Classification par Support Vector Machines

Laurent Risser
Ingénieur de Recherche CNRS

0 : Préambule – Classification

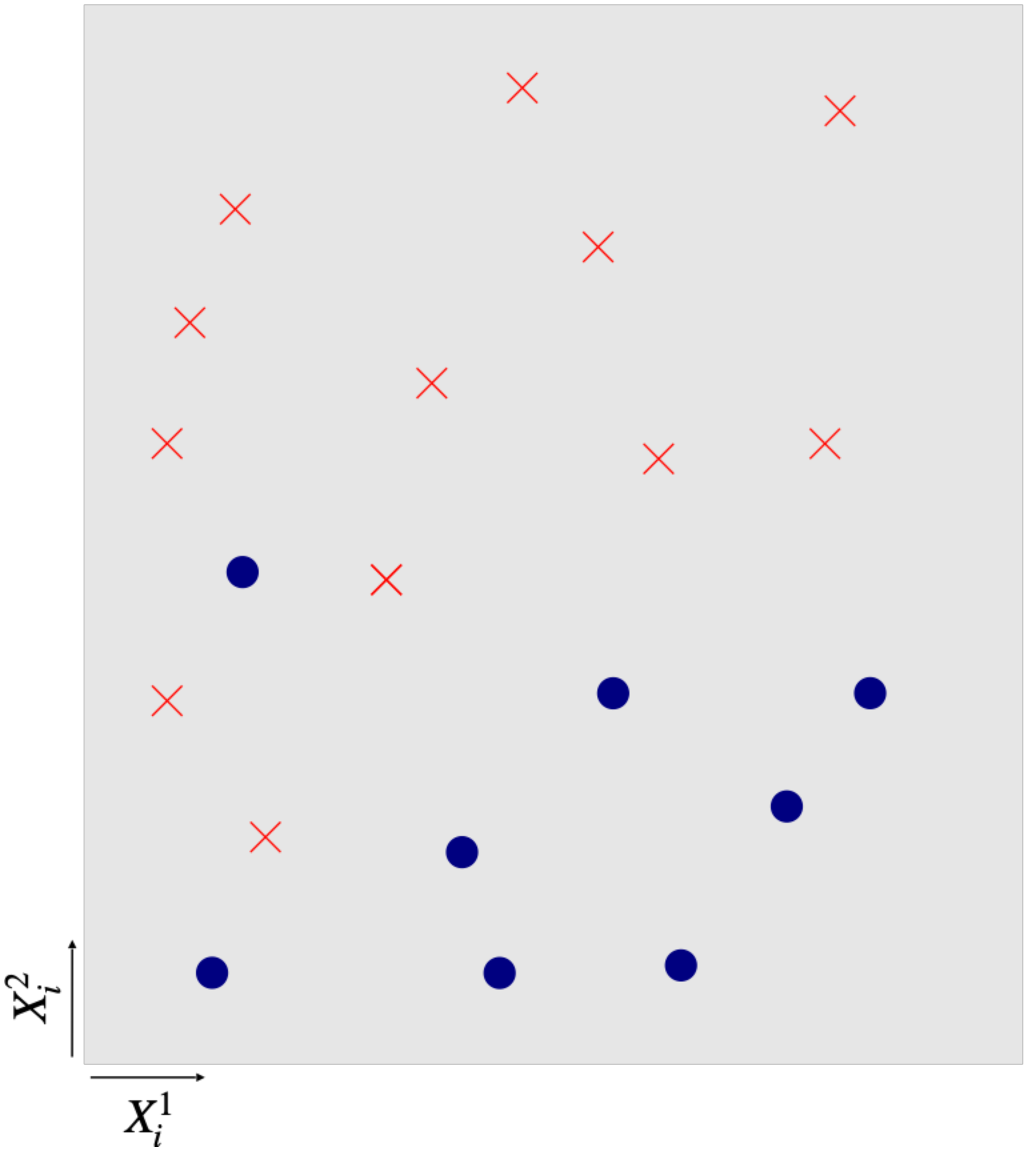
Aide au diagnostic

Base d'apprentissage	
Patient 1 : • Age = 40 • Globule Blancs/L = 6	Sain
Patient 2 : • Age = 28 • Globule Blancs/L = 12	Rhume
Patient N : • Age = 57 • Globule Blancs/L = 8	Sain

Nouveau Patient :
- Age = 34
- Globule Blancs/L = 5

→ Sain ou rhume ???

Apprentissage supervisé — classification



Observations d'entrée (*X*) :

- *n* observations $X_i \in \mathbb{R}^p$
- Ici *n = 20* et *p = 2*

Observations de sortie (*Y*) :

- *n* Labels $Y_i \in \{-1,1\}$
- × $Y_i = 1$
- ● $Y_i = -1$

<u>Dans notre exemple</u> :

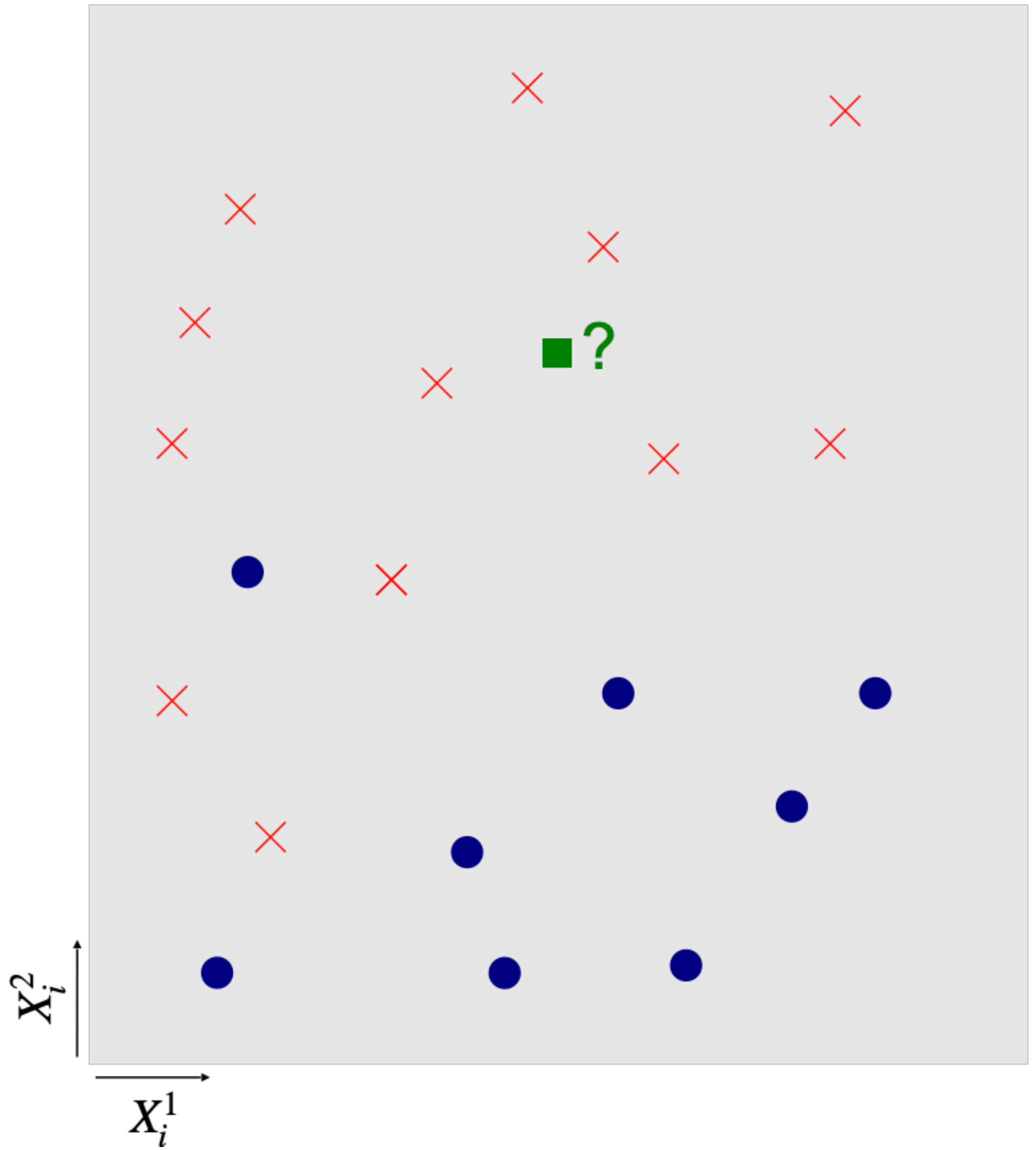
$i \longrightarrow$ Patient de la base d'apprentissage

$X_i^1 \longrightarrow$ Age

$X_i^2 \longrightarrow$ Globule Blancs/L

$Y_i \longrightarrow$ Sain ou rhume

Apprentissage supervisé — classification



Observations d'entrée (*X*) :

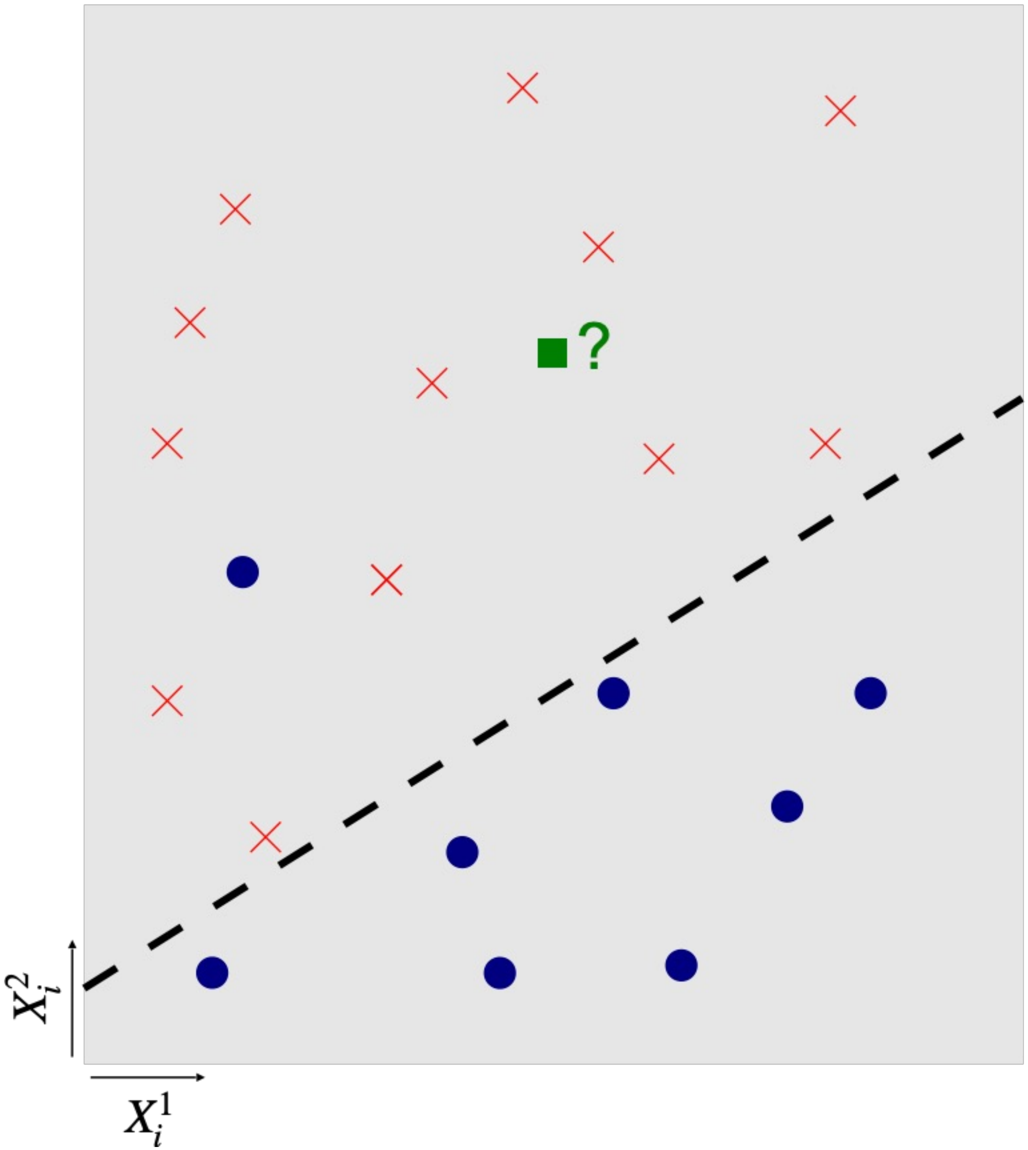
- *n* observations $X_i \in \mathbb{R}^p$
- Ici *n = 20* et *p = 2*

Observations de sortie (*Y*) :

- *n* Labels $Y_i \in \{-1,1\}$
- × $Y_i = 1$
- ● $Y_i = -1$

Label le plus probable de ■ ?

Apprentissage supervisé — classification



Observations d'entrée (*X*) :

- *n* observations $X_i \in \mathbb{R}^p$
- Ici *n = 20* et *p = 2*

Observations de sortie (*Y*) :

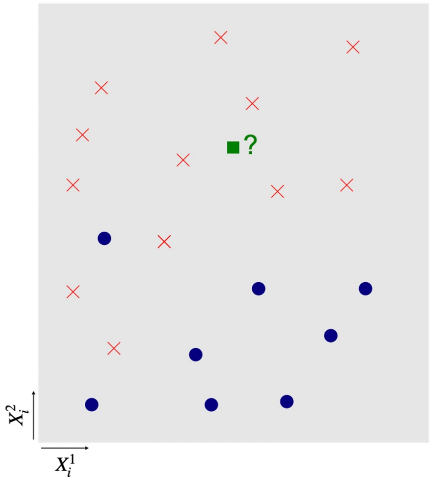
- *n* Labels $Y_i \in \{-1,1\}$
- × $Y_i = 1$
- ● $Y_i = -1$

Label le plus probable de ■ ?

1. **Choix d'un modèle** pour séparer les données d'apprentissage, i.e. les ● et les × .
2. **Apprentissage des paramètres** optimaux
3. Une fois les paramètres du modèle appris, **prédiction** extrêmement simple et rapide de ■ .

n observations d'apprentissage

- Entrée : $x_i = (x_i^1, \ldots, x_i^p) \in \mathbb{R}^p$
- Sortie : $y_i \in \{-1,1\}$

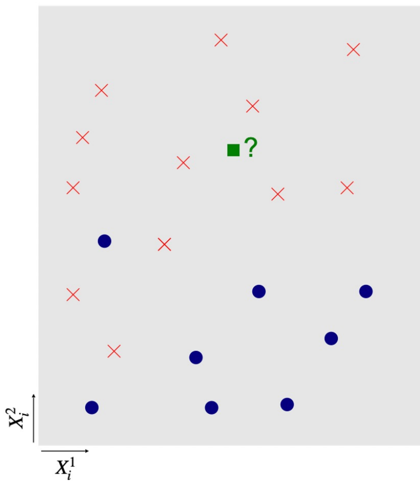


On note $\mathbb{P}(Y = 1|X)$ la loi conditionnelle que Y soit égal à 1 sachant X.

0 : Préambule – Régression logistique

n observations d'apprentissage

- Entrée : $x_i = (x_i^1, \dots, x_i^p) \in \mathbb{R}^p$
- Sortie : $y_i \in \{-1,1\}$

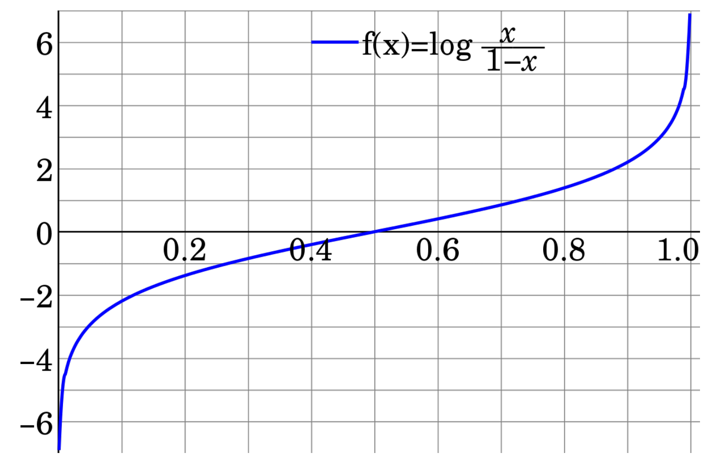


On note $\mathbb{P}(Y=1|X)$ la loi conditionnelle que Y soit égal à 1 sachant X.

