Machine à vecteurs de supports

Machine à vecteurs de supports (SVM)

CAS DES DONNÉES SÉPARABLES

Modèles linéaires pour la classification

- Une **fonction discriminante** a le rôle de prendre une entrée **x** et lui assigner une classe parmi K classes existantes.

- **Un classificateur linéaire** utilise une **frontière de décision linéaire** pour assigner les classes aux données.

- Si D est la dimension de **x**, alors:

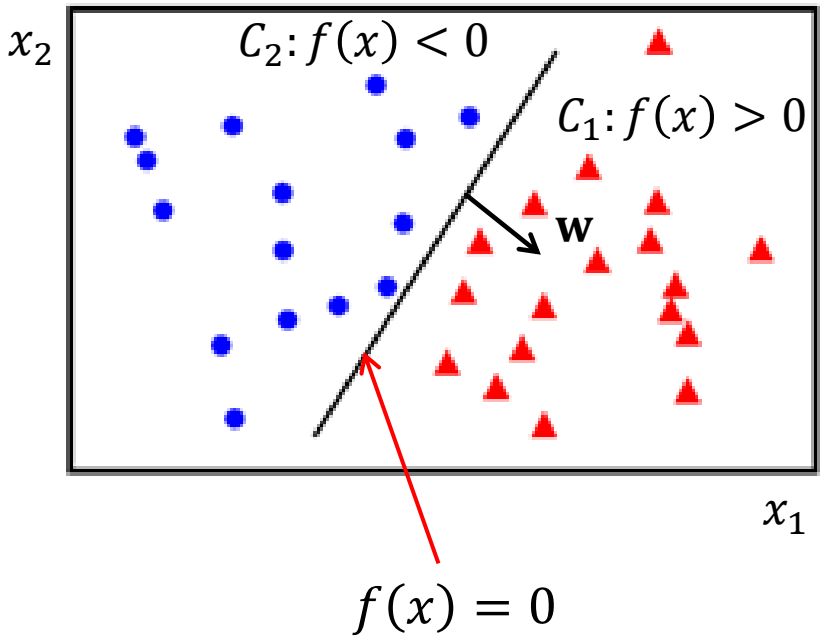
  - Pour D = 2, la frontière sera **une droite**.
  - Pour D = 3, la frontière sera **un plan**.
  - Pour D > 3, la frontière sera **un hyperplan**.

Modèles linéaires pour la classification

- Un classificateur linéaire a la forme suivante (ex. $D = 2$):

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

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

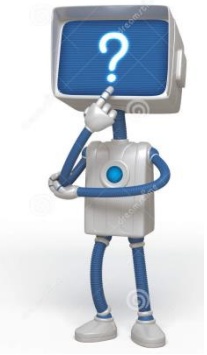


- En 2 dimensions ($D = 2$):

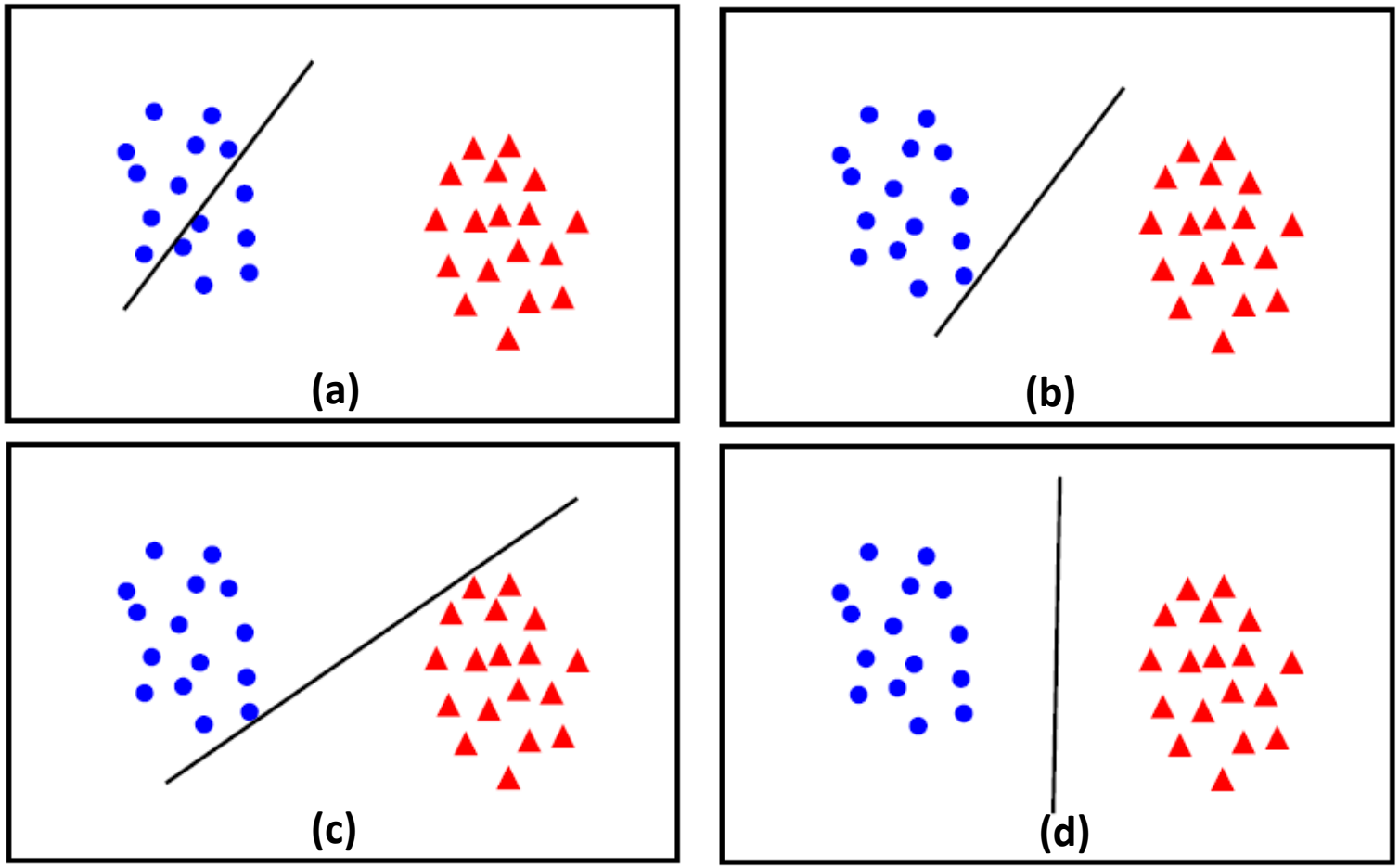
  ☞ $\mathbf{w}$ est le **vecteur normal** à la frontière de décision (ligne).

  ☞ w_0 est appelé **le biais** (ou **l'intercepte**)

Modèles linéaires pour la classification



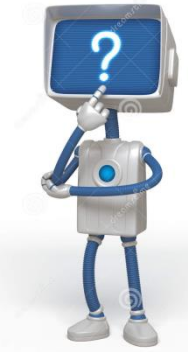
- Quelle est **la meilleure** frontière de décision?

