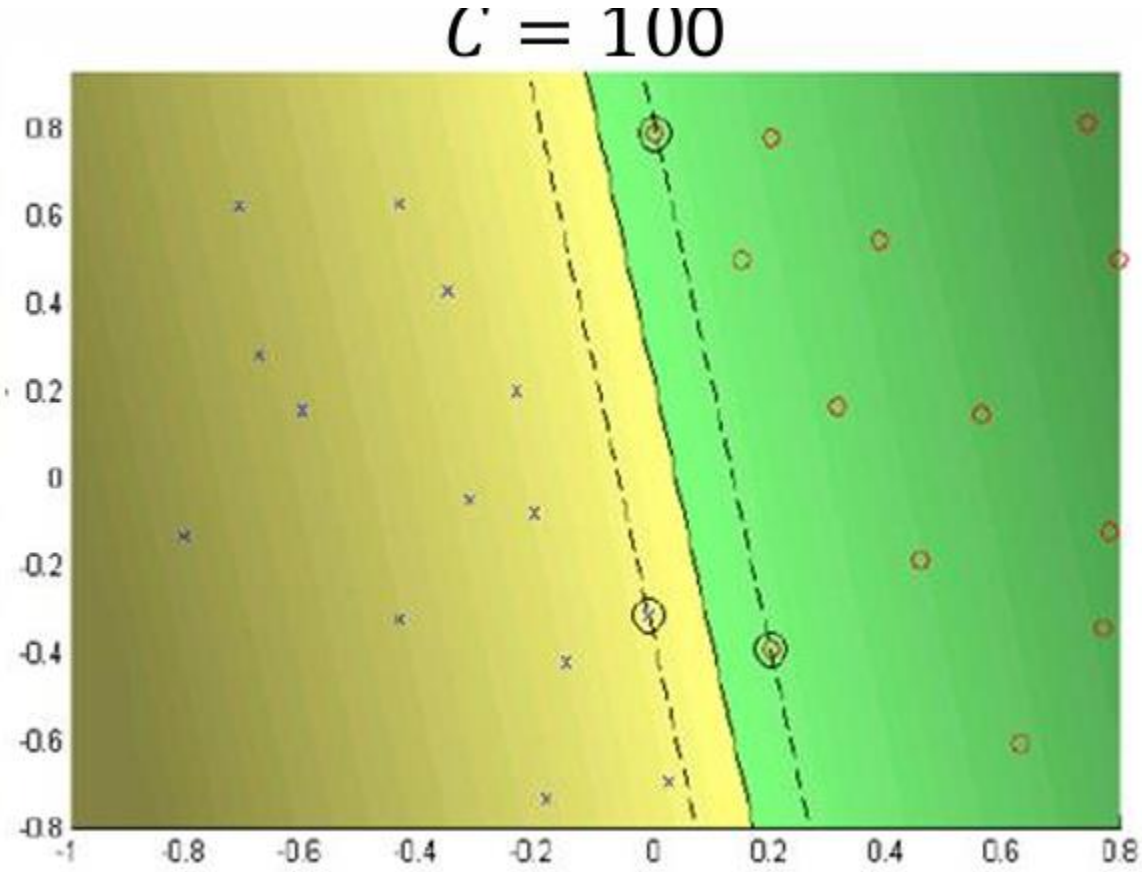
- Lorsque la constance C tend vers l'infini,

  la contrainte n'est pas ignorée ⇒ marge très étroite .

MACHINE À SUPPORTS DE VECTEURS CAS NON SÉPARABLE DE SVM

- Exemples:



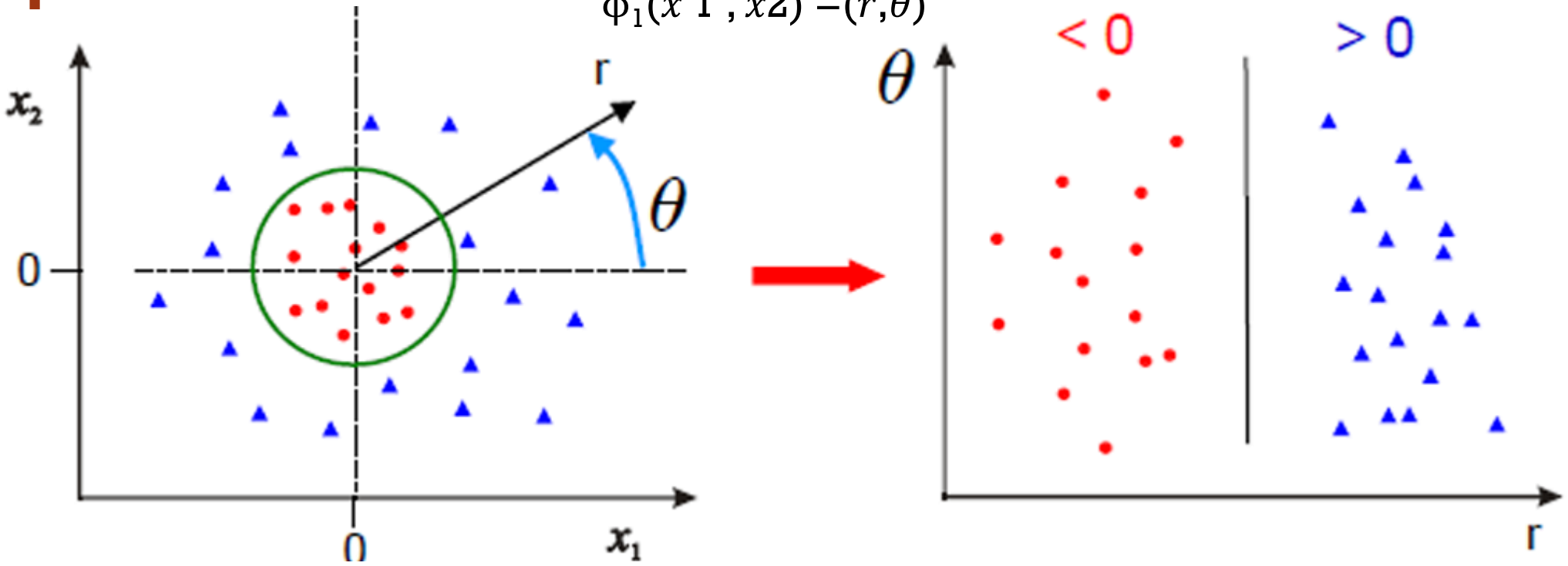


SVM ET FRONTIÈRES DE DÉCISION NON-LINÉAIRES

- Le SVM tel que présenté est capable de faire la classification linéaire des données et tolérer les erreurs de classification.
- Quand les classes de données ne sont pas linéairement séparables, la performance des SVM peut se détériorer.
- Si la frontière entre deux classes est non linéaire, on peut:
  - Soit utiliser directement un classificateur qui peut donner des frontières non-linéaires (ex. arbres, CB, etc.).
  - Soit transformer l'espace d'entrées X en un autre espace X ′ où les classes seront linéairement séparables.

- Exemple 1: Soit une fonction $\phi_1: \mathbb{R}_2 \rightarrow \mathbb{R}_2$
-

$$\phi_1(x\ 1\ , x2) = (r, \theta)$$



EXEMPLE 1:

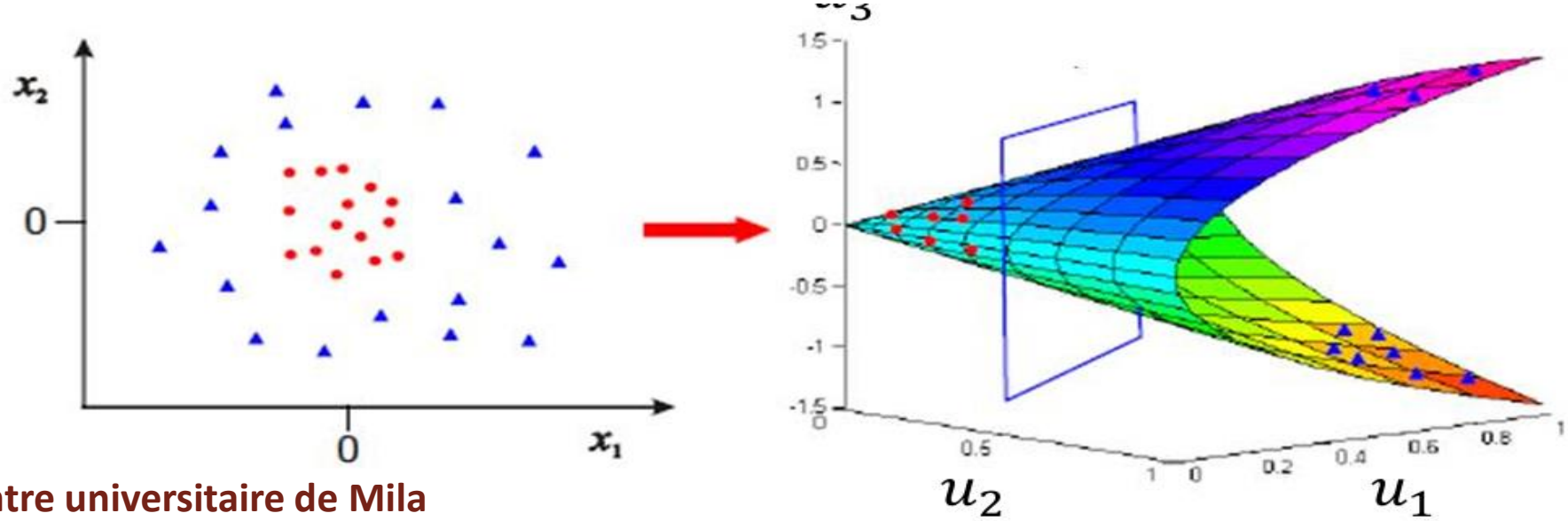
- Soit une fonction $\phi_1:\mathbb{R}_2 \rightarrow \mathbb{R}_2$
- $$\phi_1(x_1, x_2) = (r,\theta)$$
- Les deux coordonnées cartésiennes x_1 et x_2 permettent de calculer la première coordonnée polaire r par :
- $$r = \sqrt{x^2 + y^2}$$
- Pour obtenir θ dans l'intervalle $]-\pi,\pi[$, on peut utiliser la formule suivante :

$$\theta = 2arctan\left(\frac{y}{x + \sqrt{x^2 + y^2}}\right)$$

EXEMPLE 2

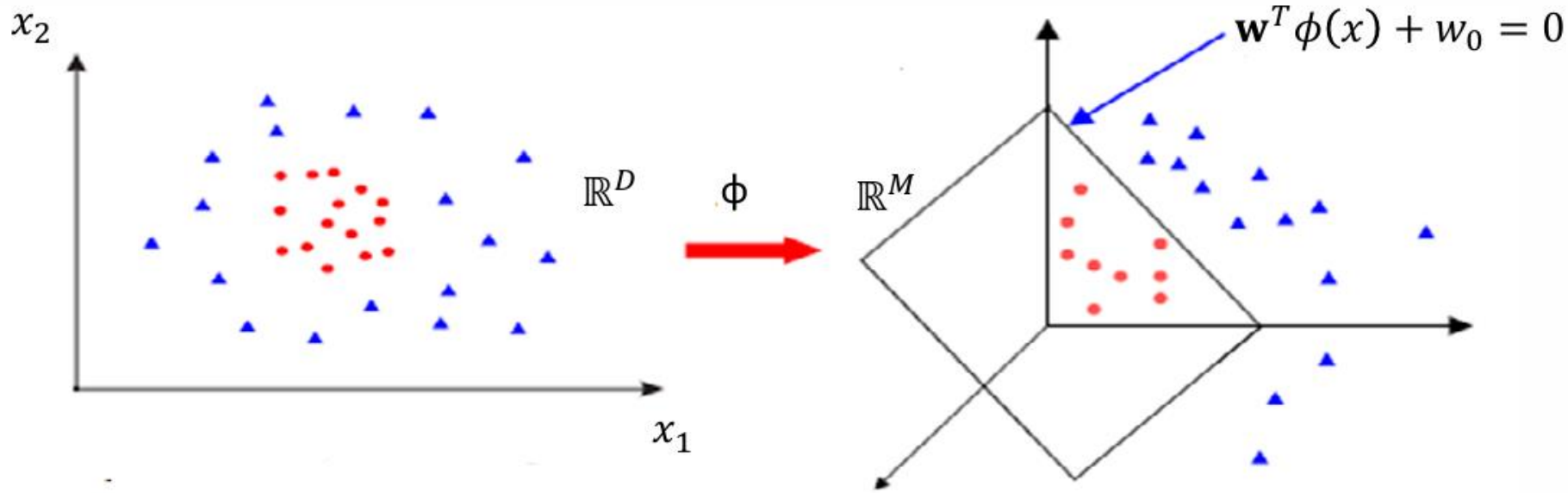
Soit une fonction $\phi_2: \mathbb{R}^2 \rightarrow \mathbb{R}^3$