On suppose alors que : $$\ln \frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

A apprendre

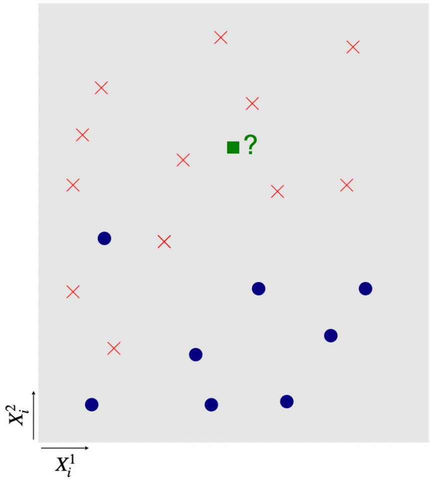
Connu



Fonction logit

n observations d'apprentissage

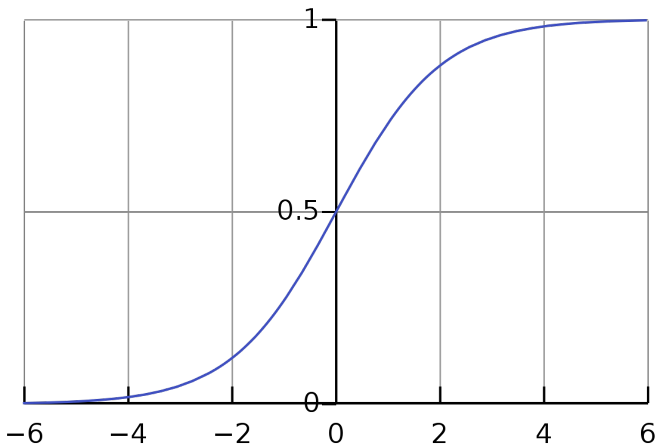
- Entrée : $x_i = (x_i^1, \dots, x_i^p) \in \mathbb{R}^p$
- Sortie : $y_i \in \{-1,1\}$



On note $\mathbb{P}(Y = 1|X)$ la loi conditionnelle que Y soit égal à 1 sachant X.

On suppose alors que : $$\ln \frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

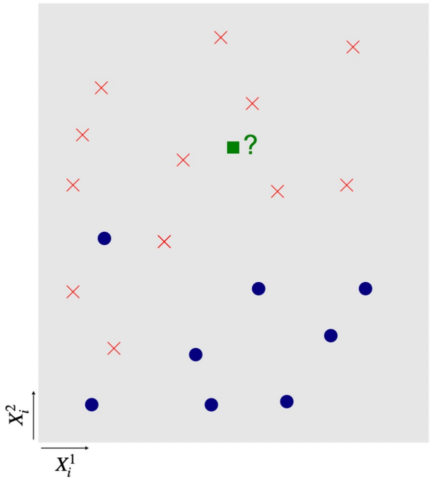
$$\mathbb{P}(Y=1|X) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}$$



Fonction logistique

n observations d'apprentissage

- Entrée : $x_i = (x_i^1, \dots, x_i^p) \in \mathbb{R}^p$
- Sortie : $y_i \in \{-1,1\}$



On note $\mathbb{P}(Y=1|X)$ la loi conditionnelle que Y soit égal à 1 sachant X.

On suppose alors que : $$\ln \frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

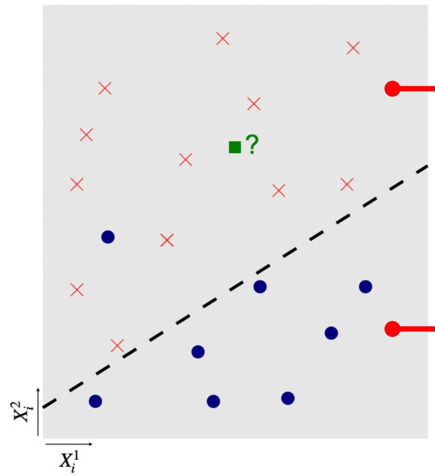
$$\mathbb{P}(Y=1|X) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}$$

pour une observation i, $i = 1, \dots, n$: $$p(y_i = 1|x_i^1, \dots, x_i^p) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_i^j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_i^j}}$$

Maximisation de la vraisemblance : $$L(\beta) = \prod_{i=1}^{n} \left[\left(p(y_i = 1|x_i^1, \dots, x_i^p) \right)^{y_i} . \left(1 - p(y_i = 1|x_i^1, \dots, x_i^p) \right)^{1-y_i} \right]$$

n observations d'apprentissage

- Entrée : $x_i = (x_i^1, \dots, x_i^p) \in \mathbb{R}^p$
- Sortie : $y_i \in \{-1,1\}$



$$\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{test}^j > 0$$

$$\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{test}^j < 0$$

On note $\mathbb{P}(Y=1|X)$ la loi conditionnelle que Y soit égal à 1 sachant X.

On suppose alors que : $$\ln \frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

$$\mathbb{P}(Y=1|X) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}$$

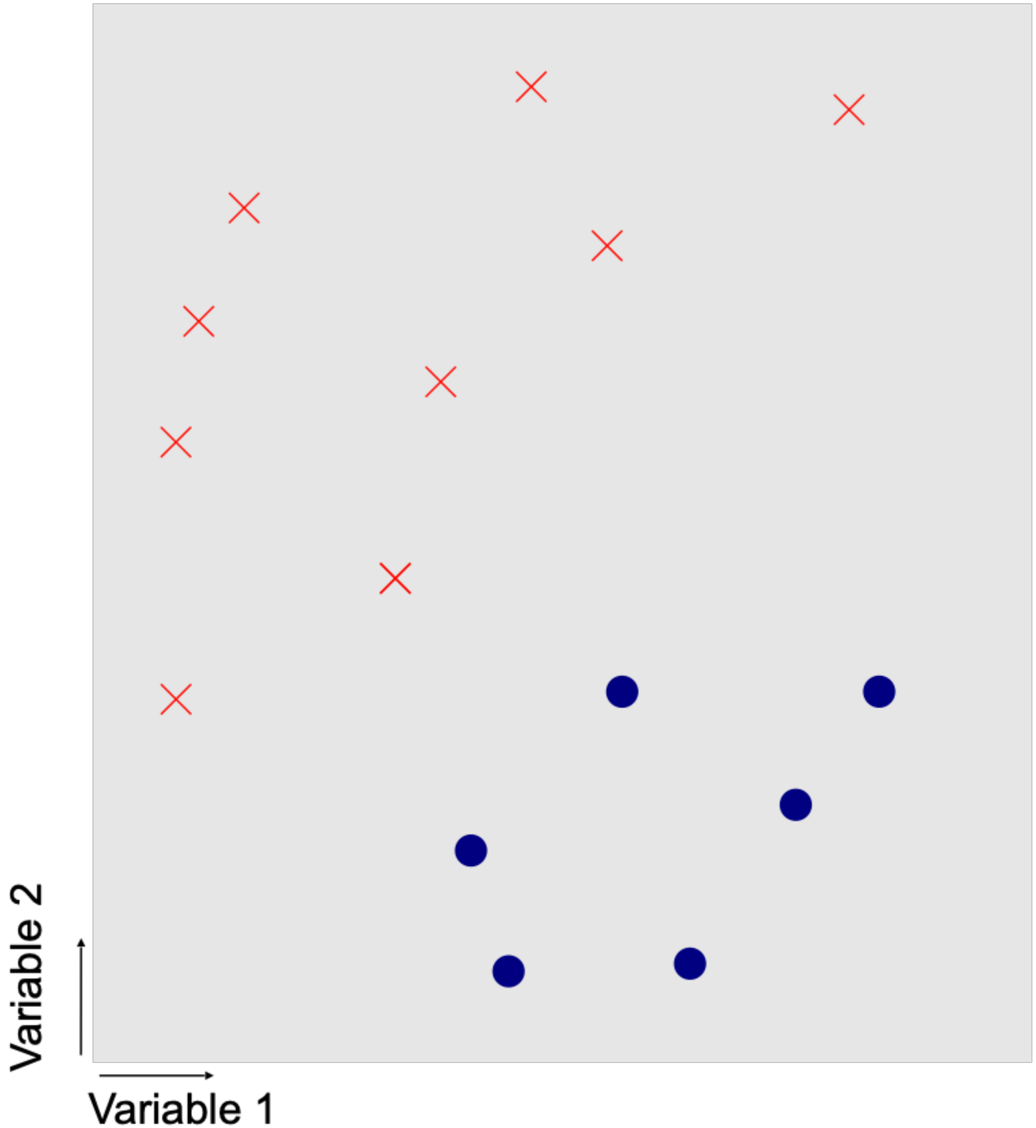
pour une observation i, $i = 1, \dots, n$: $$p(y_i = 1 | x_i^1, \dots, x_i^p) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_i^j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_i^j}}$$

Maximisation de la vraisemblance : $$L(\beta) = \prod_{i=1}^{n} \left[\left(p(y_i = 1 | x_i^1, \dots, x_i^p) \right)^{y_i} . \left(1 - p(y_i = 1 | x_i^1, \dots, x_i^p) \right)^{1-y_i} \right]$$

1 : Vision simplifiée des Support Vector Machines

Une manière plus récente d'attaquer le problème est celle des *Support Vector Machines*

Voici l'idée principale :

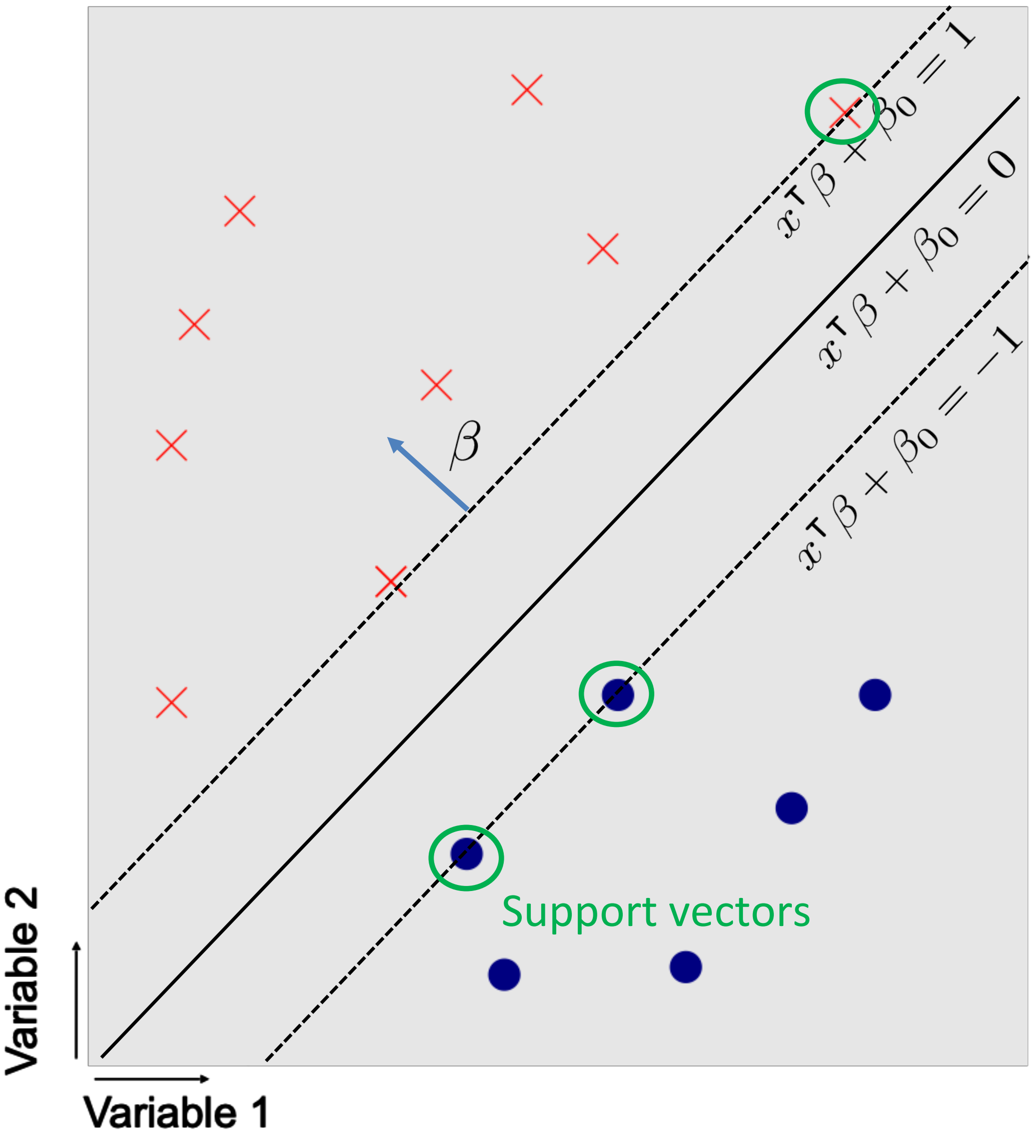


$\boldsymbol{x_1}$, $\boldsymbol{x_2}$, … , $\boldsymbol{x_N}$ les observations
(ici : x_i est la coordonnée du point en 2D)

y_1 , y_2 , … , y_N les labels
(ici : × → 1 et ● → -1)

Une manière plus récente d'attaquer le problème est celle des *Support Vector Machines*

Voici l'idée principale :



$\boldsymbol{x_1}$, $\boldsymbol{x_2}$, … , $\boldsymbol{x_N}$ les observations
(ici : x_i est la coordonnée du point en 2D)

y_1 , y_2 , … , y_N les labels
(ici : × → 1 et ● → -1)