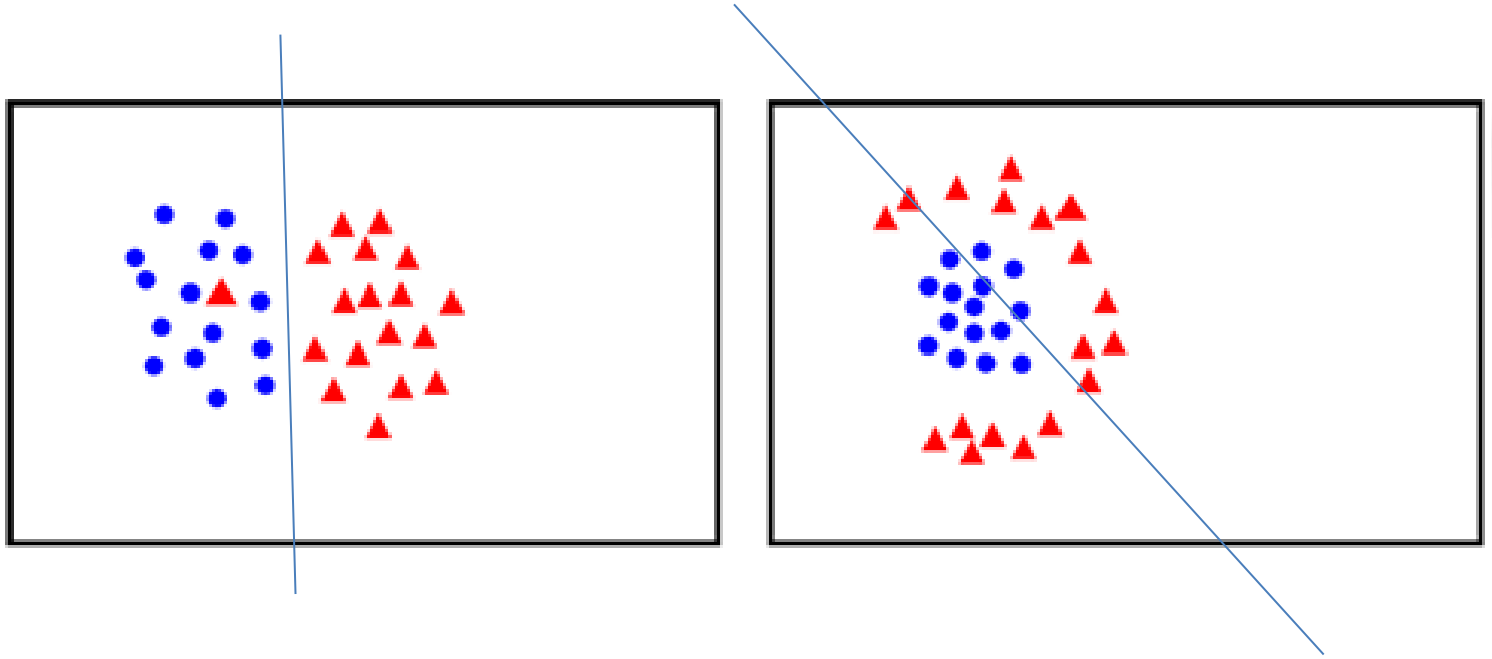


Modèles linéaires pour la classification

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

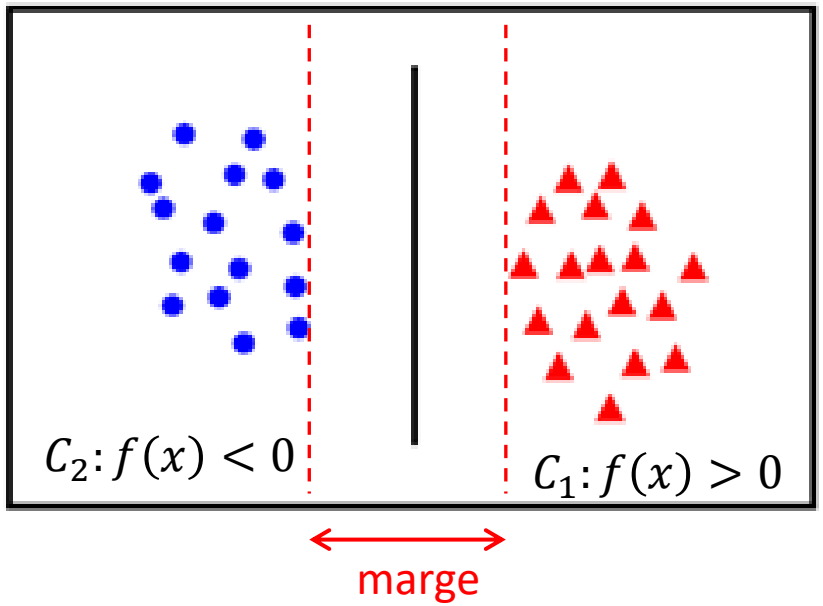
- Et si les classes ne sont pas **linéairement séparables**?





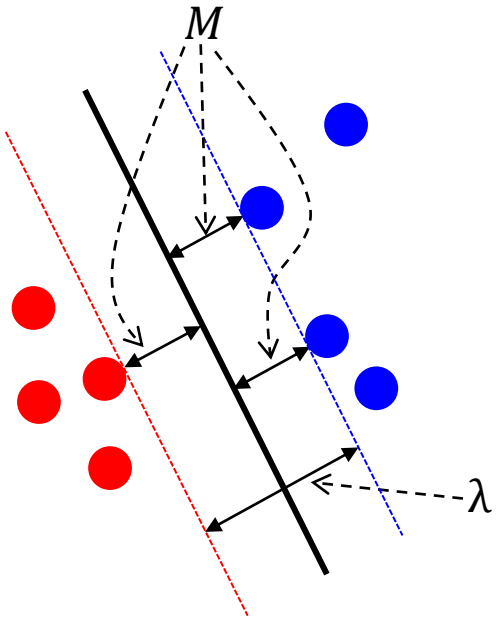
Modèles linéaires pour la classification

- La solution offrant **le maximum de marge** d'erreur peut être une bonne alternative pour une meilleure **généralisation**.
- Dans la suite, les classes C_1 et C_2 sont appelées les classes des **positifs** et des **négatifs**, respectivement.
- La cible est codée da la manière suivante: $y \in \{+1, -1\}$.

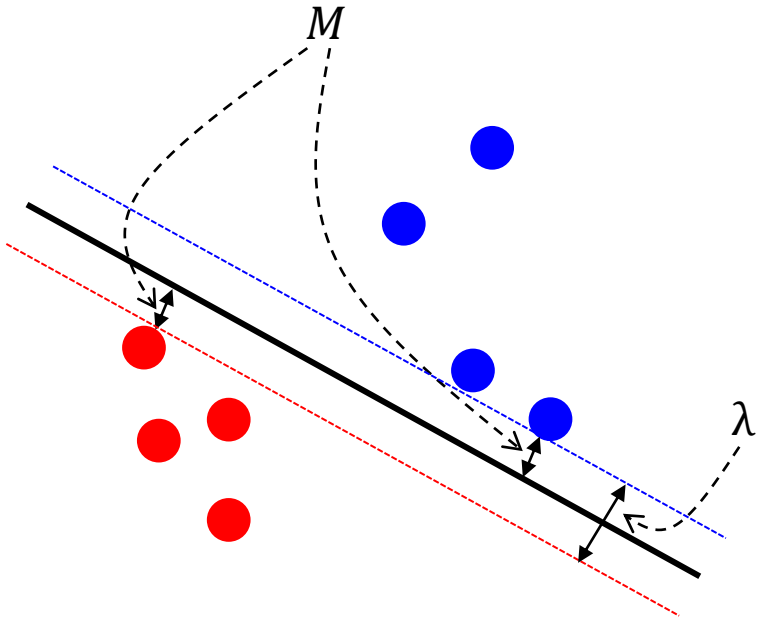


Machine à supports de vecteurs

- Soit M la distance perpendiculaire entre la frontière de décision et **les données les plus proches** de chaque classe.
- La **marge** est indiquée par $\lambda = 2 \times M$.