$$\phi_2(x_1, x_2) = ({x_1}^2, {x_2}^2, \sqrt{2}x_1x_2)$$
$$= (u_1, u_2, u_3)$$



SVM NON-LINÉAIRE

- Plus généralement, soit la fonction: $\phi : \mathbb{R}^D \rightarrow \mathbb{R}^M$.
- Supposons que les classes dans l'espace $\mathbb{R}^M$ sont linéairement séparables. • Trouver un classificateur linéaire: $f(x) = \mathbf{w}^T\phi(x) + w_0$.



L'ASTUCE DES NOYAUX

- Plus généralement, SVM dans l'espace $\mathbb{R}^D$ maximise:

$$Q = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\, y^{(i)}y^{(j)}x^{(i)^T}x^{(j)} + \sum_{i=1}^{N}\alpha_i$$

- Dans l'espace $\mathbb{R}^M$, SVM maximise:

$$Q = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\, y^{(i)}y^{(j)}\underbrace{\phi^T\left(x^{(i)}\right)\phi\left(x^{(j)}\right)}_{k\left(x^{(i)},x^{(j)}\right)} + \sum_{i=1}^{N}\alpha_i$$

☞**Astuce du noyau!**

- Si on démontre que $\boldsymbol{k\left(x^{(i)},x^{(j)}\right) = \phi^T\left(x^{(i)}\right)\phi\left(x^{(j)}\right)}$, donc on peut utiliser le noyau $\boldsymbol{k\left(x^{(i)},x^{(j)}\right)}$ dans notre expression à la place de tout produit scalaire intérieur dans l'espace d'origine.

L'ASTUCE DES NOYAUX

- La caractéristique la plus importante du SVM est d'éviter de faire explicitement la transformation de $\mathbb{R}^D$ vers $\mathbb{R}^M$, et remplacer le produit scalaire par le noyau $\boldsymbol{k}$.
- Il existe de nombreuses fonctions de noyau prêtes à utiliser (chacune équivalente à un produit scalaire après une certaine transformation). :