On optimise

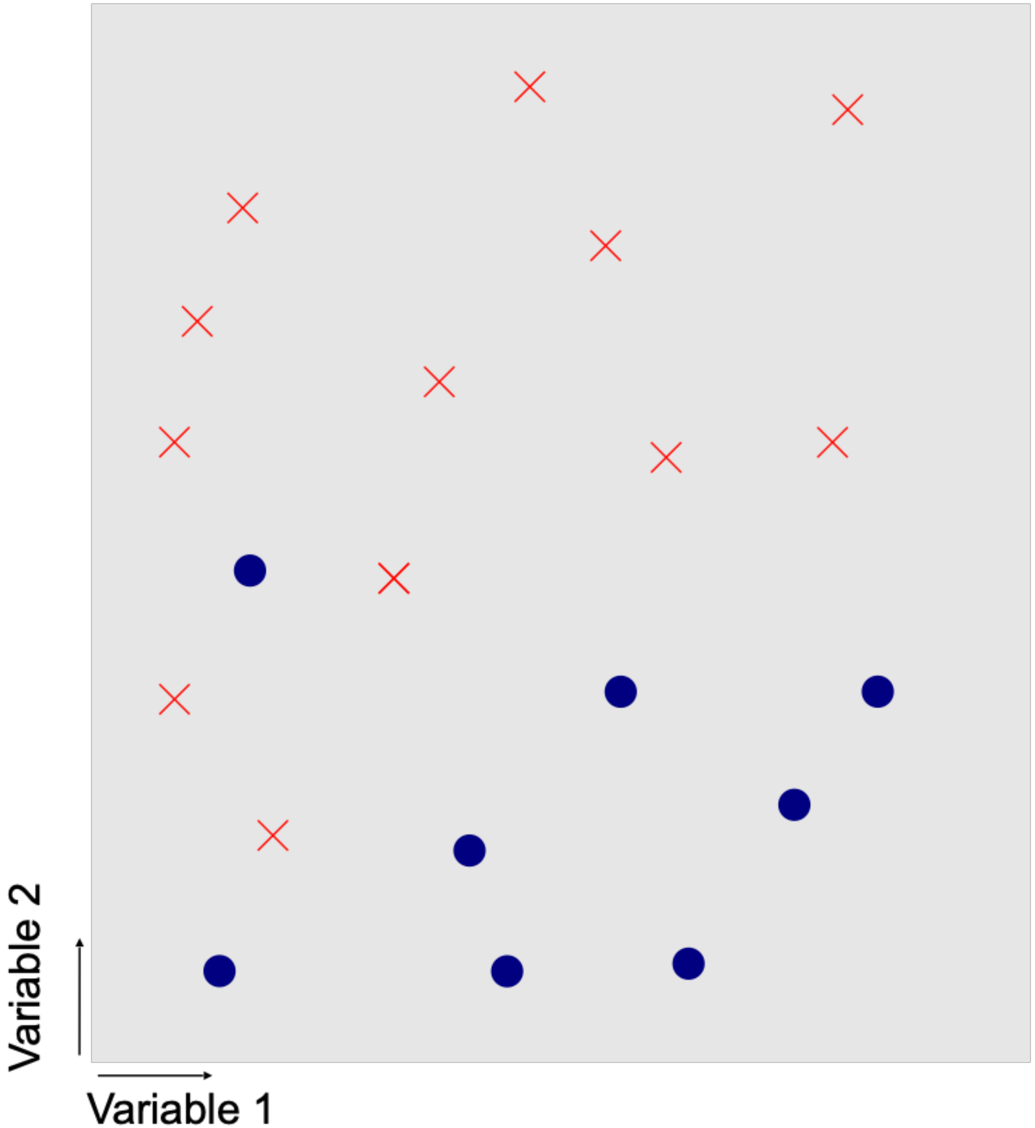
$$y_i \left(\sum_{j=1}^{p} \beta^j x_i^j + \beta_0 \right) \geq 1, \forall i \in \{1, \dots, n\}$$

en fonction de $\beta = (\beta^1, \dots, \beta^p)^\intercal$ et β^0

1 : Vision simplifiée des Support Vector Machines

Une manière plus récente d’attaquer le problème est celle des *Support Vector Machines*

Voici l’idée principale :



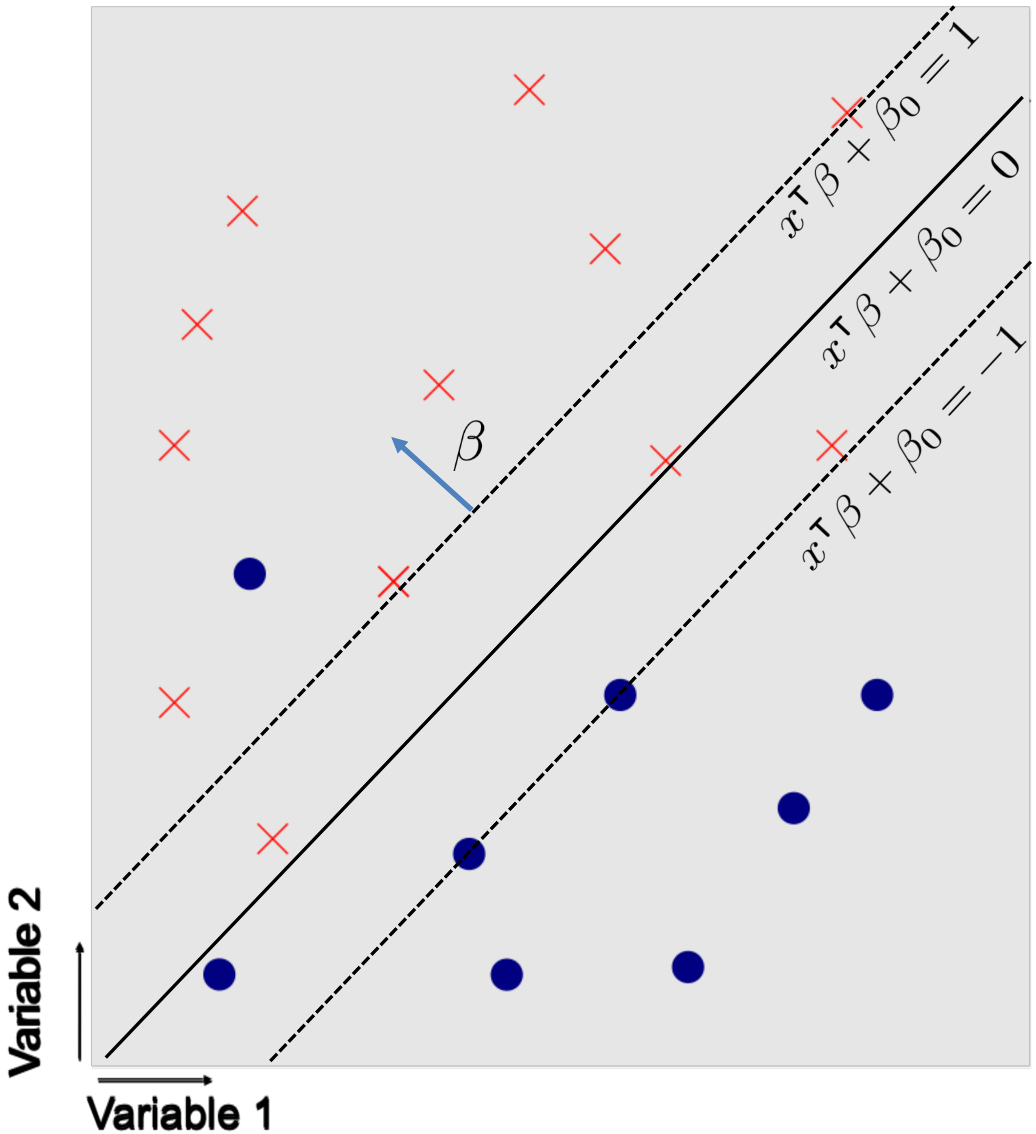
$\boldsymbol{x}_1$, $\boldsymbol{x}_2$, … , $\boldsymbol{x}_N$ les observations
(ici : x_i est la coordonnée du point en 2D)

y_1 , y_2 , … , y_N les labels
(ici : × → 1 et ● → -1)

Que faire si il est impossible de séparer tous les points?

Une manière plus récente d'attaquer le problème est celle des *Support Vector Machines*

Voici l'idée principale :



x_1 , x_2 , … , x_N les observations
(ici : x_i est la coordonnée du point en 2D)

y_1 , y_2 , … , y_N les labels
(ici : × → 1 et ● → -1)

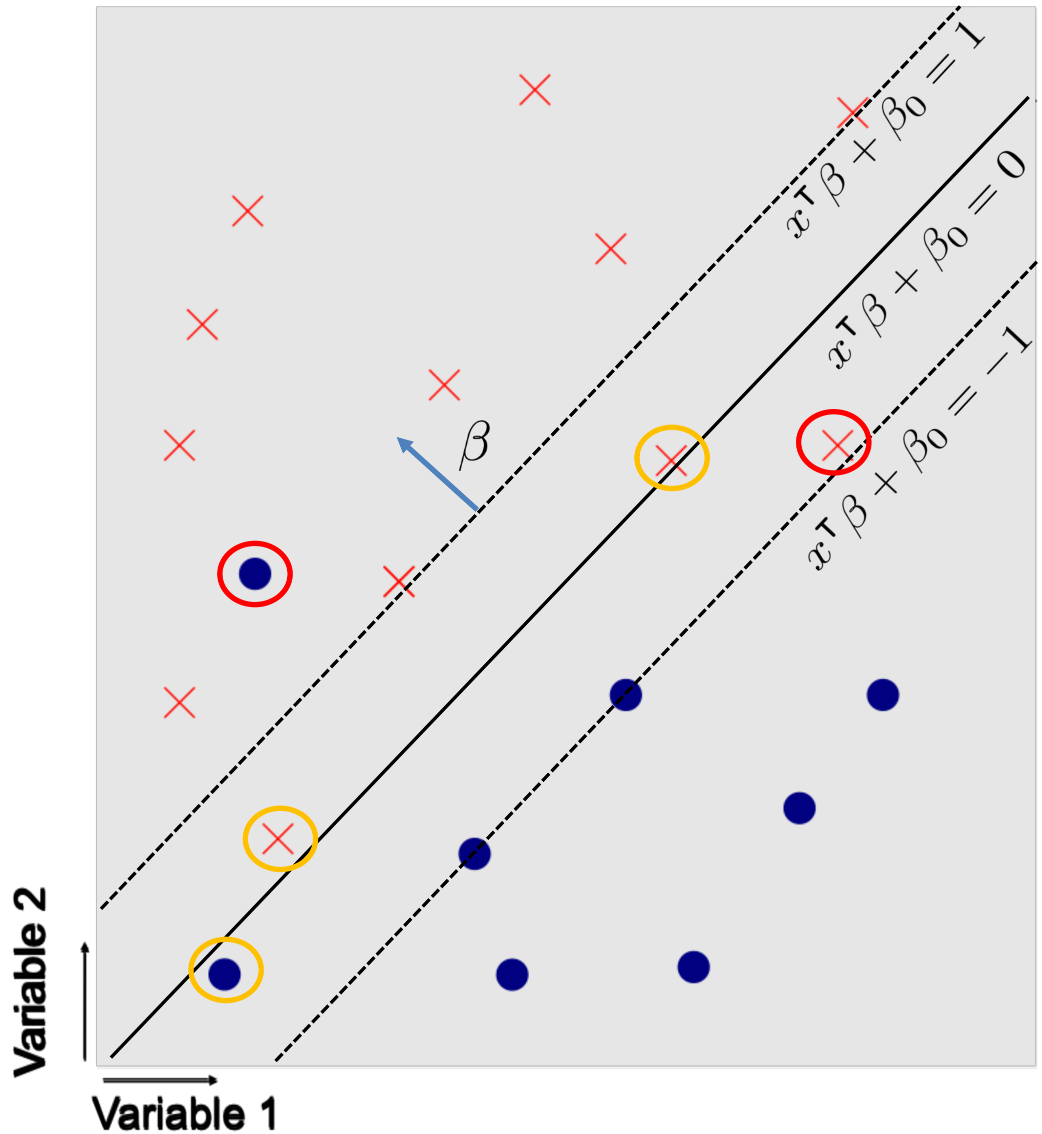
Que faire si il est impossible de séparer tous les points?

On peut optimiser en fct de β^0 et $\beta = (\beta^1, \dots, \beta^p)^\intercal$

$$\frac{1}{n}\sum_{i=1}^{n} \max\left(0, y_i\left(\sum_{j=1}^{p} \beta^j x_i^j - b\right)\right) + ||\beta||_2$$

Une manière plus récente d'attaquer le problème est celle des *Support Vector Machines*

Voici l'idée principale :



$\boldsymbol{x_1}$, $\boldsymbol{x_2}$, … , $\boldsymbol{x_N}$ les observations
(ici : x_i est la coordonnée du point en 2D)

y_1 , y_2 , … , y_N les labels
(ici : × → 1 et ● → -1)

Que faire si il est impossible de séparer tous les points?

On peut optimiser en fct de β^0 et $\beta = (\beta^1, \ldots, \beta^p)^\top$

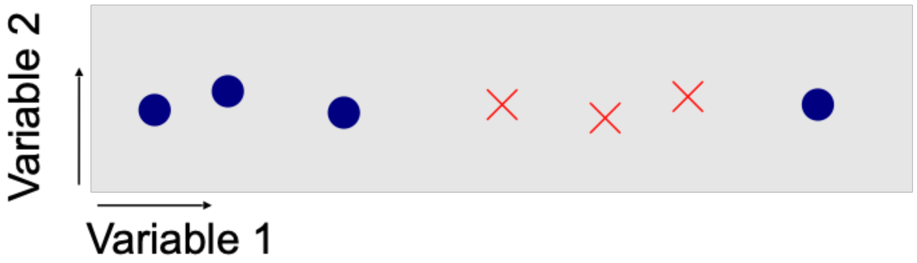
$$\frac{1}{n}\sum_{i=1}^{n} \max\left(0, y_i\left(\sum_{j=1}^{p}\beta^j x_i^j - b\right)\right) + ||\beta||_2$$

< 0 si mal classé

Contrainte sur *w*

Une manière plus récente d'attaquer le problème est celle des *Support Vector Machines*

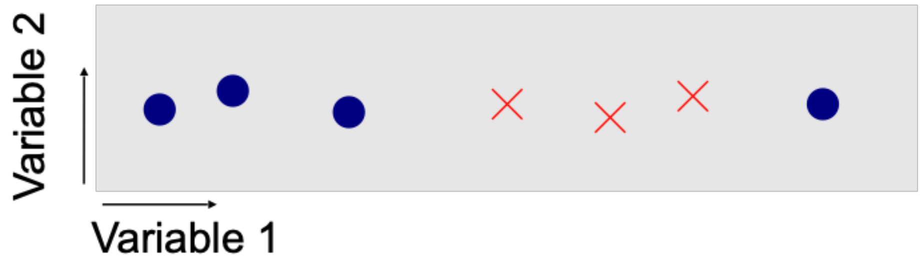
Un aspect important avec les SVM est l'utilisation des noyaux :



Que faire dans ce cas là ?

Une manière plus récente d'attaquer le problème est celle des *Support Vector Machines*

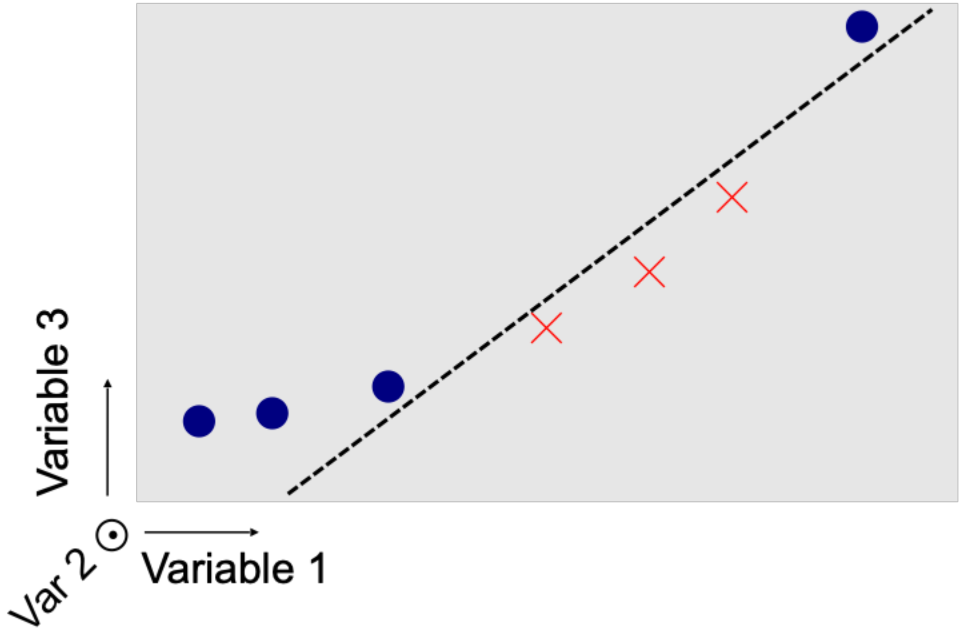
Un aspect important avec les SVM est l'utilisation des noyaux :



Que faire dans ce cas là ?