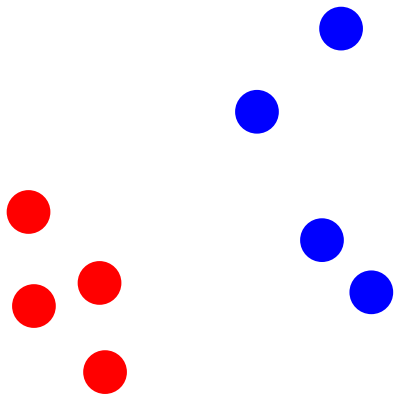


(a) La frontière de décision qui maximise la marge



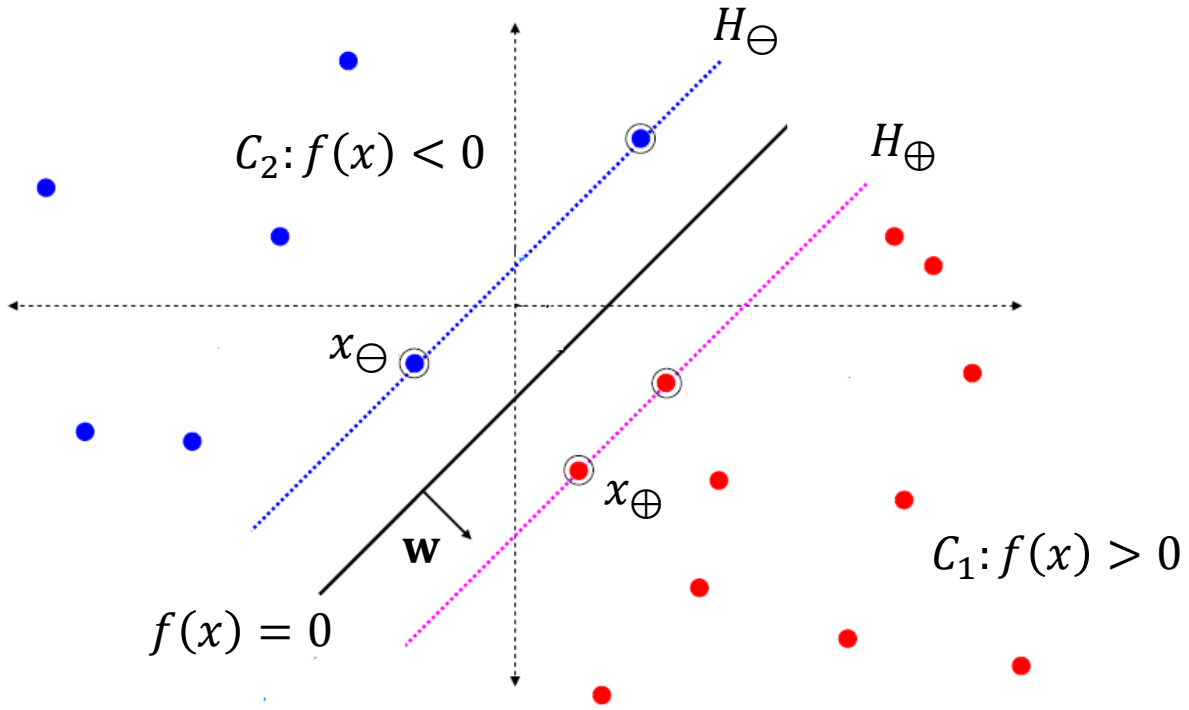
(b) Une frontière de décision non-optimale

Machine à supports de vecteurs



Machine à supports de vecteurs

- Pour un cas général, soit $H_{\oplus}$ et $H_{\ominus}$ les hyperplans contenant **les données les plus proches** dans la classe C_1 et C_2, qui sont $x_{\oplus}$ et $x_{\ominus}$, respectivement.



Machine à supports de vecteurs

- On peut alors choisir des coefficients $\mathbf{w}$ et w_0 de sorte que:

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

- La valeur de la marge est alors donnée géométriquement par:

$$\lambda = \frac{\mathbf{w}^T}{\|\mathbf{w}\|}\left(x_{\oplus} - x_{\ominus}\right) = \frac{\mathbf{w}^T\left(x_{\oplus} - x_{\ominus}\right)}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}$$

- Pour **maximiser la marge**, il faut **minimiser** la valeur de $\|\mathbf{w}\|$.

Note: $\|\mathbf{w}\|$ représente la norme du vecteur $\mathbf{w}$.

Machine à supports de vecteurs

- La valeur de la marge peut être déduite comme suit:

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

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

Note: $\|\mathbf{w}\|$ représente la norme du vecteur $\mathbf{w}$.

Machine à supports de vecteurs (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:

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

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

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

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

Note: $\|\mathbf{w}\|$ représente la norme du vecteur $\mathbf{w}$.

Machine à supports de vecteurs (SVM)

- Le problème d'optimisation est devenu:

$$argmin_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 \text{ sujet à : } \left(\mathbf{w}^T x^{(i)} + w_0\right) y^{(i)} > 1$$

- Pour obtenir $\mathbf{w}$ et w_0, nous devons intégrer les contraintes dans la fonction objective via un ensemble de **multiplicateurs de Lagrange**.
- Les **multiplicateurs de Lagrange** ajoutent un nouveau terme pour chaque contrainte de sorte que l'optimum de la nouvelle fonction corresponde à l'optimum du problème d'origine.

Note: $\|\mathbf{w}\|$ représente la norme du vecteur $\mathbf{w}$.

Machine à supports de vecteurs

- **La fonction objective** à optimiser combinant **les deux contraintes** est donnée par:

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

- Les $\alpha_i \geq 0$ sont des **multiplicateurs de Lagrange**. La fonction Q doit être **minimisée** sur $\mathbf{w}$ et w_0 et **maximisée** sur les α_i. En utilisant les dérivées, on aura:

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

Note: $\|\mathbf{w}\|$ représente la norme du vecteur $\mathbf{w}$.

Machine à supports de vecteurs

- En remplaçant la valeur de $\mathbf{w}$ dans **la fonction objective** Q, on obtient:

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

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

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

- Cette fonction sera **maximisée** en fonction des α_i avec la contrainte: $\sum_{i=1}^{N}\alpha_i y^{(i)} = 0$ et $\alpha_i >= 0, \forall i = 1, .., N$

Machine à supports de vecteurs

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

- 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.
- Par exemple, on peut utiliser la fonction MATLAB **quadprog** pour le résoudre.

Machine à supports de vecteurs

- Une fois les valeurs optimale des α_i obtenues, la plupart d'elles vont s'annuler $\alpha_i \to 0$ 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 $\mathbf{w}$ est réécrit comme une **somme pondérée** des **vecteurs de supports**. Ces **vecteurs se trouvent sur la marge** et donc satisfont l'équation:

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

Machine à supports de vecteurs

- 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.

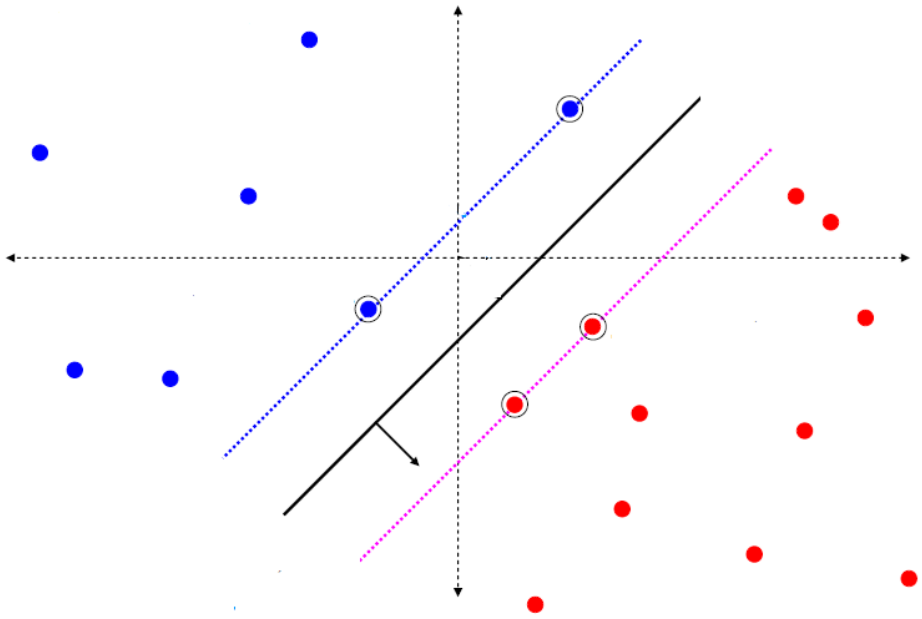
- La plupart des α_i sont nuls et on aura pour leurs données $\left(\mathbf{w}^T x^{(i)} + w_0\right) y^{(i)} > 1$.