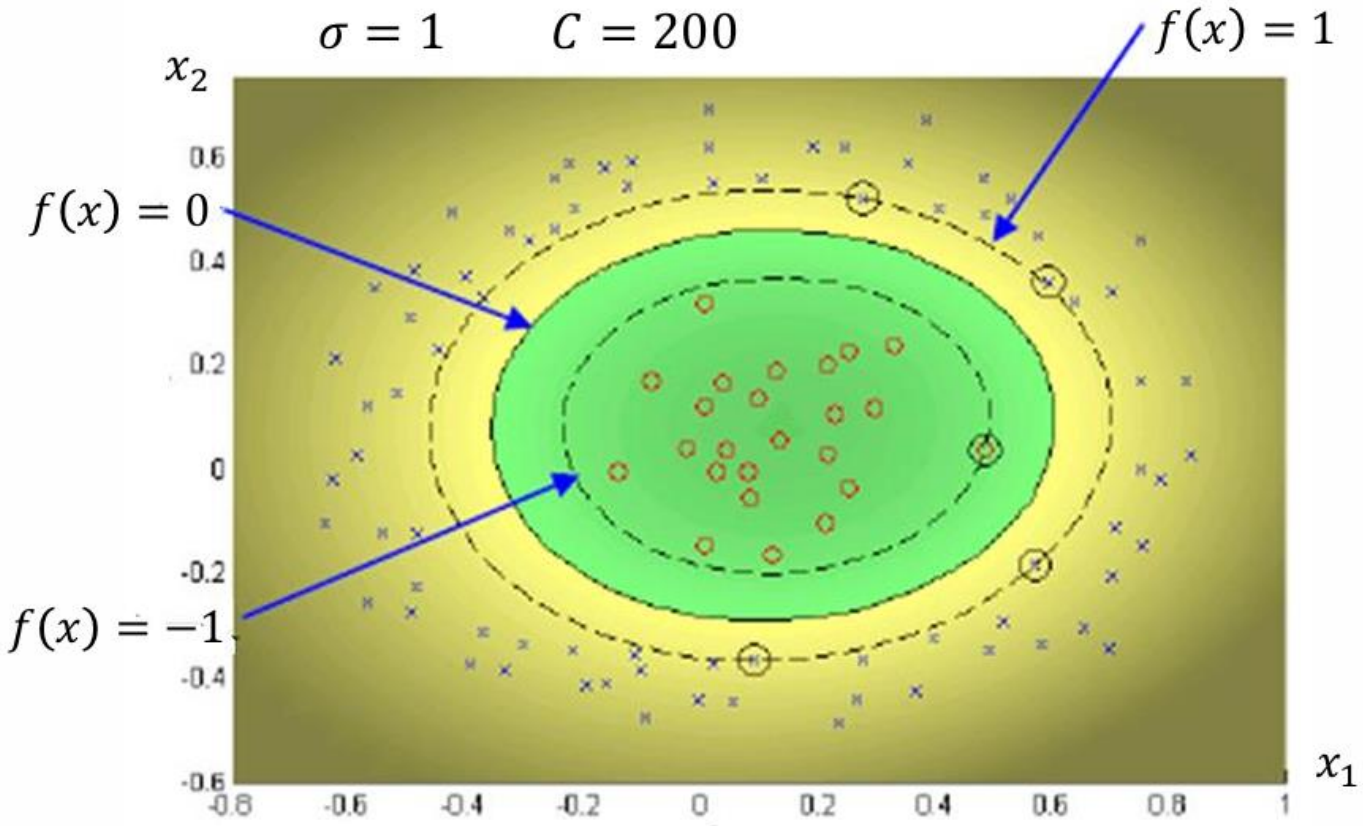
Linéaire: $$k\left(x^{(i)}, x^{(j)}\right) = x^{(i)^T} x^{(j)}$$

Polynomiale: $$k\left(x^{(i)}, x^{(j)}\right) = \left(1 + x^{(i)^T} x^{(j)}\right)^m \quad m > 0$$

Gaussien: $$k\left(x^{(i)}, x^{(j)}\right) = \exp\left(-\left\|x^{(i)} - x^{(j)}\right\|^2 / 2\sigma^2\right)$$