On note $x_i = (x_i^1, x_i^2)$ une observation

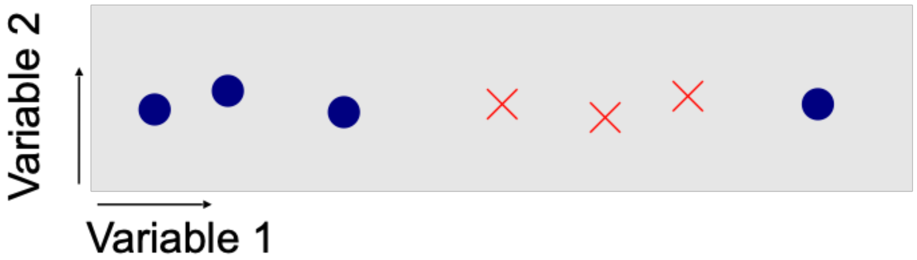
On va séparer les $\Phi(x_i) = \left(x_i^1, x_i^2, (x_i^1)^2\right)$ plutôt que les x_i



1 : Vision simplifiée des Support Vector Machines

Une manière plus récente d'attaquer le problème est celle des *Support Vector Machines*

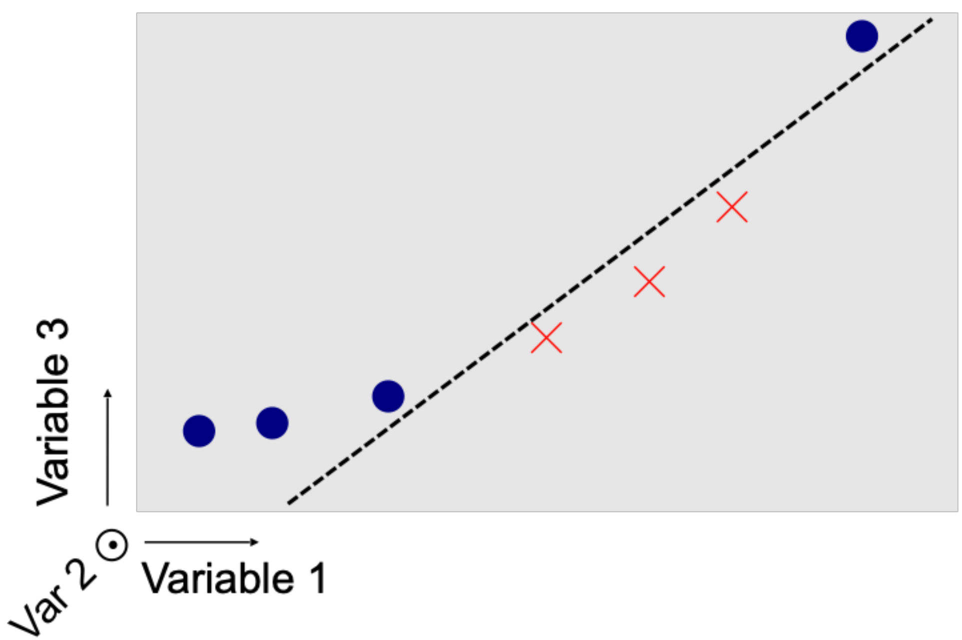
Un aspect important avec les SVM est l'utilisation des noyaux :



Que faire dans ce cas là ?

On note $x_i = (x_i^1, x_i^2)$ une observation

On va séparer les $\Phi(x_i) = \left(x_i^1, x_i^2, (x_i^1)^2\right)$ plutôt que les x_i

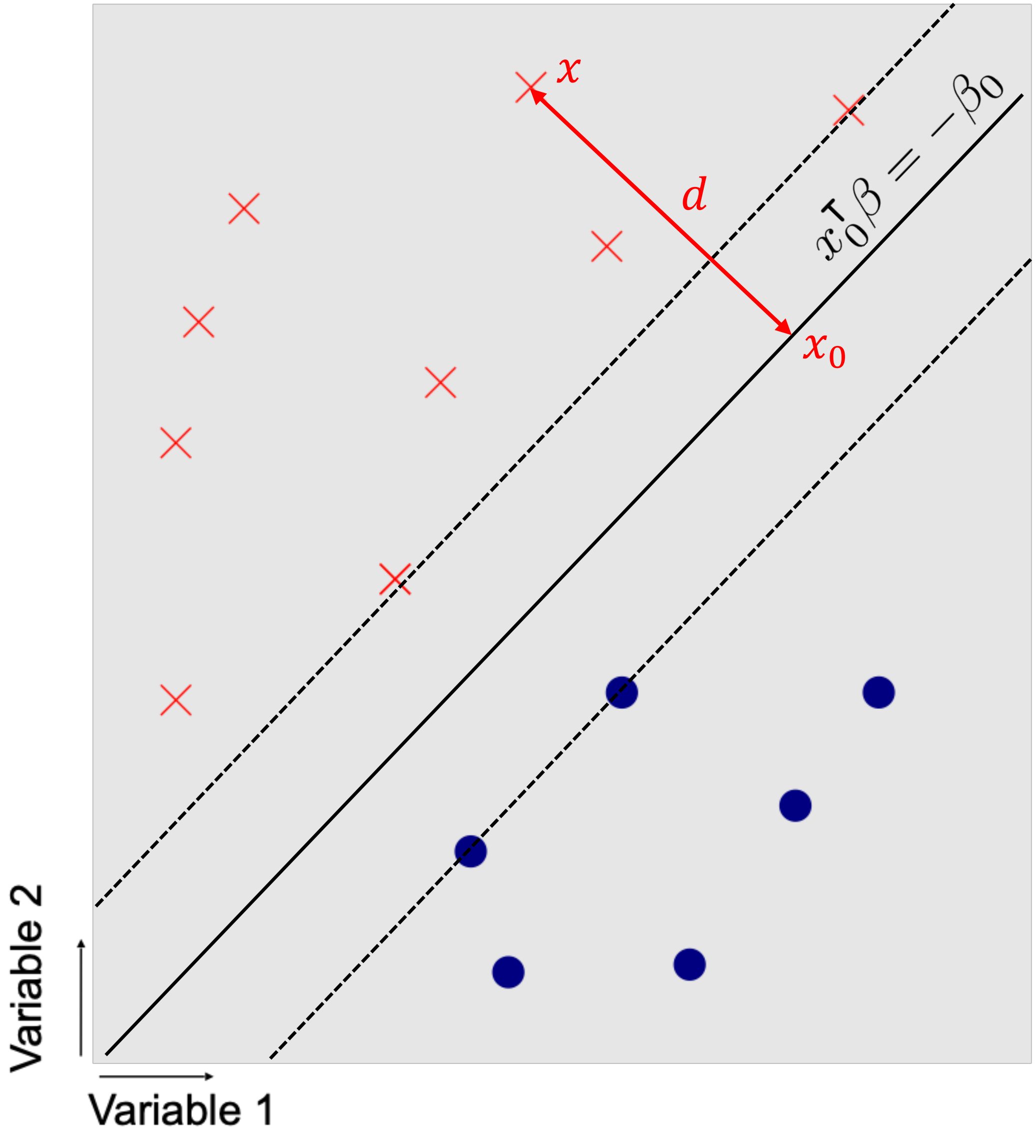


… optimisation efficace ?

… explosion de la dimension avec des noyaux pertinents ?

Voyons cela plus en détail !

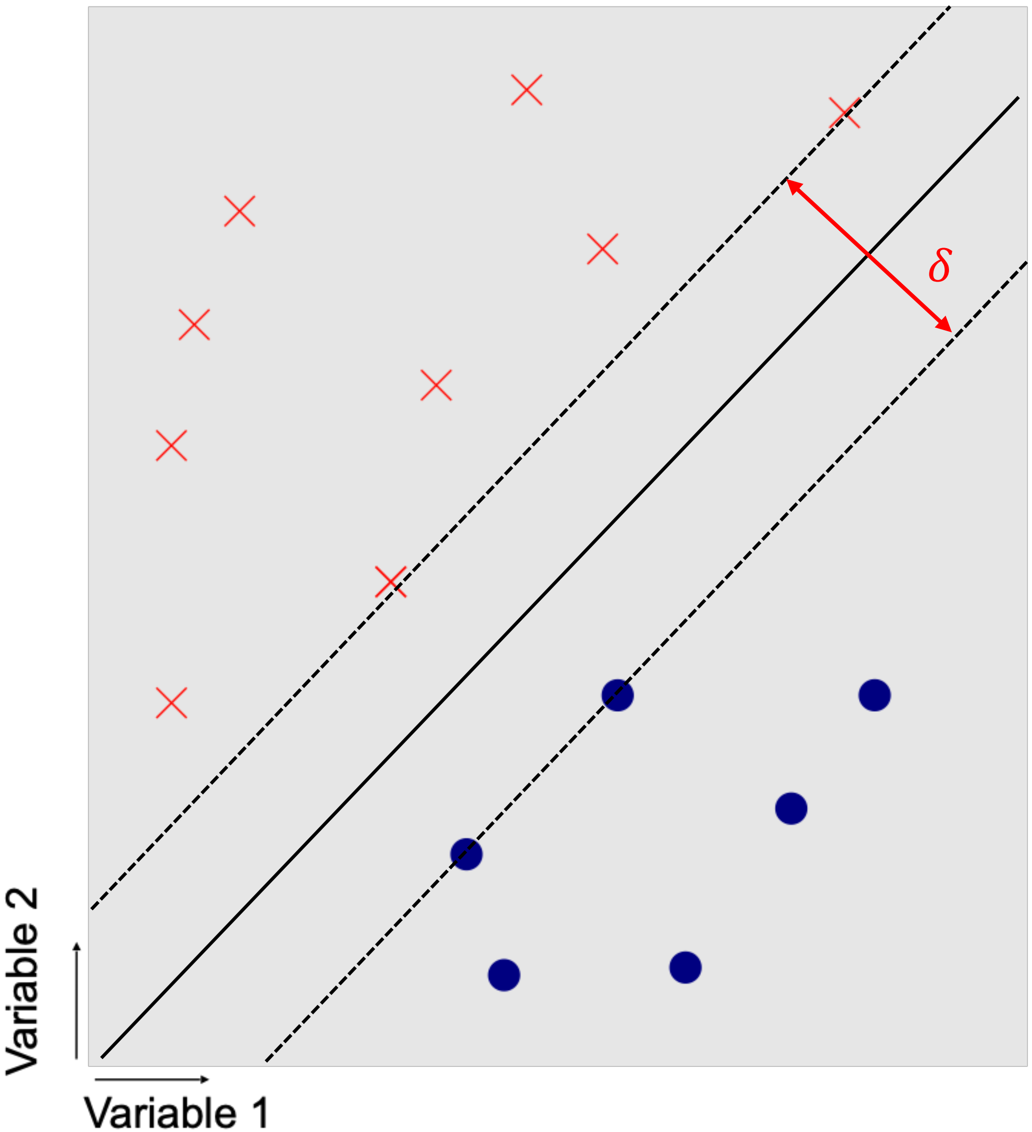
Creusons la construction mathématique des SVM → Distance d'un point x à la frontière



On peut montrer que : $d = \dfrac{|x^{\mathsf{T}}\beta + \beta_0|}{||\beta||}$

avec $||\beta|| = \sqrt{\sum_{j=1}^{p} (\beta^j)^2}$

Creusons la construction mathématique des SVM → Marge maximale

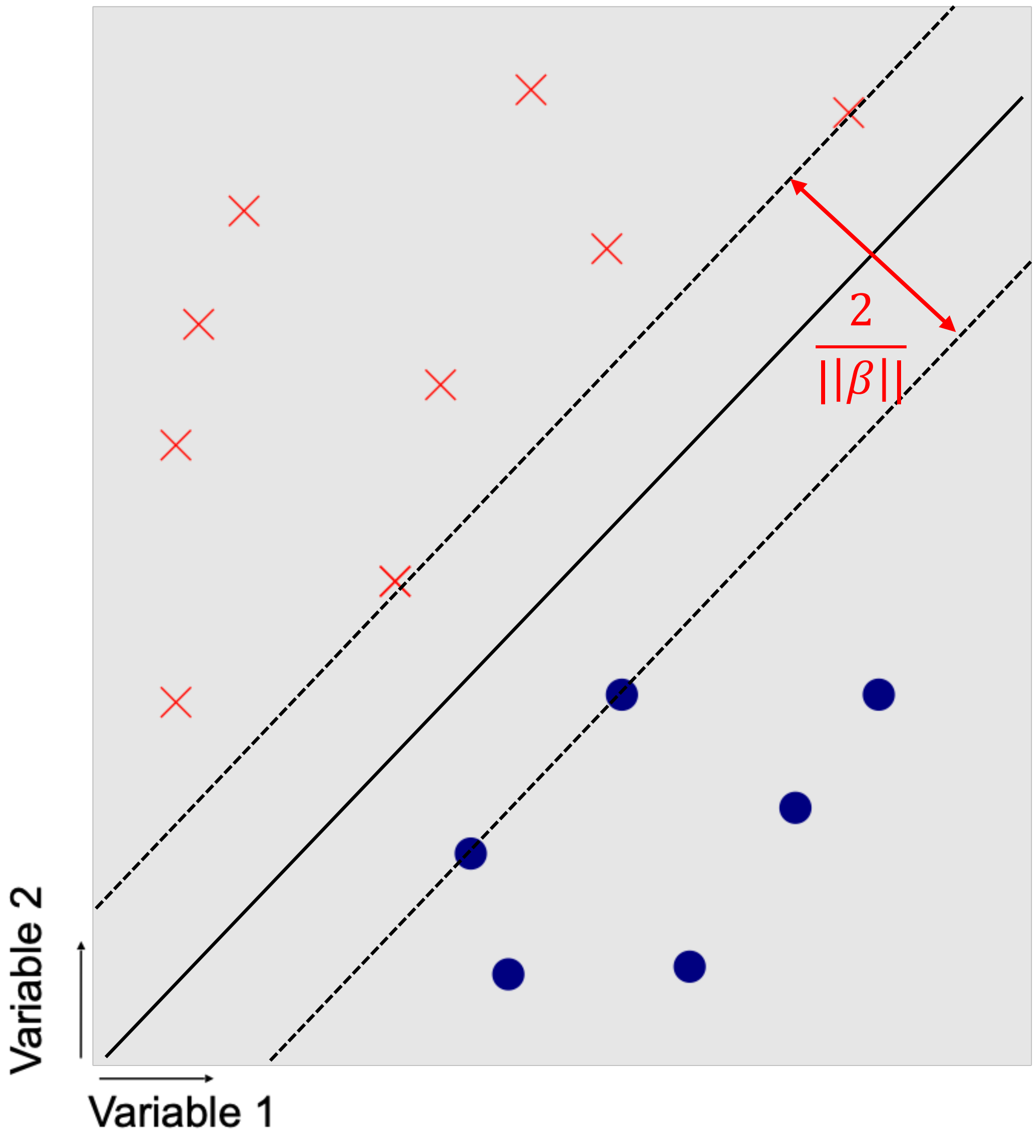


On peut montrer que : $d = \dfrac{|x^\intercal \beta + \beta_0|}{||\beta||}$

avec $||\beta|| = \sqrt{\sum_{j=1}^{p} (\beta^j)^2}$

Ainsi la marge maximale est : $\delta = \dfrac{2}{||\beta||}$

Creusons la construction mathématique des SVM → Problème brute



On peut montrer que : $d = \frac{|x^\intercal \beta + \beta_0|}{||\beta||}$

avec $||\beta|| = \sqrt{\sum_{j=1}^{p} (\beta^j)^2}$

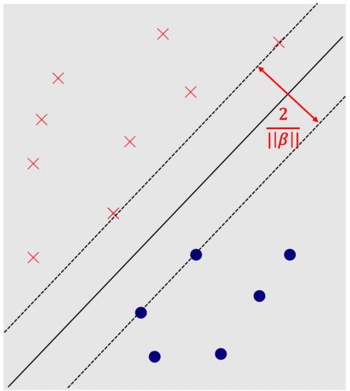
Ainsi la marge maximale est : $\delta = \frac{2}{||\beta||}$

Pour avoir **tous les points bien classés** avec une **marge aussi large que possible** :

$$\{\hat{\beta}, \hat{\beta}_0\} = \arg\min_{\beta, \beta_0} \frac{1}{2} ||\beta||$$

$$\text{s.t. } y_i(x_i^\intercal \beta + \beta_0) \geq 1, \forall i \in \{1, \ldots, n\}$$

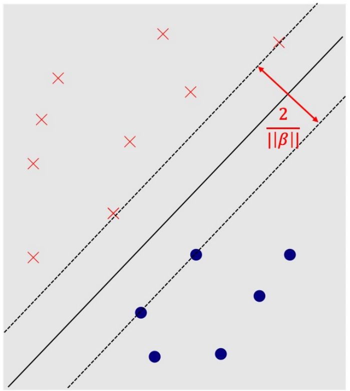
Problème d'optimisation sous contrainte :



$$\{\hat{\beta}, \hat{\beta}_0\} = \arg\min_{\beta, \beta_0} \frac{1}{2} ||\beta||$$

$$\text{s.t. } \; y_i(x_i^{\intercal}\beta + \beta_0) \geq 1, \forall i \in \{1, \ldots, n\}$$