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

Machine à supports de vecteurs

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

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

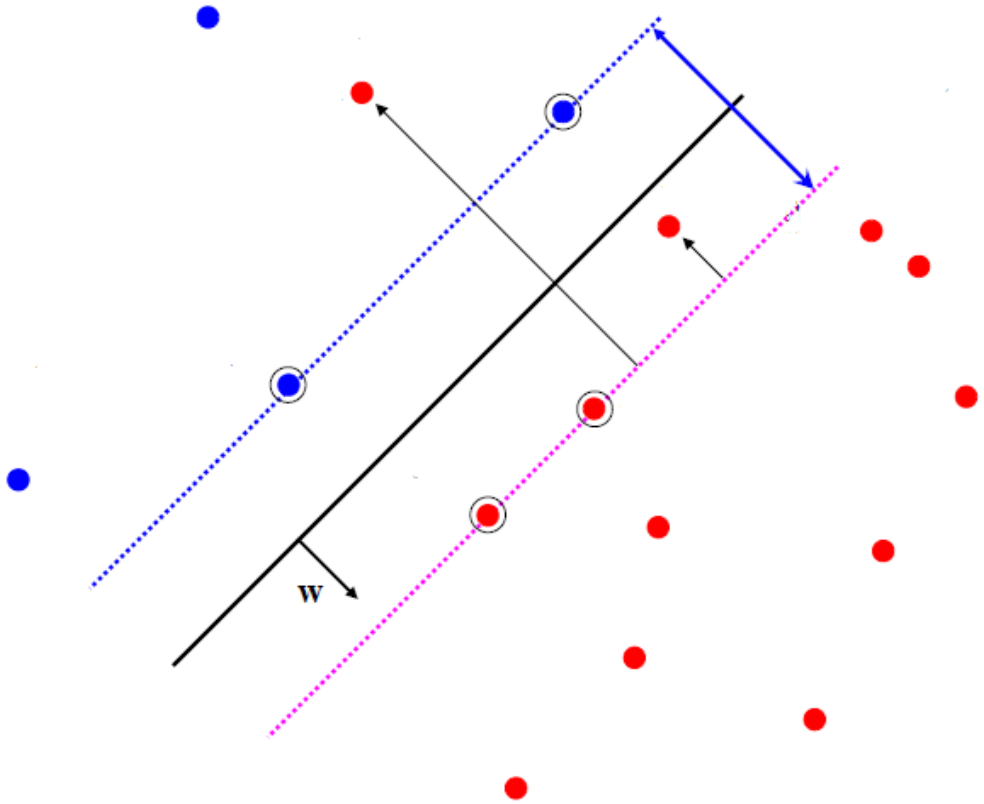


Machine à supports de vecteurs

CAS NON SÉPARABLE DE SVM

Cas non séparable de SVM

Parfois les deux classes C_1 et C_2 **ne sont pas linéairement séparables**, comme dans l'exemple suivant:



Cas non séparable de SVM

- Si les classes C_1 et C_2 ne sont pas **linéairement séparables**, l'algorithme SVM présenté auparavant ne fonctionnera pas.

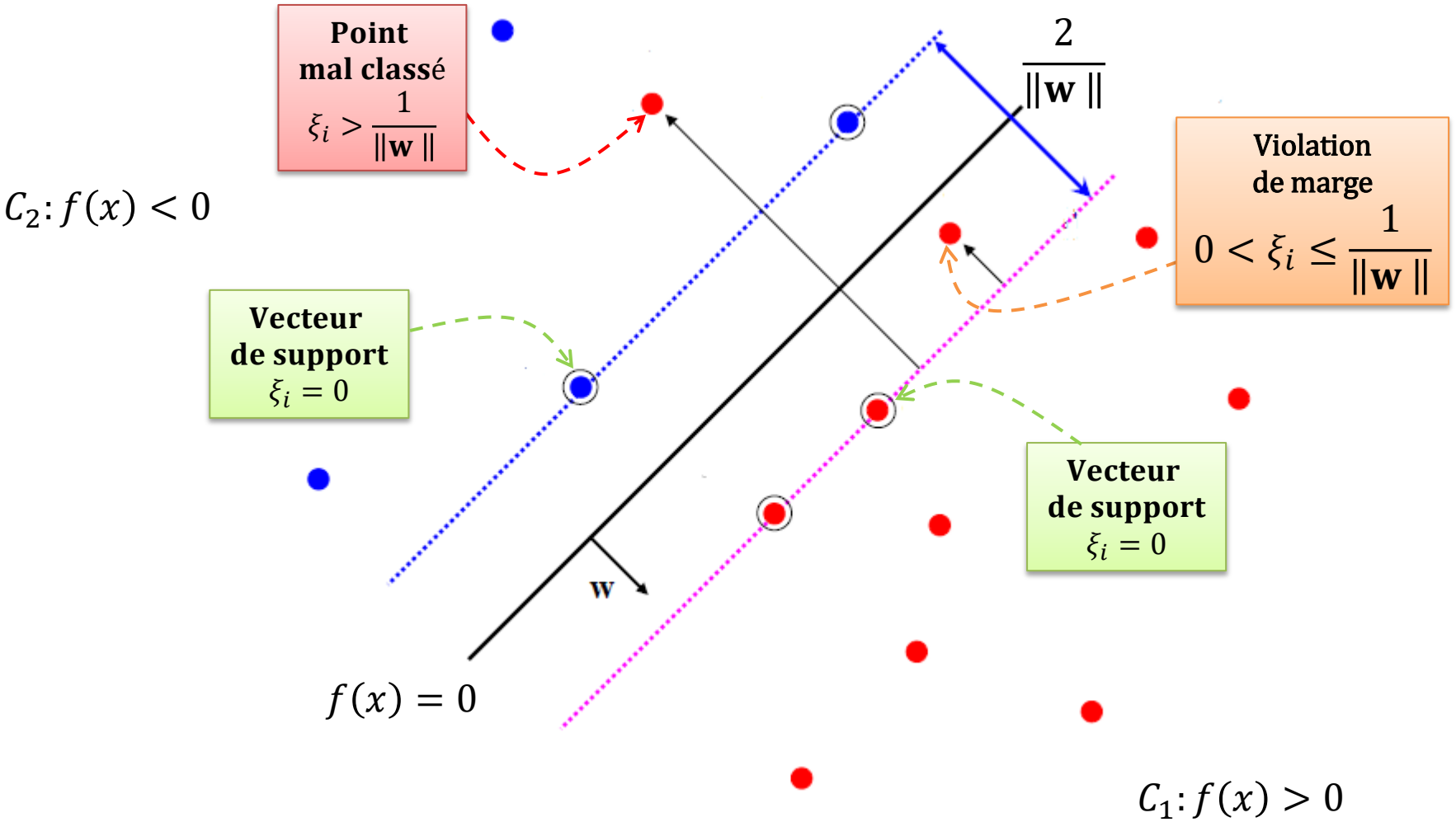
- On peut forcer l'algorithme **à tolérer des erreurs** de classification (**les moindres possible**).

- Pour chaque donnée $x^{(i)}$, on crée une variable $\xi_i \geq 0$ qui mesure **la déviation** de la donnée par rapport à la marge:

  ☞ Si $0 < \xi_i \leq \frac{1}{\|\mathbf{w}\|}$, alors le point est entre la frontière de décision et la marge (**violation de marge**).

  ☞ Si $\xi_i > \frac{1}{\|\mathbf{w}\|}$, le point est alors **mal classé**.

Cas non séparable de SVM



Cas non séparable de SVM

- On doit adapter l'équation des contraintes afin qu'elle admette la possibilité que certains points violent la marge ou affectés à la mauvaise classe. La contrainte devient

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

- L'erreur de classification globale est définie par: $\sum_{i=1}^{N} \xi_i$.
- En ajoutant cette erreur, on aura une pénalité combinée qui, d'une part, **maximise la marge** et, d'autre part, **minimise l'erreur** de classification:

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

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

Note: $\|\mathbf{w}\|$ représente la norme du vecteur $\mathbf{w}$.