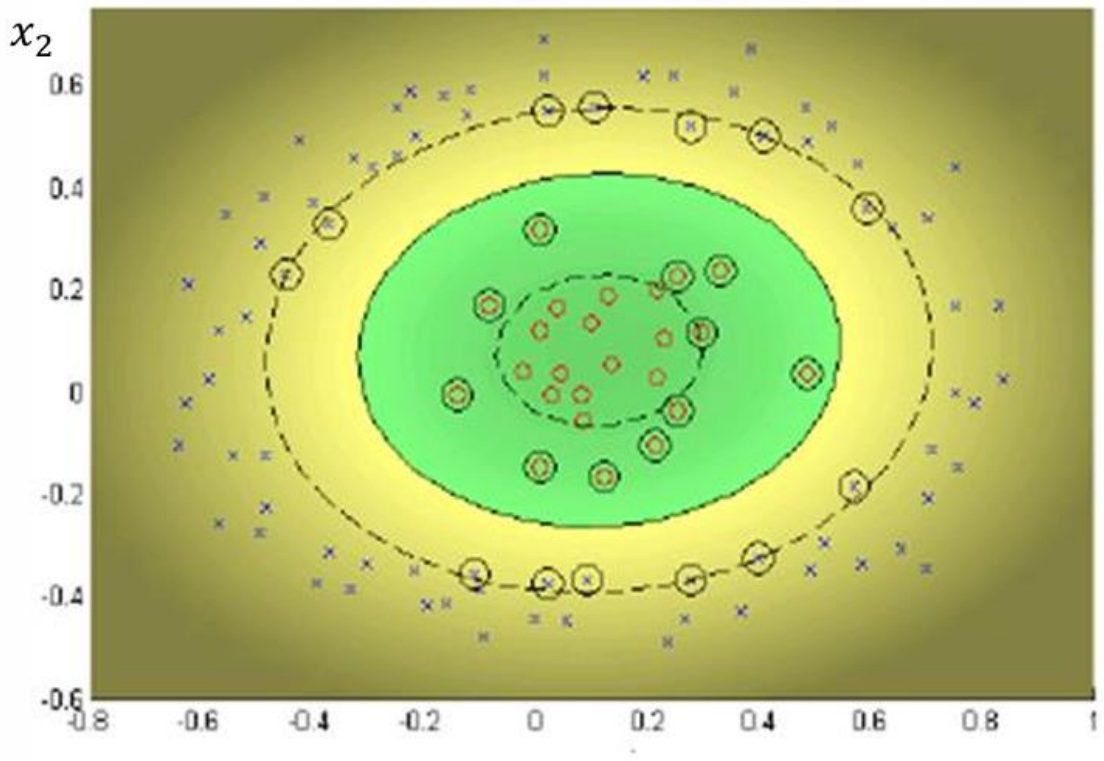
EXEMPLE

$$\boldsymbol{k}\left(x^{(i)}, x^{(j)}\right) = \exp\left(-\left\|x^{(i)} - x^{(j)}\right\|^2 / 2\sigma^2\right)$$

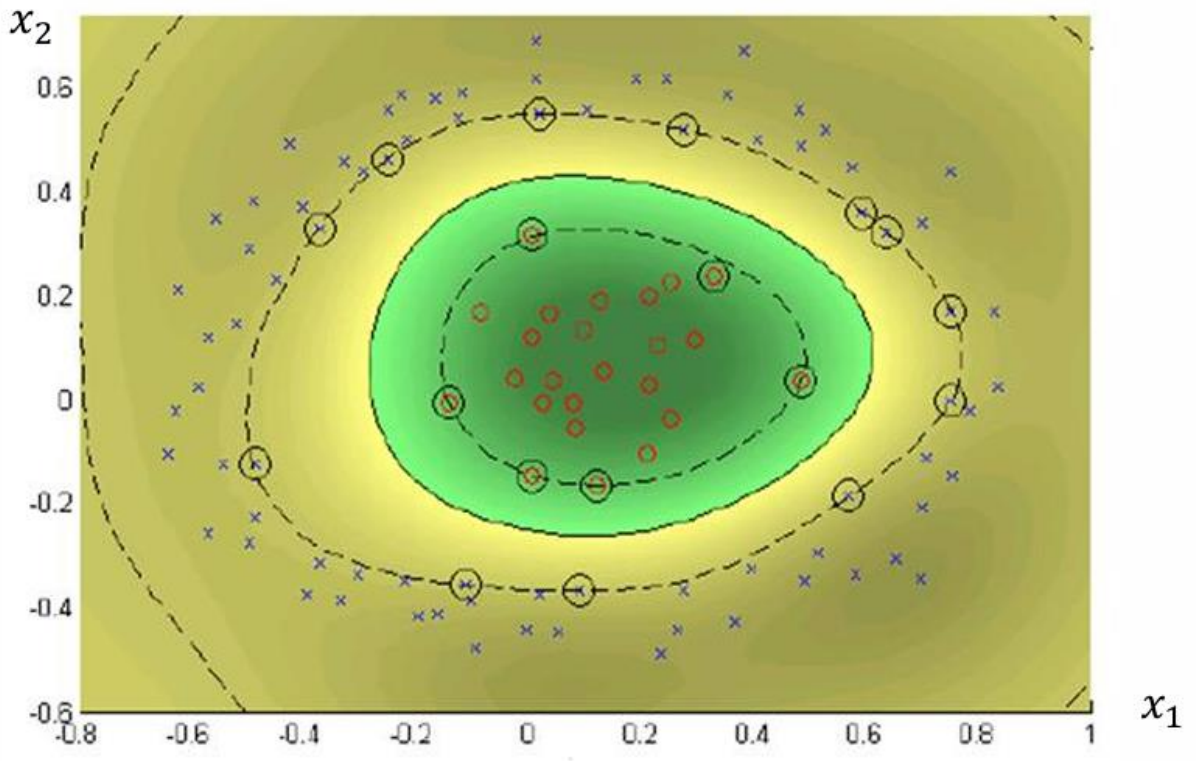


EXEMPLE

$\sigma = 1 \quad C = 10$



$\sigma = 0.25 \quad C = 200$



FIN DE CHAPITRE 06