Problème d'optimisation sous contrainte :



$$\{\hat{\beta}, \hat{\beta}_0\} = \arg\min_{\beta,\beta_0} \frac{1}{2}||\beta||$$
$$\text{s.t. } \; y_i(x_i^\intercal \beta + \beta_0) \geq 1, \forall i \in \{1, \ldots, n\}$$

Passage au Lagrangien :

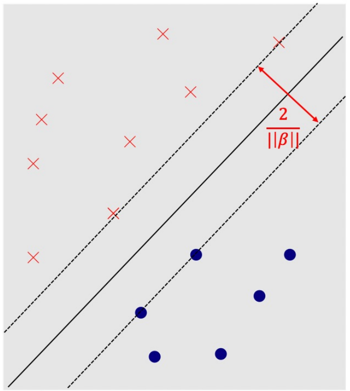
$$\mathcal{L}(\beta, \beta_0, \alpha) = \frac{1}{2}||\beta||^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(x_i^\intercal \beta + \beta_0) - 1\right)$$ avec $$\alpha_i \geq 0 \, , \forall i$$

Ainsi on résout :

$$\left\{ \begin{array}{l} \frac{\partial \mathcal{L}}{\partial \beta} = \beta - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \\ \frac{\partial \mathcal{L}}{\partial \beta_0} = \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \frac{\partial \mathcal{L}}{\partial \alpha_i} = y_i(x_i^\intercal \beta + \beta_0) - 1 = 0 \, , \forall i \end{array} \right.$$

2 : Construction mathématique des Support Vector Machines

Problème d'optimisation sous contrainte :



$$\{\hat{\beta}, \hat{\beta}_0\} = \arg\min_{\beta,\beta_0} \frac{1}{2}||\beta||$$
$$\text{s.t. } y_i(x_i^\intercal \beta + \beta_0) \geq 1, \forall i \in \{1, \ldots, n\}$$

Passage au Lagrangien :

$$\mathcal{L}(\beta, \beta_0, \alpha) = \frac{1}{2}||\beta||^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(x_i^\intercal \beta + \beta_0) - 1 \right)$$

avec $\alpha_i \geq 0 \, , \forall i$

Ainsi on résout :

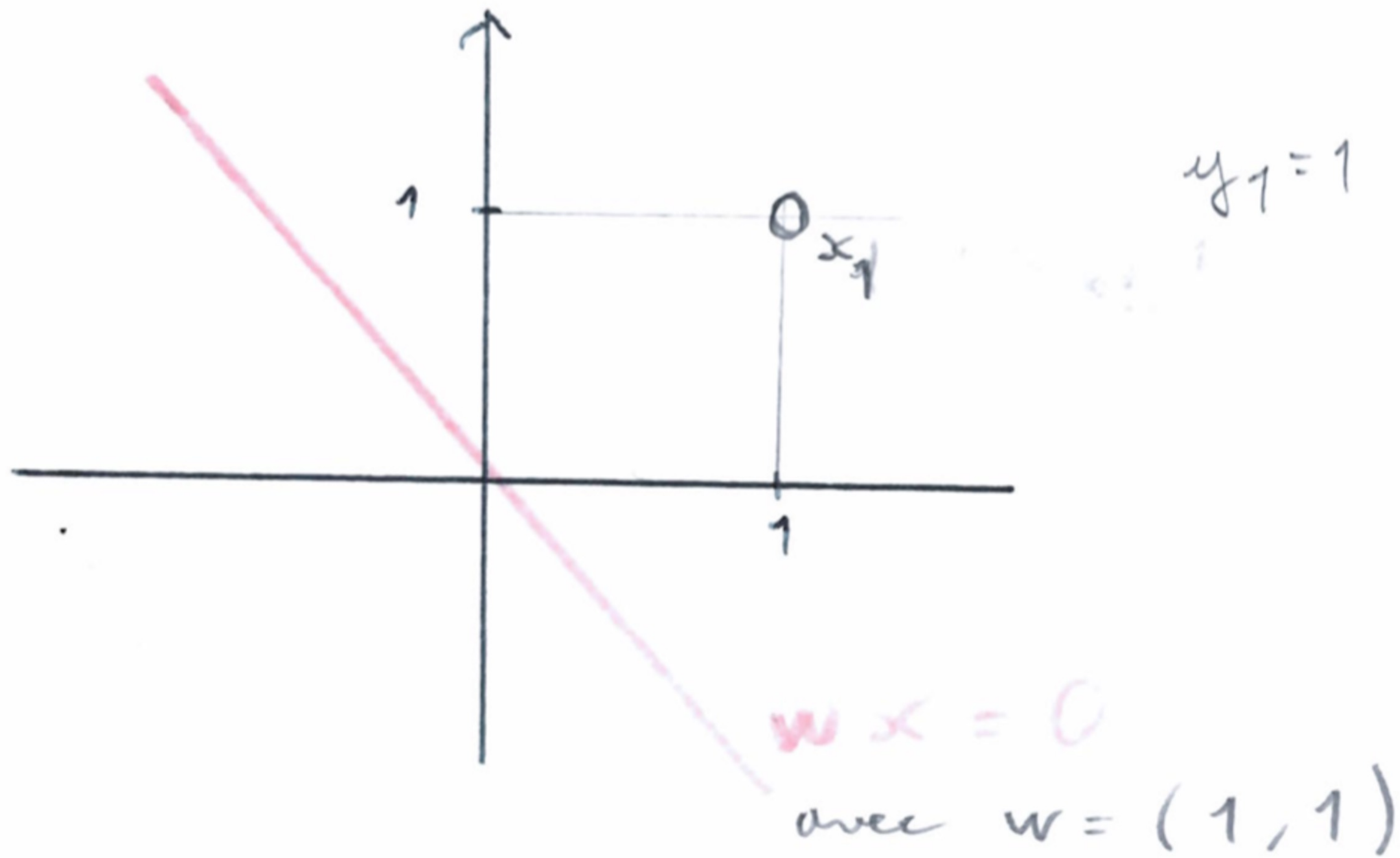
$$\begin{cases} \boxed{\frac{\partial \mathcal{L}}{\partial \beta} = \beta - \sum_{i=1}^{n} \alpha_i y_i x_i = 0} \\ \frac{\partial \mathcal{L}}{\partial \beta_0} = \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \frac{\partial \mathcal{L}}{\partial \alpha_i} = y_i(x_i^\intercal \beta + \beta_0) - 1 = 0 \, , \forall i \end{cases}$$

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

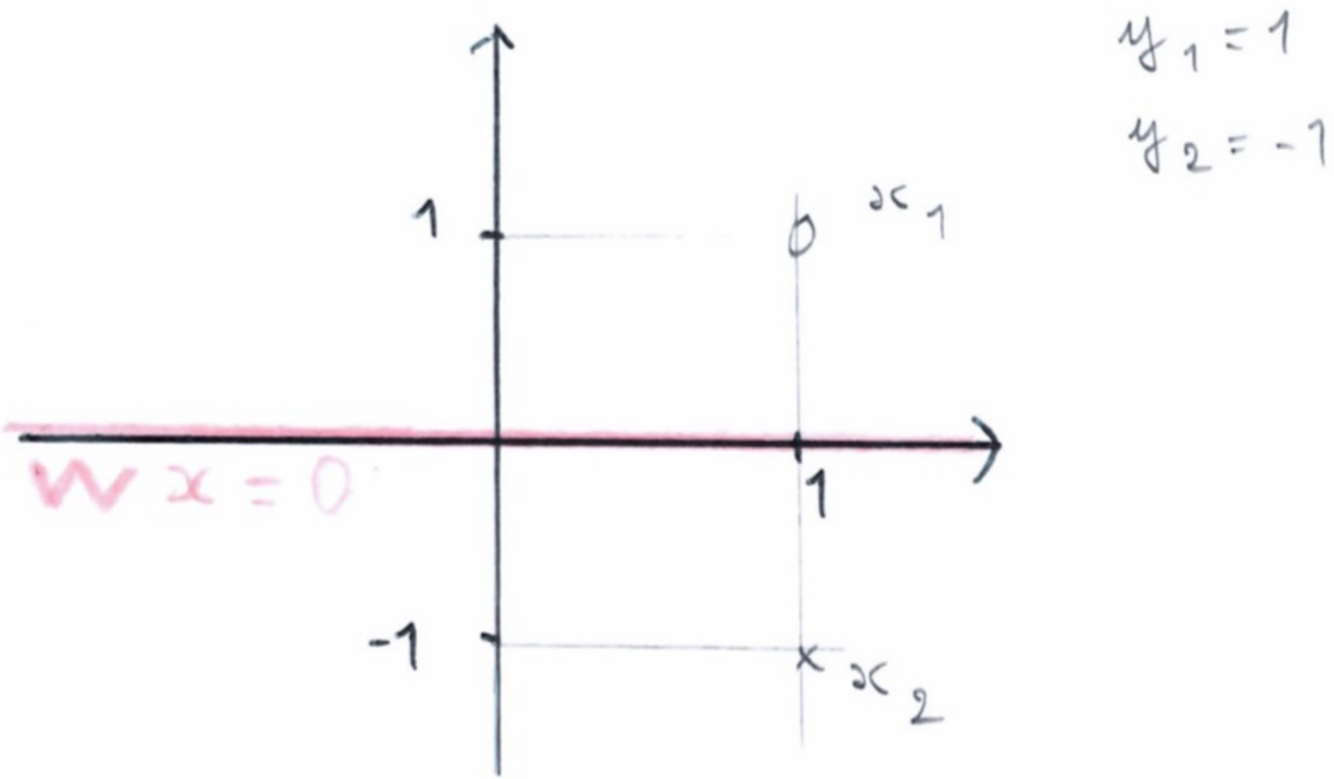
L'estimation de β peut être remplacée par celle des α_i !!!

(lié au *representer theorem*)

Intérêt pratique de $\beta = \sum_{i=1}^{n} \alpha_i y_i x_i$ pour séparer deux groupes d'observations

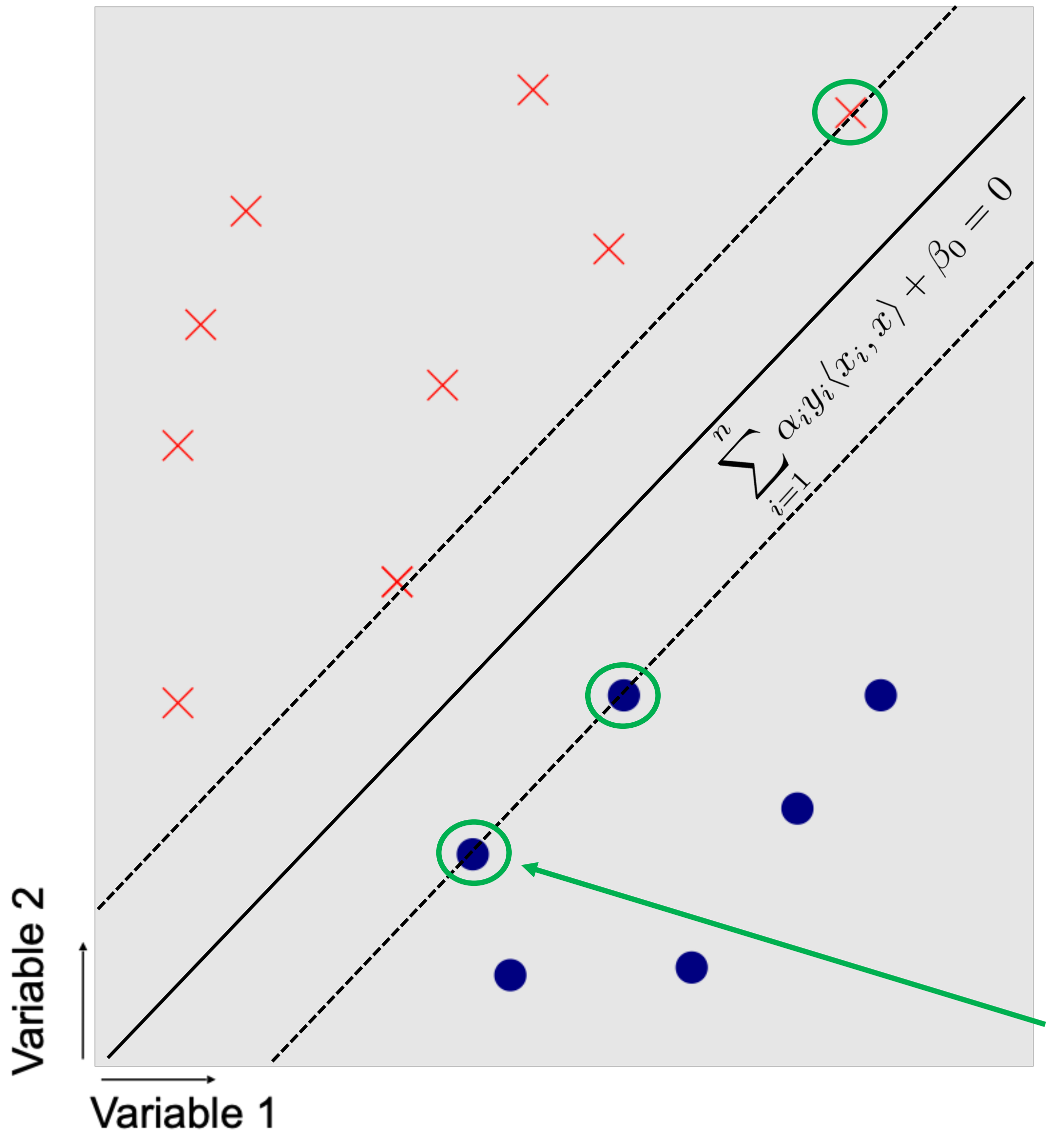


Intérêt pratique de $\beta = \sum_{i=1}^{n} \alpha_i y_i x_i$ pour séparer deux groupes d'observations



avec $W = 1 \times (1, 1) + (-1) \times (1, -1) = (0, 2)$

On va ainsi remplacer $x^\mathsf{T}\beta + \beta_0$ par $\sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle + \beta_0$, avec $\langle x_i, x \rangle = x_i^\mathsf{T} x$



Chaque α_i représente l'influence d'une observation !