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:

Cas non séparable de SVM

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

- 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 : $\sum_{i=1}^{N} y^{(i)} \alpha_i = 0$ et $0 \leq \alpha_i \leq C, \forall i = 1, .., N$

Cas non séparable de SVM

- 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 $\mathbf{w}$.
- Les vecteurs à support ayant $\alpha_i < C$ seront **sur la marge** et auront $\xi_i = 0$. Ils satisfont $\left(\mathbf{w}^T x^{(i)} + w_0\right) y^{(i)} = 1$. On peut les utiliser pour calculer w_0.
- Les vecteurs qui seront **à l'intérieur de la marge ou mal classés** auront $\alpha_i = C$.

Cas non séparable de SVM

- Lorsque la constance C est petite, la contrainte est ignorée

  $\Rightarrow$ **grande marge**.

- Lorsque la constance C est grande, la contrainte n'est pas ignorée

  $\Rightarrow$ **petite marge**.

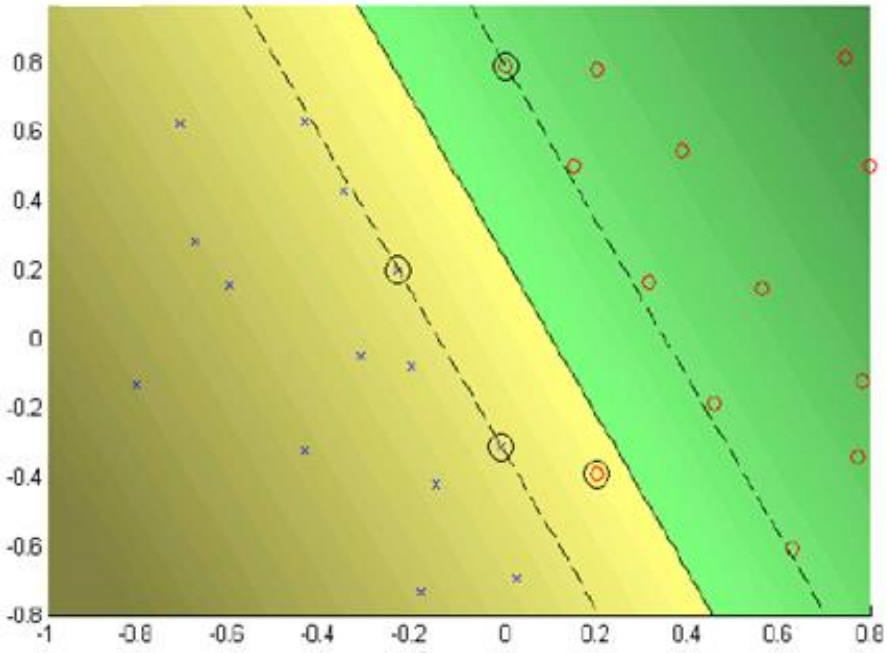
- Lorsque la constance C tend vers l'infini, la contrainte n'est pas ignorée

  $\Rightarrow$ **marge très étroite** .

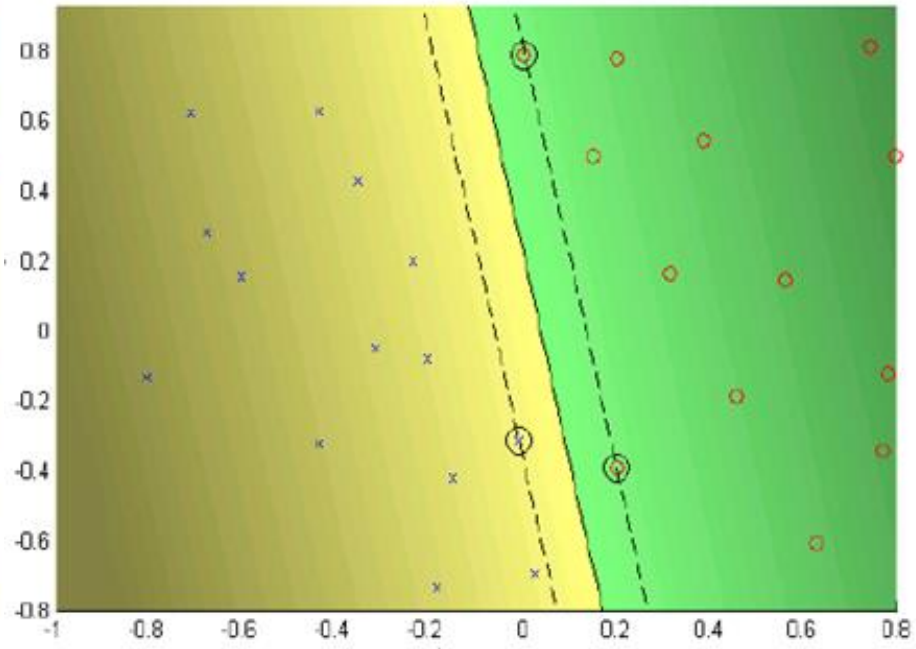
Cas non séparable de SVM

Exemples:

$C = 10$



$C = 100$



Machine à supports de vecteurs

FRONTIÈRES DE DÉCISION NON-LINÉAIRES

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:

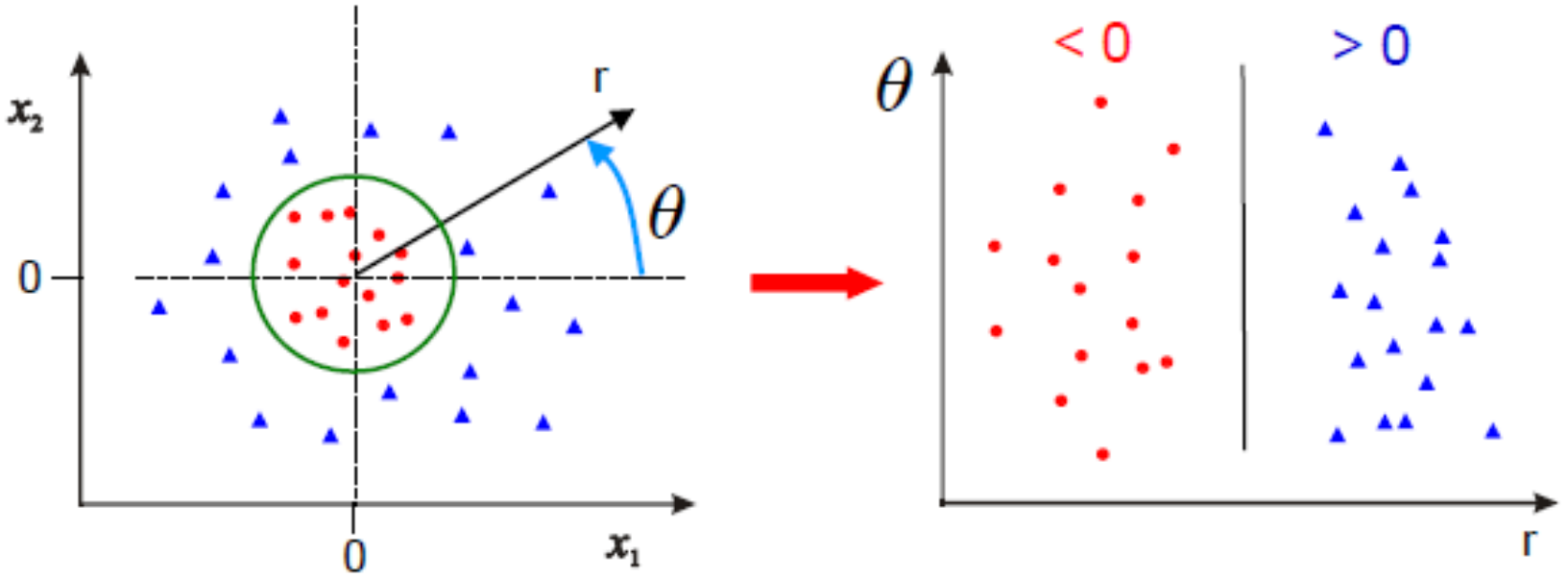
  ☞ Soit utiliser directement un classificateur qui peut donner **des frontières non-linéaires** (ex. arbres, CB, etc.).