Avec (rappel) :

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

$$\alpha_i \geq 0 \, , \, \forall i$$

Observations support avec $\alpha_i > 0$
(les autres $\alpha_i = 0$)

On a alors les conditions KKT suivantes pour les SVM

- Stationarité $\beta - \sum_{i=1}^{n} \alpha_i y_i x_i = 0$

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

- Admissibilité primale $y_i(\beta^\mathsf{T} x_i + \beta_0) \geq 1 \ , \forall i \in \{1, \dots, n\}$

- Admissibilité duale $\alpha_i \geq 0 \ , \forall i \in \{1, \dots, n\}$

- Complémentarité $\alpha_i \left(y_i(\beta^\mathsf{T} x_i + \beta_0) \right) = 0 \ , \forall i \in \{1, \dots, n\}$

Formulation duale des SVM (formulation de Wolfe – admise ici)

- On résout $\arg\max_{\beta,\beta_0,\alpha} \frac{1}{2}||\beta||^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(\beta^\intercal x_i + \beta_0) - 1 \right)$

- avec $\alpha_i \geq 0 \,,\, \forall i \in \{1, \ldots, n\}$

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

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

Formulation duale des SVM (formulation de Wolfe – admise ici)

- On résout $\arg\max_{\beta,\beta_0,\alpha} \frac{1}{2}||\beta||^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(\beta^{\mathsf{T}} x_i + \beta_0) - 1 \right)$

- avec $\alpha_i \geq 0 \, , \, \forall i \in \{1, \ldots, n\}$

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

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

Réécriture avec $\beta = \sum_{i=1}^{n} \alpha_i y_i x_i$ **et sans** $||\beta||^2$ **et** β_0

- On résout $\arg\max_{\alpha} -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_{i=1}^{n} \alpha_i$

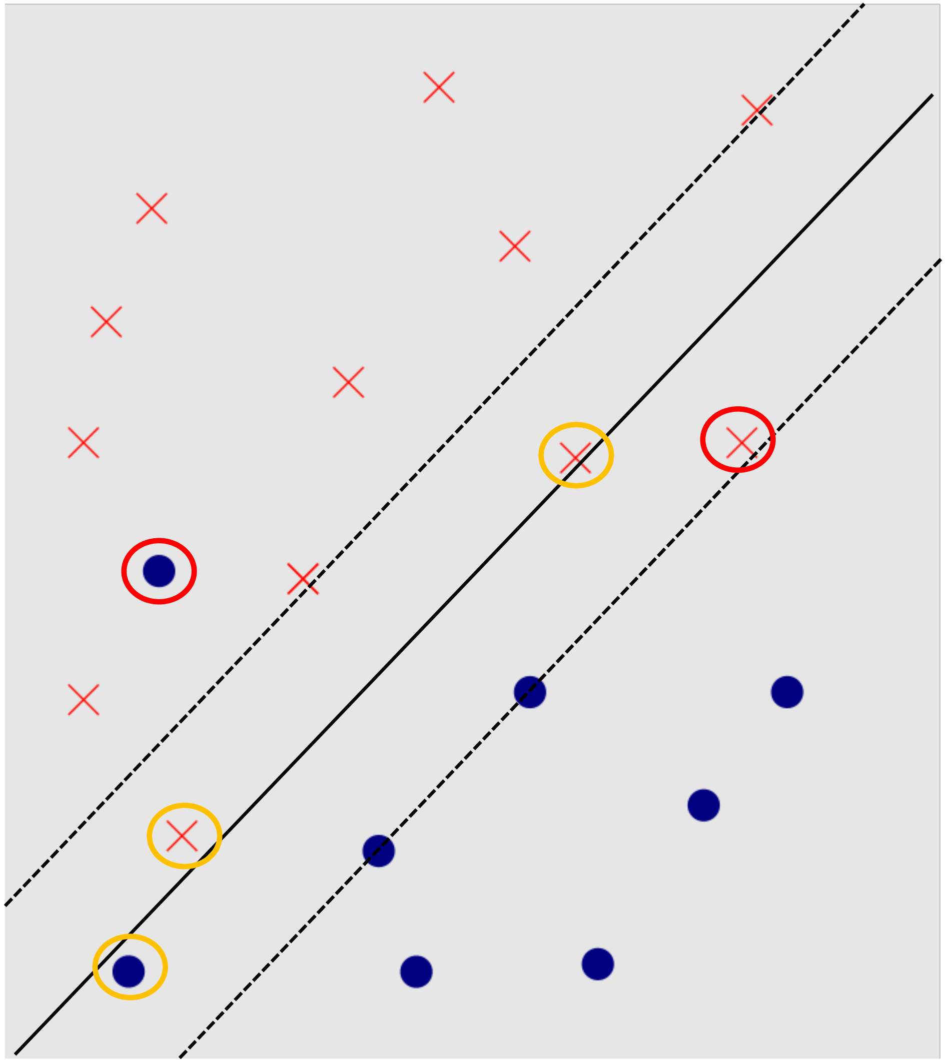
- avec $\alpha_i \geq 0 \, , \, \forall i \in \{1, \ldots, n\}$

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

Problème quadratique !

Peut être résolu de manière analytique ou avec une descente de gradient…

Et le cas non séparable ???

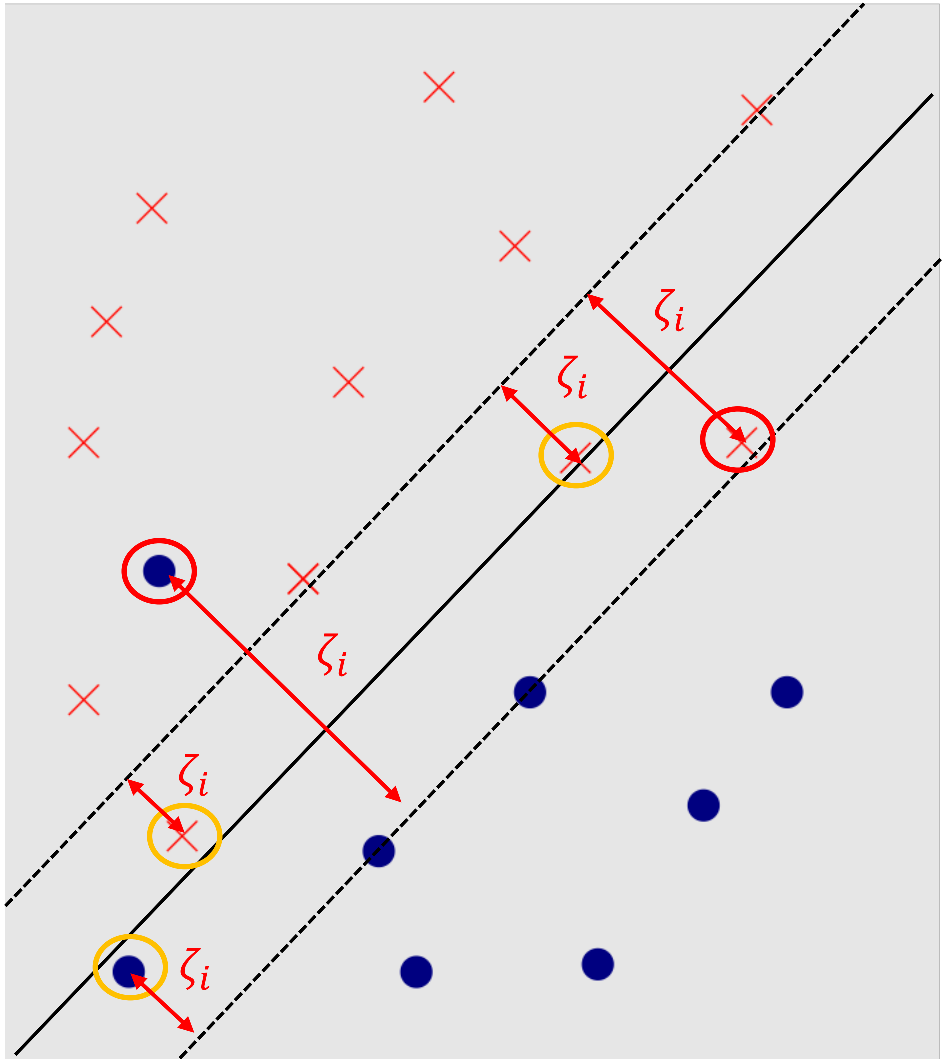


Et le cas non séparable ???

On introduit une « slack variable » : ζ_i

$$\text{Si } y_i(\beta^\intercal x_i + \beta_0) \geq 1 \text{ alors } \xi_i = 0$$

$$\text{Sinon } \xi_i = 1 - y_i(\beta^\intercal x_i + \beta_0) > 0$$

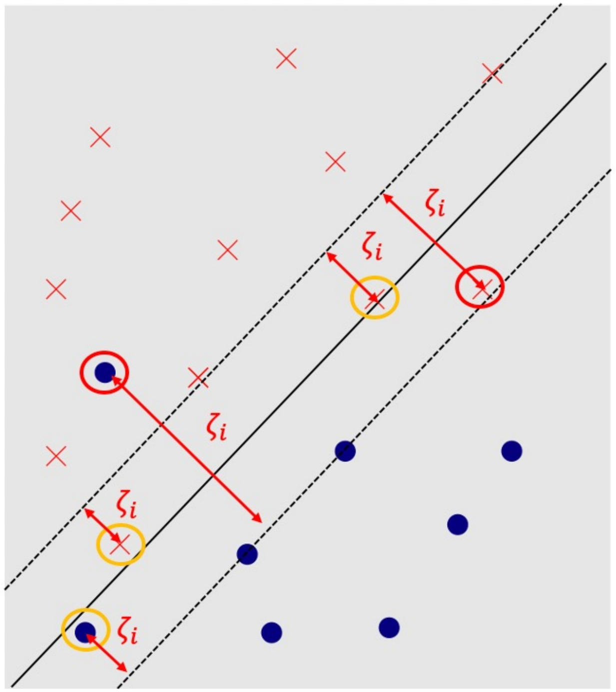


Problème d'optimisation sous contrainte

$$\{\hat{\beta}, \hat{\beta}_0\} = \arg\min_{\beta,\beta_0} \frac{1}{2}||\beta|| + \frac{C}{p}\sum_{i=1}^{p} \xi_i$$

$$\text{s.t. } y_i(x_i^\intercal \beta + \beta_0) \geq 1 - \xi_i, \; \forall i \in \{1, \ldots, n\}$$

$$\xi_i \geq 0, \; \forall i \in \{1, \ldots, n\}$$

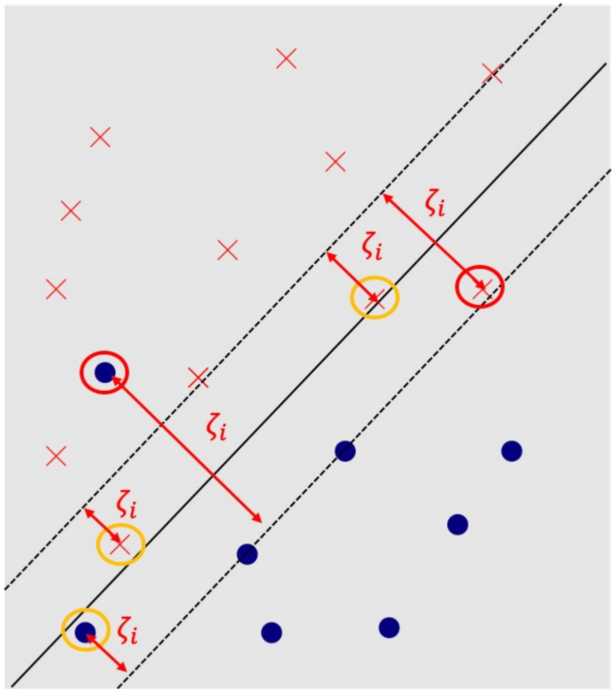


Problème d'optimisation sous contrainte

$$\{\hat{\beta}, \hat{\beta}_0\} = \arg\min_{\beta,\beta_0} \frac{1}{2}||\beta|| + \frac{C}{p}\sum_{i=1}^{p} \xi_i$$

$$\text{s.t. } y_i(x_i^\intercal \beta + \beta_0) \geq 1 - \xi_i, \ \forall i \in \{1, \ldots, n\}$$

$$\xi_i \geq 0, \ \forall i \in \{1, \ldots, n\}$$



Formulation Lagrangienne (avec $\alpha_i \geq 0$ et $\gamma_i \geq 0$)

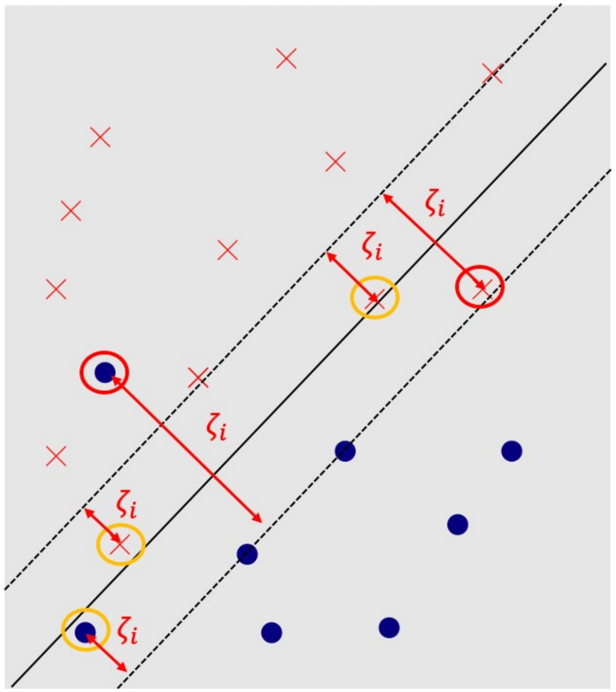
$$\mathcal{L}(\beta, \beta_0, \alpha, \gamma) = \frac{1}{2}||\beta||^2 + \frac{C}{p}\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left(y_i(x_i^\intercal \beta + \beta_0) - 1 + \xi_i\right) - \sum_{i=1}^{n} \gamma_i \xi_i$$

Problème d'optimisation sous contrainte

$$\{\hat{\beta}, \hat{\beta}_0\} = \arg\min_{\beta,\beta_0} \frac{1}{2}||\beta|| + \frac{C}{p}\sum_{i=1}^{p} \xi_i$$

$$\text{s.t. } y_i(x_i^\intercal \beta + \beta_0) \geq 1 - \xi_i, \; \forall i \in \{1,\ldots,n\}$$

$$\xi_i \geq 0, \; \forall i \in \{1,\ldots,n\}$$

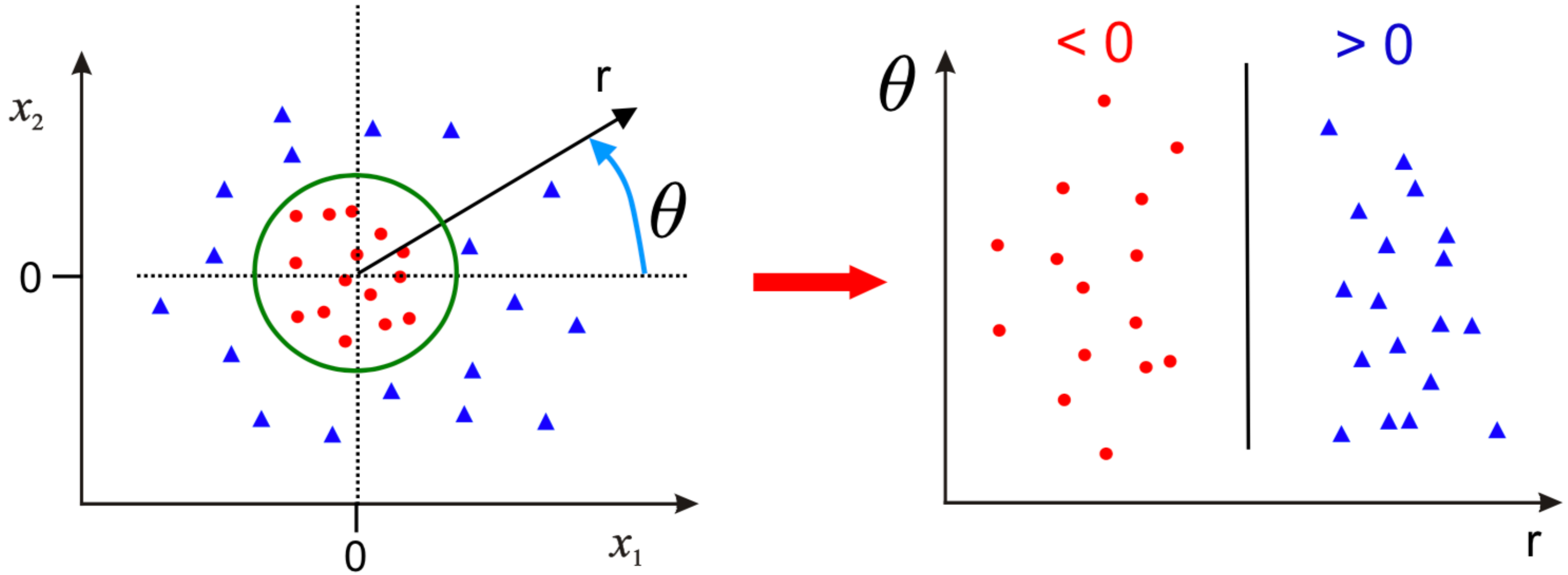


Formulation Lagrangienne (avec $\alpha_i \geq 0$ et $\gamma_i \geq 0$)

$$\mathcal{L}(\beta, \beta_0, \alpha, \gamma) = \frac{1}{2}||\beta||^2 + \frac{C}{p}\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left(y_i(x_i^\intercal \beta + \beta_0) - 1 + \xi_i\right) - \sum_{i=1}^{n} \gamma_i \xi_i$$

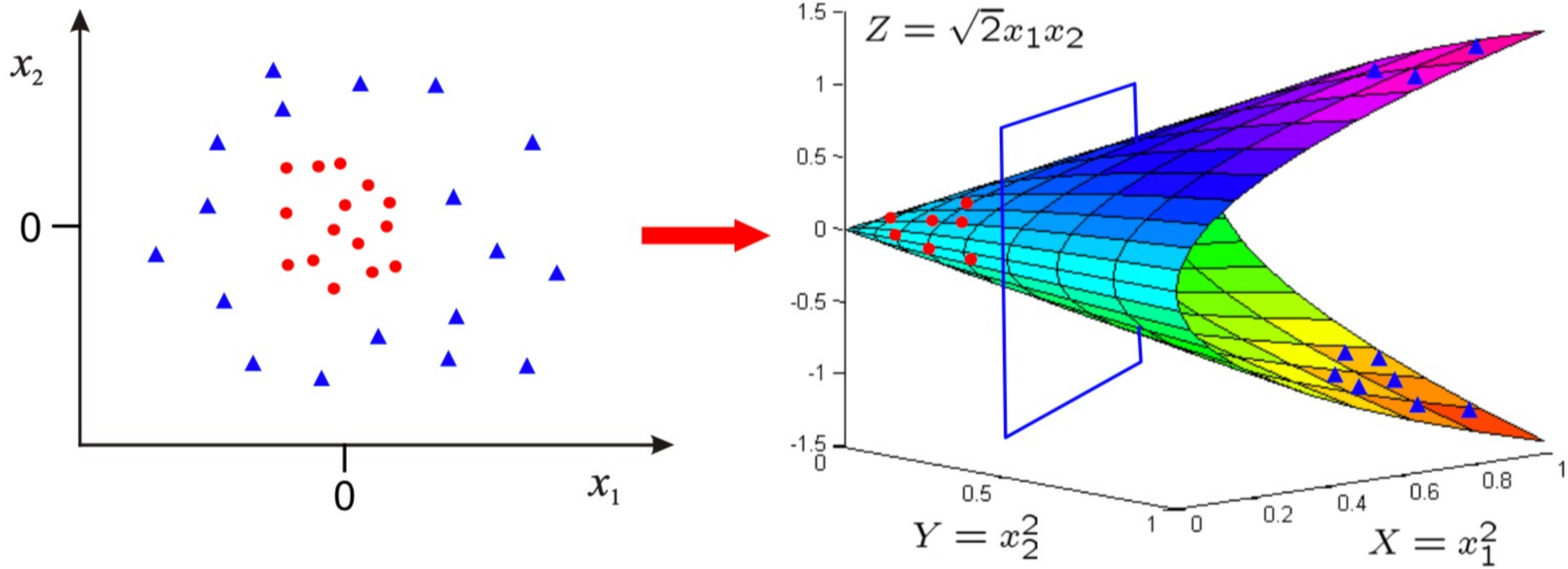
Nous passerons la formulation duale… au final le problème est résolu comme dans le cas séparable, mais $0 \leq \alpha_i \leq C$ et pas seulement $0 \leq \alpha_i$ … problème toujours quadratique !

Voyons deux exemples élémentaires où des transformations non-linéaires permettent de séparer des données facilement



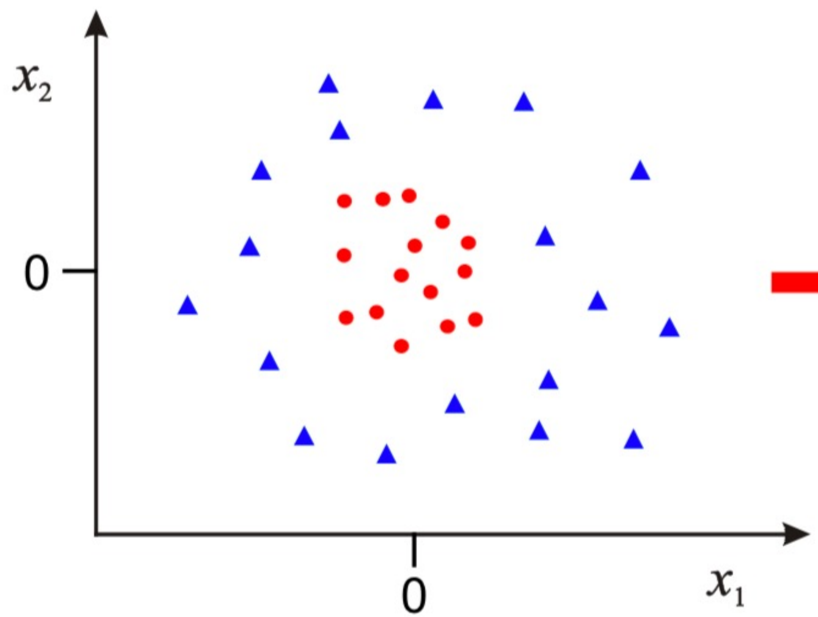
$$\Phi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \rightarrow \begin{pmatrix} r \\ \theta \end{pmatrix} \qquad \mathbb{R}^2 \rightarrow \mathbb{R}^2$$

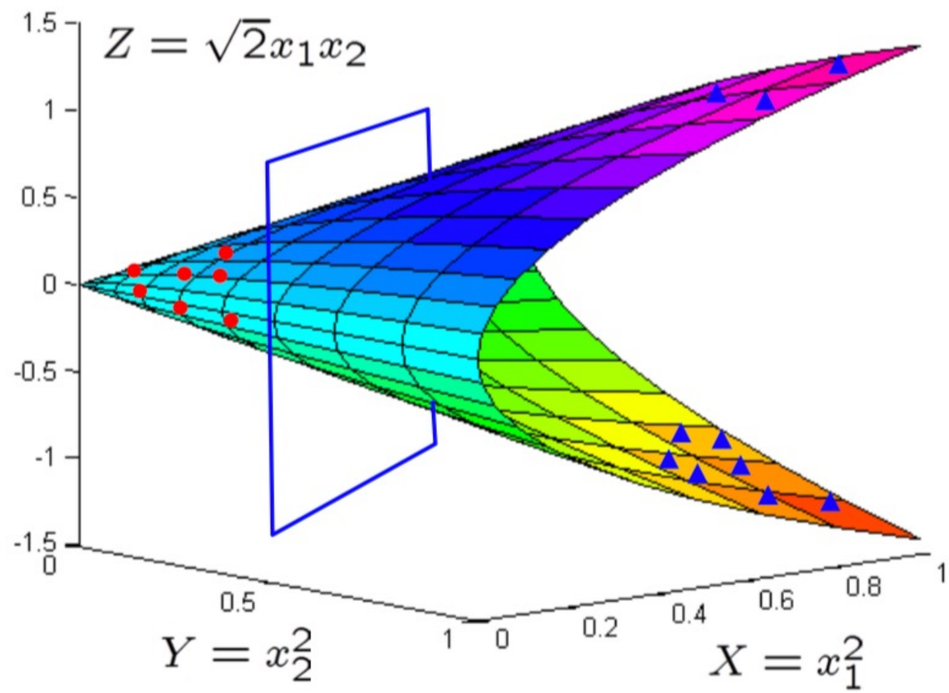
Voyons deux exemples élémentaires où des transformations non-linéaires permettent de séparer des données facilement



$$\Phi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \rightarrow \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{pmatrix} \quad \mathbb{R}^2 \rightarrow \mathbb{R}^3$$

Voyons deux exemples élémentaires où des transformations non-linéaires permettent de séparer des données facilement





$$\Phi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \to \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{pmatrix} \qquad \mathbb{R}^2 \to \mathbb{R}^3$$

… oui mais :

- On a augmenté la dimension
- On savait visuellement où chercher et on a ainsi utilisé une transformation ad hoc.

Revenons à notre formulation duale des SVM (cas séparable pour ne pas alourdir les notations)

On a $\beta = \sum_{i=1}^{n} \alpha_i y_i x_i$

et on optimise $\arg\max_{\alpha} -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_{i=1}^{n} \alpha_i$

sous les contraintes $\alpha_i \geq 0 \, , \, \forall i \in \{1, \ldots, n\}$ et $\sum_{i=1}^{n} \alpha_i y_i = 0$

Revenons à notre formulation duale des SVM (cas séparable pour ne pas alourdir les notations)

On a $\beta = \sum_{i=1}^{n} \alpha_i y_i x_i$

et on optimise $\arg\max_{\alpha} -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_{i=1}^{n} \alpha_i$

sous les contraintes $\alpha_i \geq 0 \, , \forall i \in \{1, \ldots, n\}$ et $\sum_{i=1}^{n} \alpha_i y_i = 0$

En utilisant une transformation $\phi(x)$ sur chaque coordonnée x , nous avons vu qu'il a été possible de résoudre des problèmes où une non-linéarité était nécessaire. La fonctionnelle suivante a ainsi été optimisée :

$$\arg\max_{\alpha} -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle + \sum_{i=1}^{n} \alpha_i$$

Revenons à notre formulation duale des SVM (cas séparable pour ne pas alourdir les notations)

On a $\beta = \sum_{i=1}^{n} \alpha_i y_i x_i$

et on optimise $\arg\max_{\alpha} -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_{i=1}^{n} \alpha_i$

sous les contraintes $\alpha_i \geq 0 \, , \, \forall i \in \{1, \ldots, n\}$ et $\sum_{i=1}^{n} \alpha_i y_i = 0$

En utilisant une transformation $\phi(x)$ sur chaque coordonnée x , nous avons vu qu'il a été possible de résoudre des problèmes où une non-linéarité était nécessaire. La fonctionnelle suivante a ainsi été optimisée :

$$\arg\max_{\alpha} -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle + \sum_{i=1}^{n} \alpha_i$$

Par exemple, si $x \in \mathbb{R}$ et $\phi(x) = (x, x^2)$ alors $\langle \phi(x_i), \phi(x) \rangle$ demande 3 multiplications et 1 addition alors que $\langle x_i, x \rangle$ ne demande qu'une multiplication.

→ augmentation rapide des temps de calcul avec des noyaux complexes !!!

Idée fondamentale (kernel trick) :

Pour certaines transformations, on sait que : $\langle \phi(x_i), \phi(x) \rangle = K(x_i, x)$

où K est un noyaux non-linéaire. Pour aller plus loin, on sait même que certains noyaux K respectent cette propriété sans même chercher à connaitre ϕ .

Idée fondamentale (kernel trick) :

Pour certaines transformations, on sait que : $\langle \phi(x_i), \phi(x) \rangle = K(x_i, x)$

où K est un noyaux non-linéaire. Pour aller plus loin, on sait même que certains noyaux K respectent cette propriété sans même chercher à connaitre ϕ .

Des noyaux admissibles sont par exemple :

$$K(x_i, x) = (1 + x_i^\intercal x)^s$$

$$K(x_i, x) = \tanh\left(\kappa x_i^\intercal x - \delta\right)$$

$$K(x_i, x) = \exp \frac{-(x_i^\intercal - x)^2}{2\sigma^2}$$

Idée fondamentale (kernel trick) :

Pour certaines transformations, on sait que : $\langle \phi(x_i), \phi(x) \rangle = K(x_i, x)$

où K est un noyaux non-linéaire. Pour aller plus loin, on sait même que certains noyaux K respectent cette propriété sans même chercher à connaitre ϕ .

Des noyaux admissibles sont par exemple :

$$K(x_i, x) = (1 + x_i^\intercal x)^s$$

$$K(x_i, x) = \tanh\left(\kappa x_i^\intercal x - \delta\right)$$

$$K(x_i, x) = \exp \frac{-(x_i^\intercal - x)^2}{2\sigma^2}$$

Classification non-linaire efficace en optimisant :

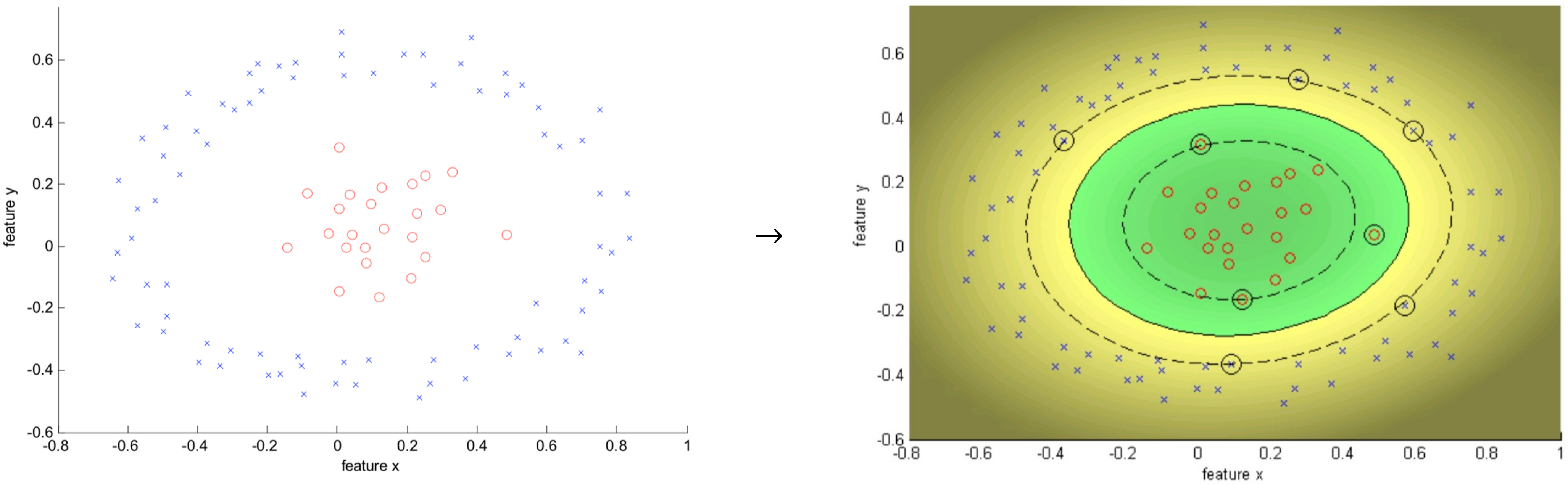
$$\arg\max_{\alpha} -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) + \sum_{i=1}^{n} \alpha_i$$

→ pratiquement le même coût et la même complexité que le problème linéaire !