  ☞ Soit **transformer l'espace d'entrées** $\mathcal{X}$ en un autre espace $\mathcal{X}'$ où les classes seront linéairement séparables.

SVM non-linéaire

Exemple 1: Soit une fonction $\phi_1 : \mathbb{R}^2 \rightarrow \mathbb{R}^2$

$$\phi_1(x_1, x_2) = (r, \theta)$$



(a) Espace de coordonnées cartésiennes. Un point est définit par deux valeur x_1 et x_2

(b) Espace de coordonnées polaires. Un point est définit par un angle θ et une distance r

SVM non-linéaire

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 :

$$r = \sqrt{{x_1}^2 + {x_2}^2}$$

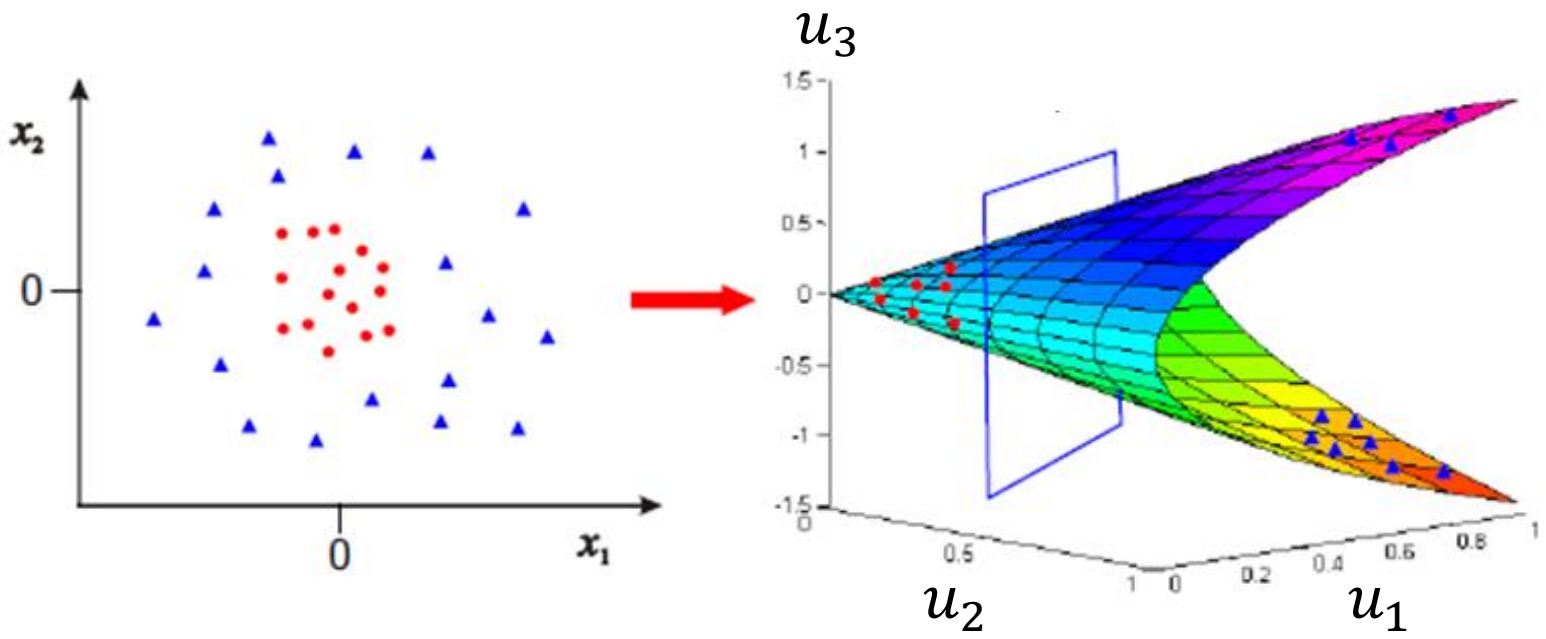
Pour obtenir θ dans l'intervalle $]-\pi, \pi[$, on peut utiliser la formule suivante :

$$\theta = 2arctan\left(\frac{x_2}{x_1 + \sqrt{{x_1}^2 + {x_2}^2}}\right)$$

SVM non-linéaire

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

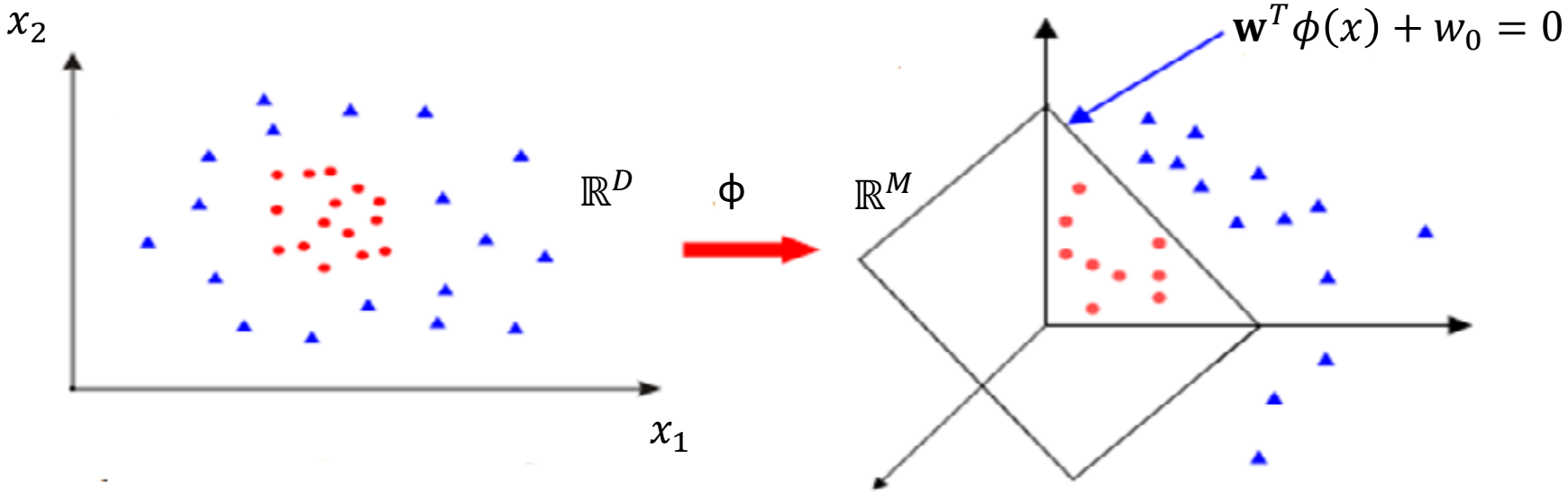
$$\phi_2(x_1, x_2) = ({x_1}^2, {x_2}^2, \sqrt{2}x_1x_2)$$
$$= (u_1, u_2, u_3)$$



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:

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

☞**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). :

- Voici les trois noyaux les plus utilisés:

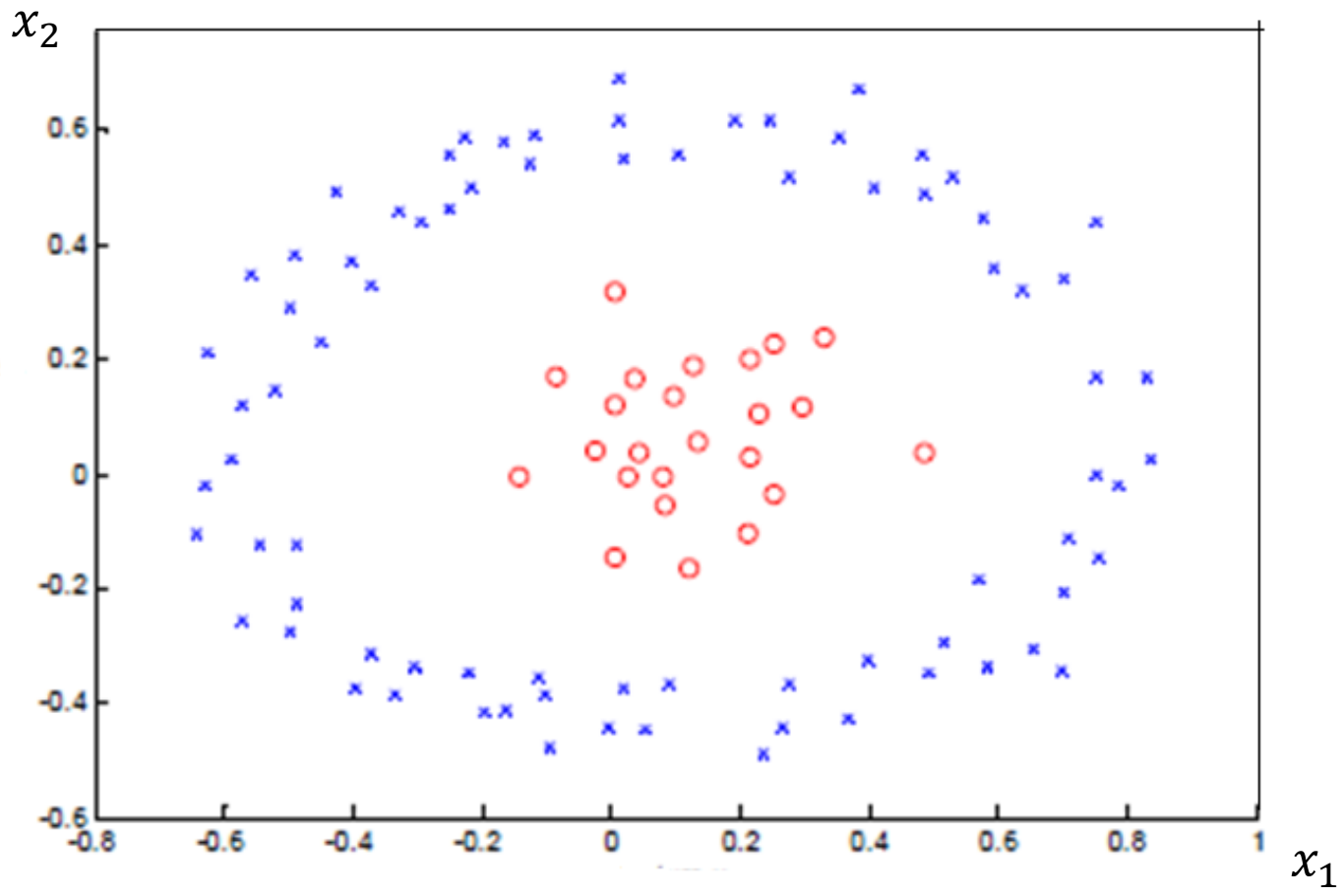
Linéaire: $\boldsymbol{k}\left(x^{(i)}, x^{(j)}\right) = {x^{(i)}}^T x^{(j)}$

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

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

L’astuce des noyaux

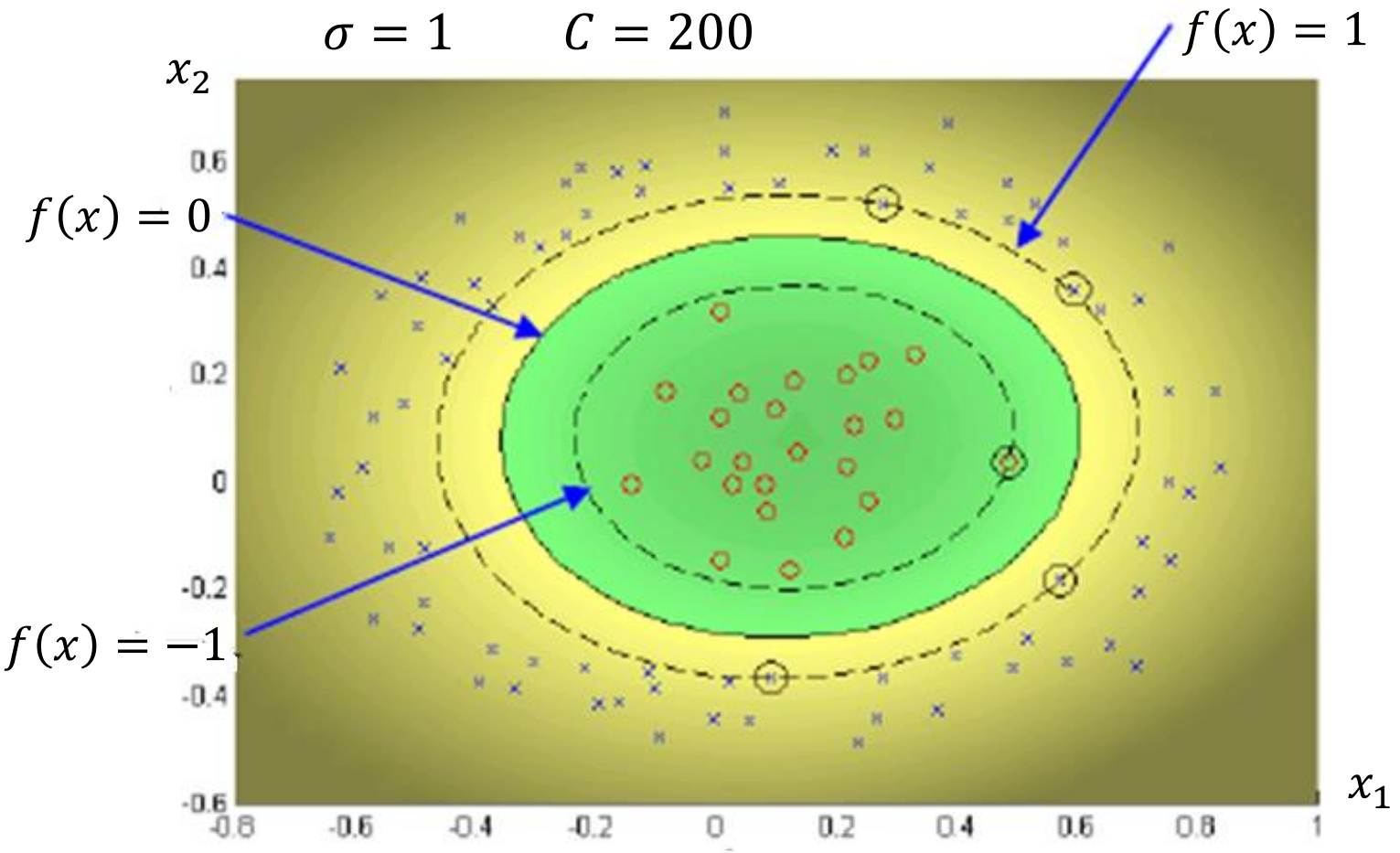
Exemple de classification:



L'astuce des noyaux

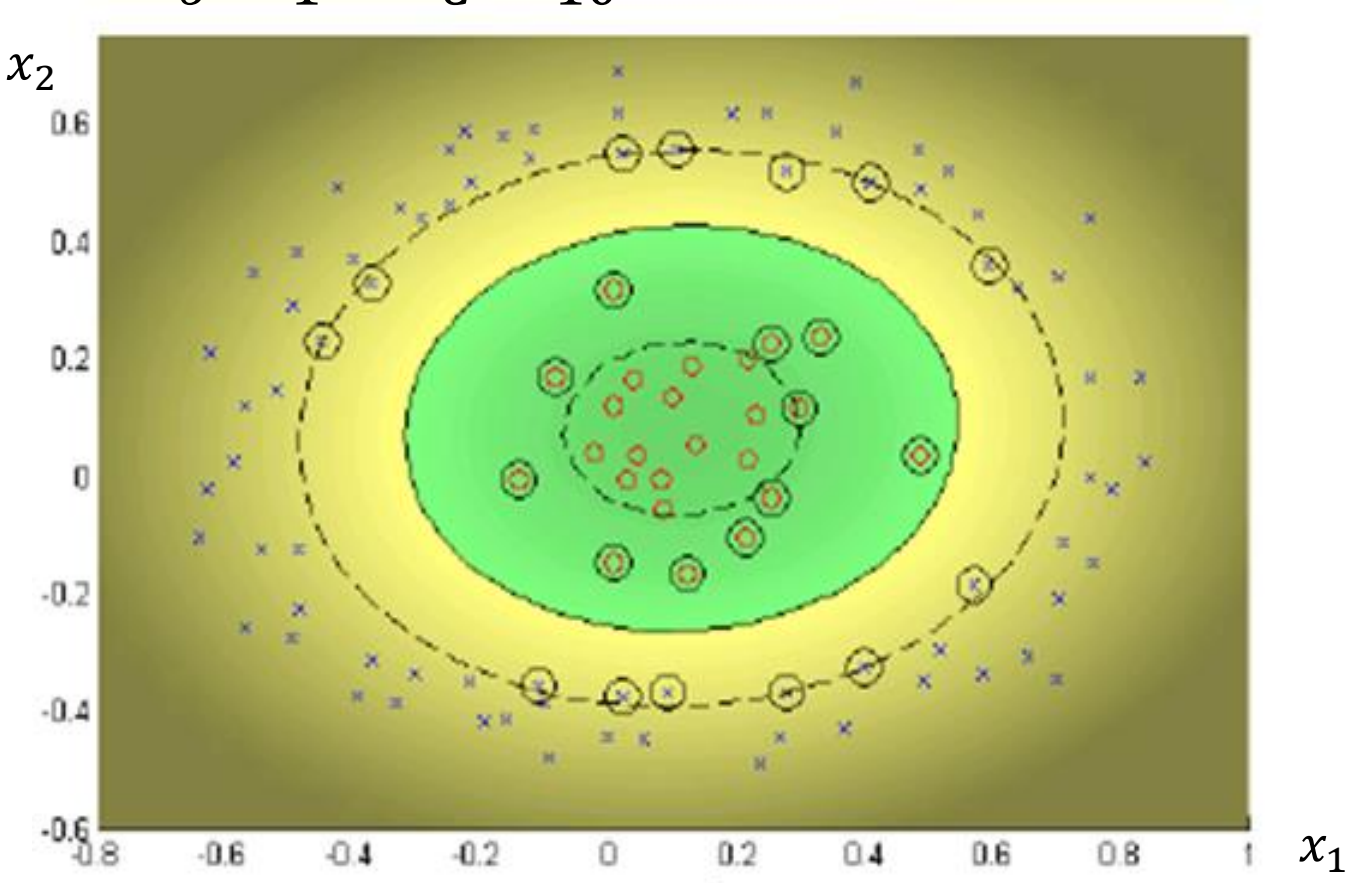
$$\boldsymbol{k}\left(x^{(i)}, x^{(j)}\right) = \exp\left(-\left\|x^{(i)} - x^{(j)}\right\|^2 / 2\sigma^2\right)$$

$\sigma = 1$ $\quad C = 200$



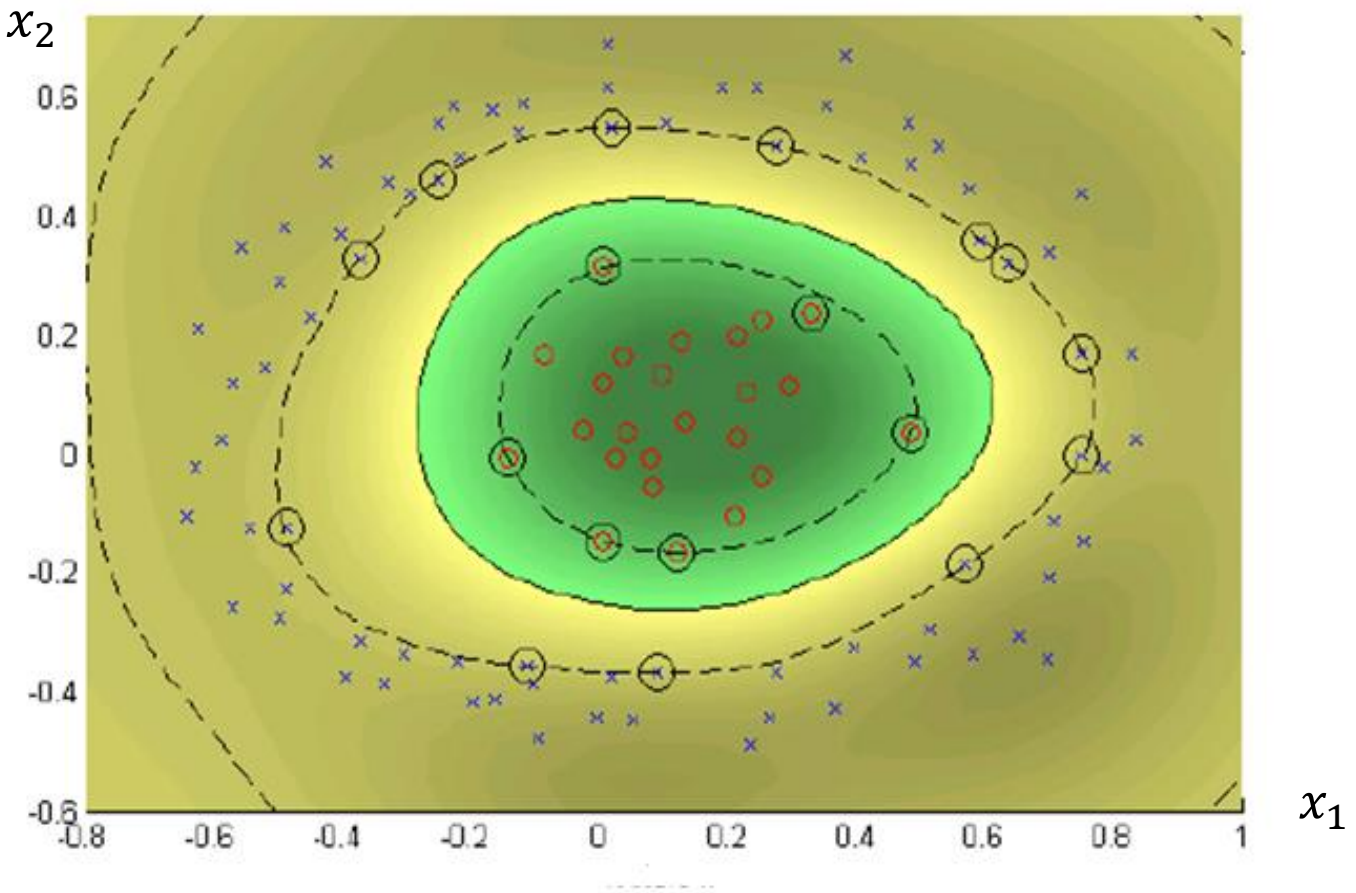
L’astuce des noyaux

$\sigma = 1 \qquad C = 10$



L’astuce des noyaux

$\sigma = 0.25 \quad C = 200$



Références

1. M. S. Allili. Techniques d'apprentissage automatique (Cours de 2e cycle). Université du Québec en Outaouais (UQO), Québec, Canada. Hivers 2015.
2. S. Rogers et M Girolami. A first Course in machine learning, CRC press, 2012.
3. C. Bishop. Pattern Recognition and Machine learning. Springer 2006.
4. R. Duda, P. Storck et D. Hart. Pattern Classification. Prentice Hall, 2002.