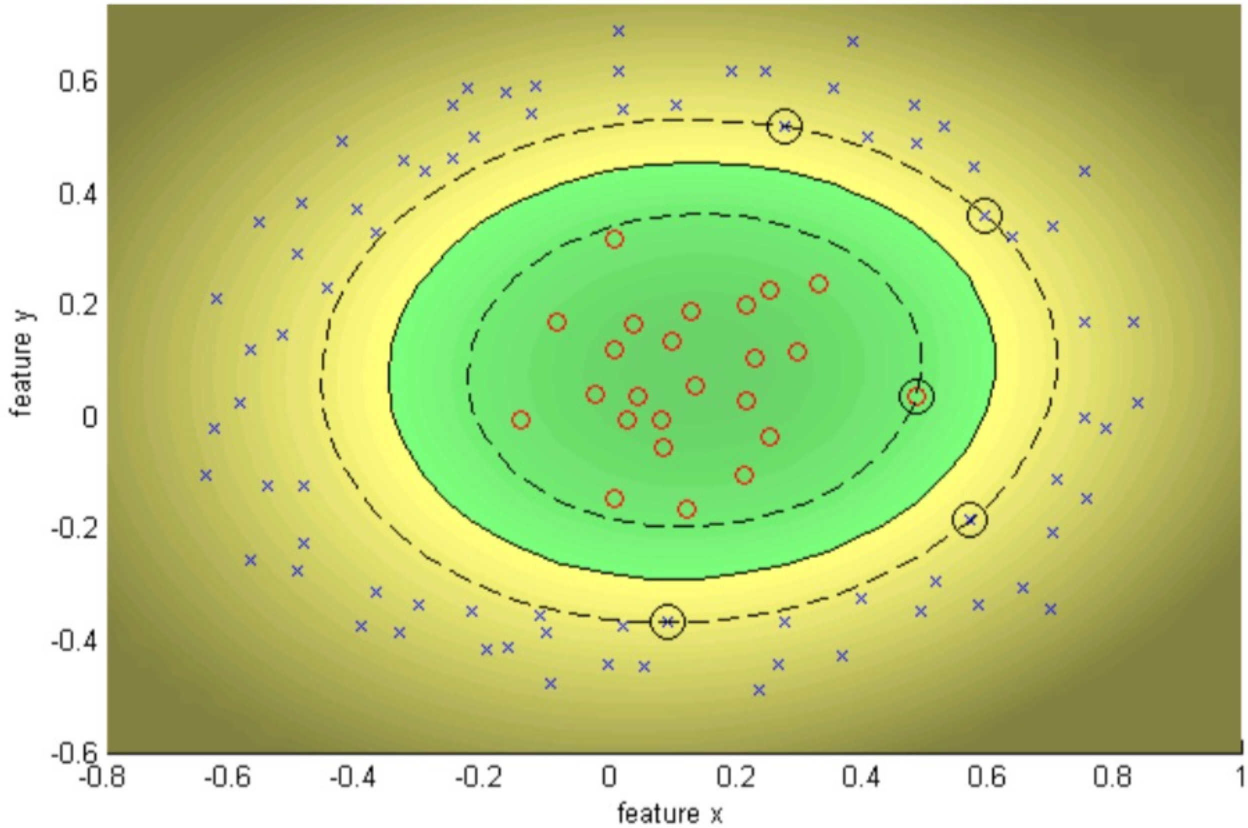
Exemple : Résolution avec un noyau Gaussien

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

avec $K(x_i, x) = \exp \dfrac{-(x_i^{\mathsf{T}} - x)^2}{2\sigma^2}$

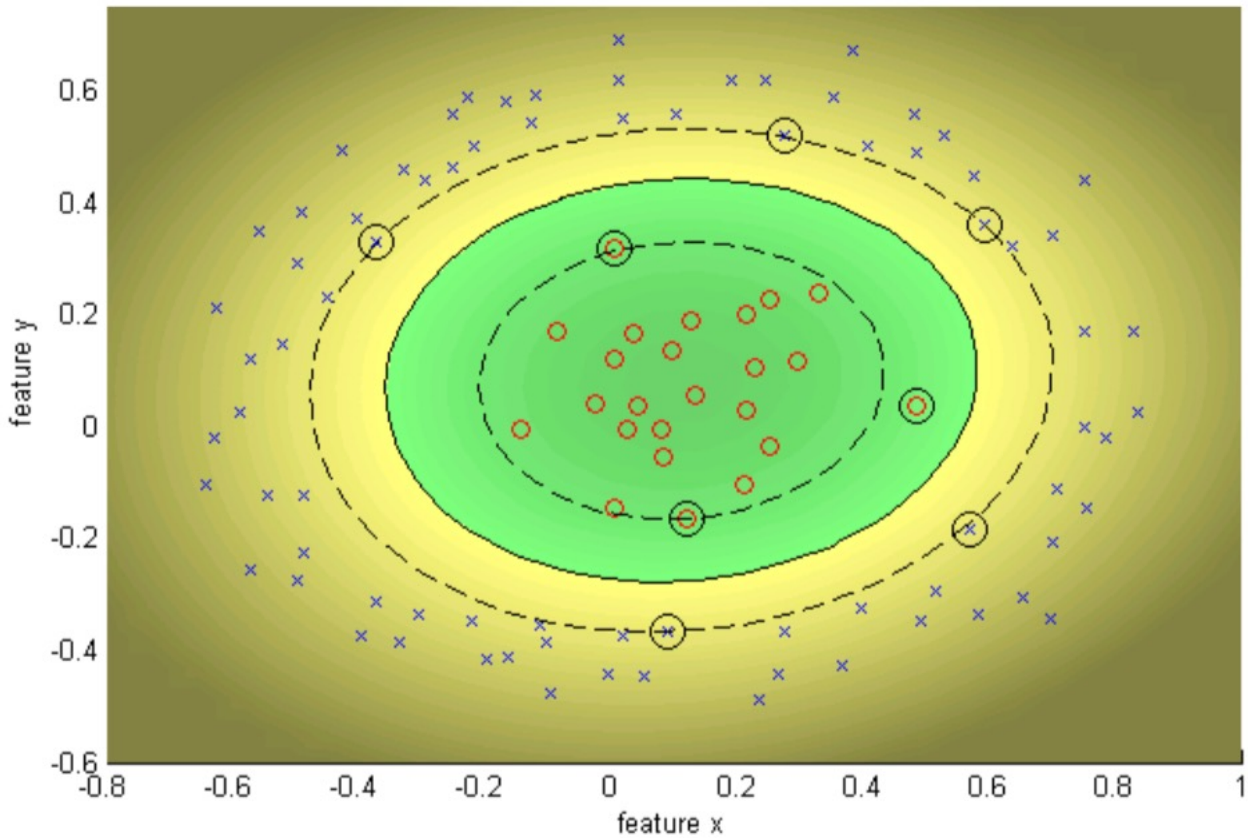


Etudions pour finir l'impact de σ et $\boldsymbol{C}$ dans le cas d'un noyaux gaussien dans le cas ci-dessous :



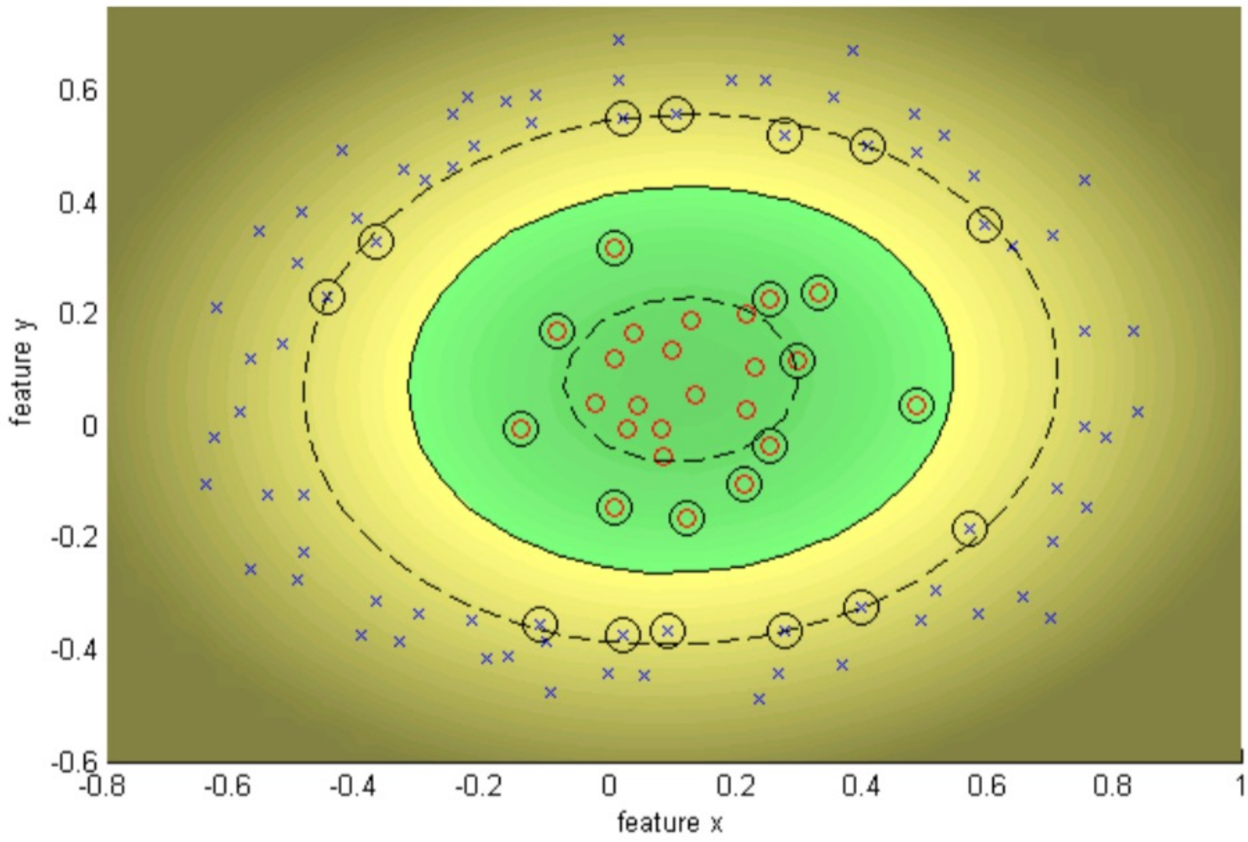
$$\boldsymbol{\sigma = 1.0 \text{ et } C = \infty}$$

Etudions pour finir l'impact de σ et $\boldsymbol{C}$ dans le cas d'un noyaux gaussien dans le cas ci-dessous :



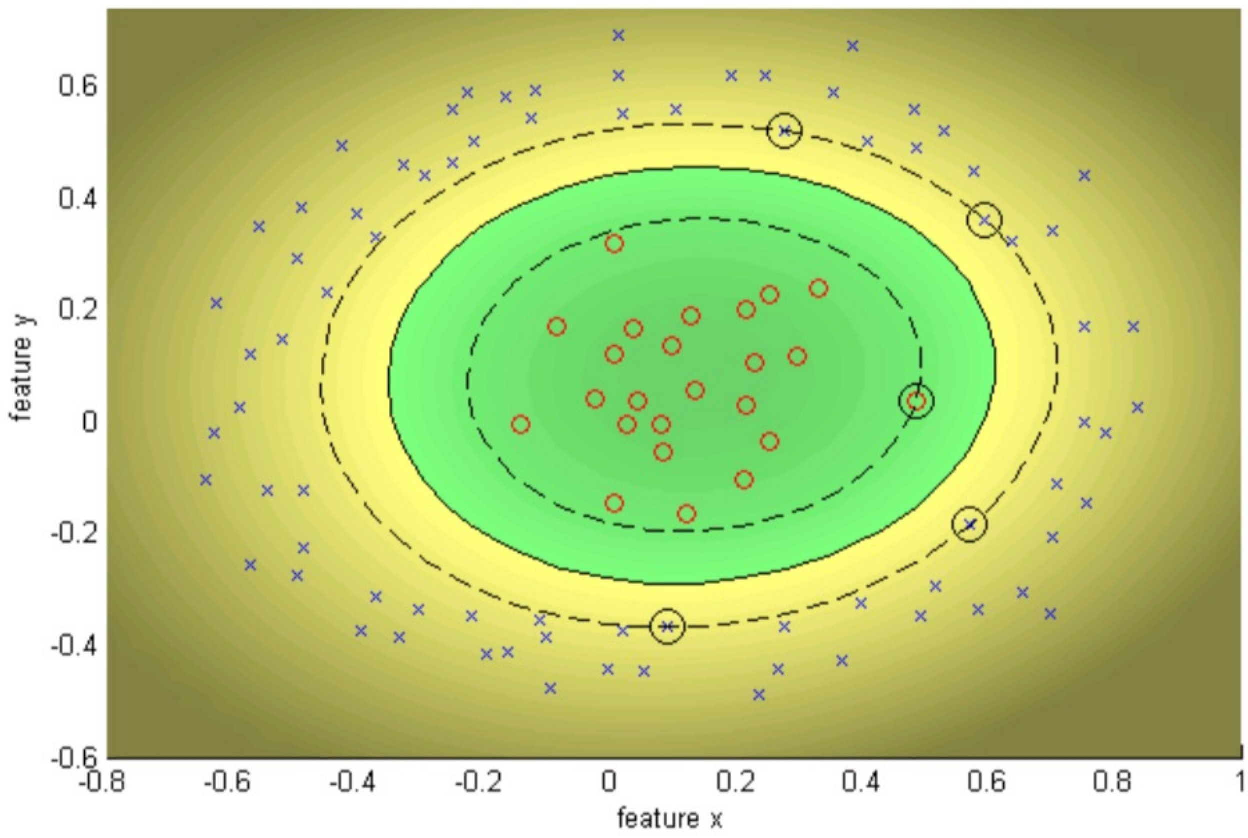
$$\boldsymbol{\sigma = 1.0 \text{ et } C = 100}$$

Etudions pour finir l'impact de σ et $\boldsymbol{C}$ dans le cas d'un noyaux gaussien dans le cas ci-dessous :



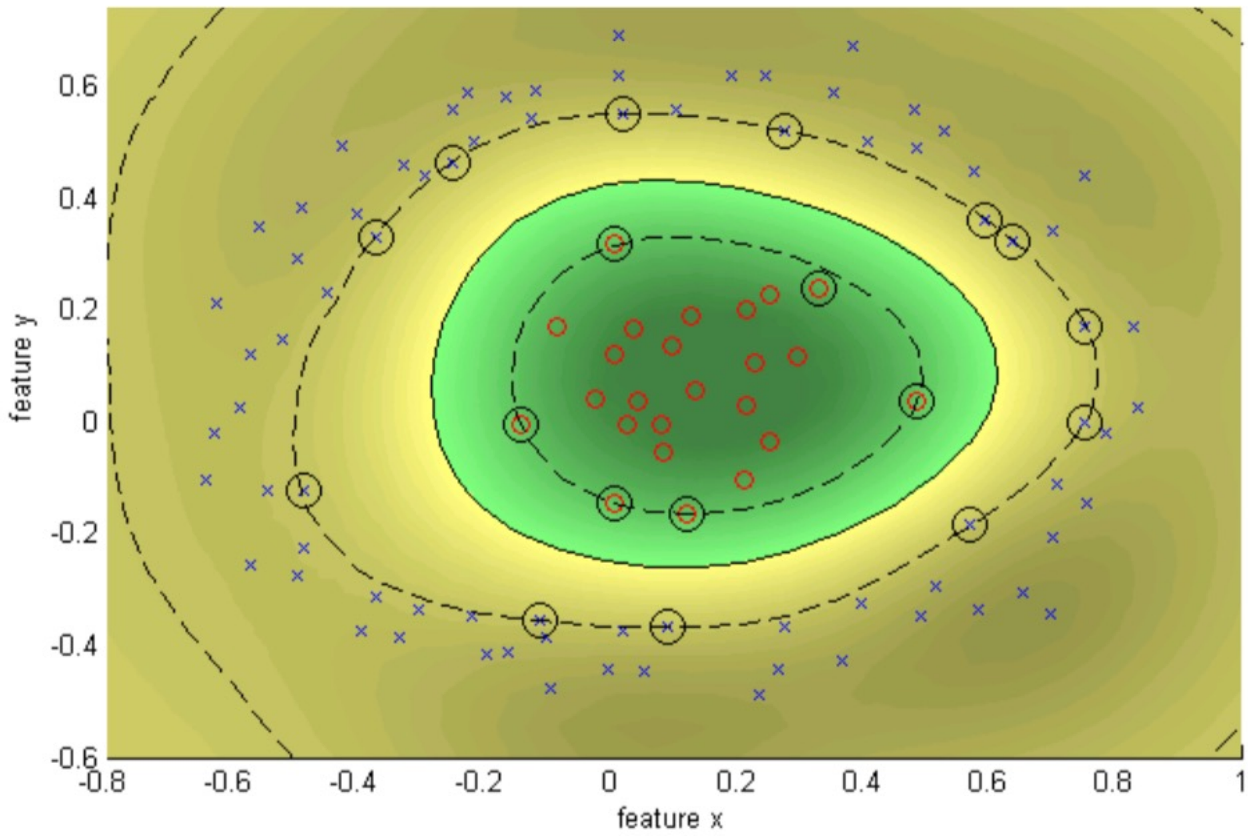
$$\boldsymbol{\sigma = 1.0 \text{ et } C = 10}$$

Etudions pour finir l'impact de $\boldsymbol{\sigma}$ et C dans le cas d'un noyaux gaussien dans le cas ci-dessous :



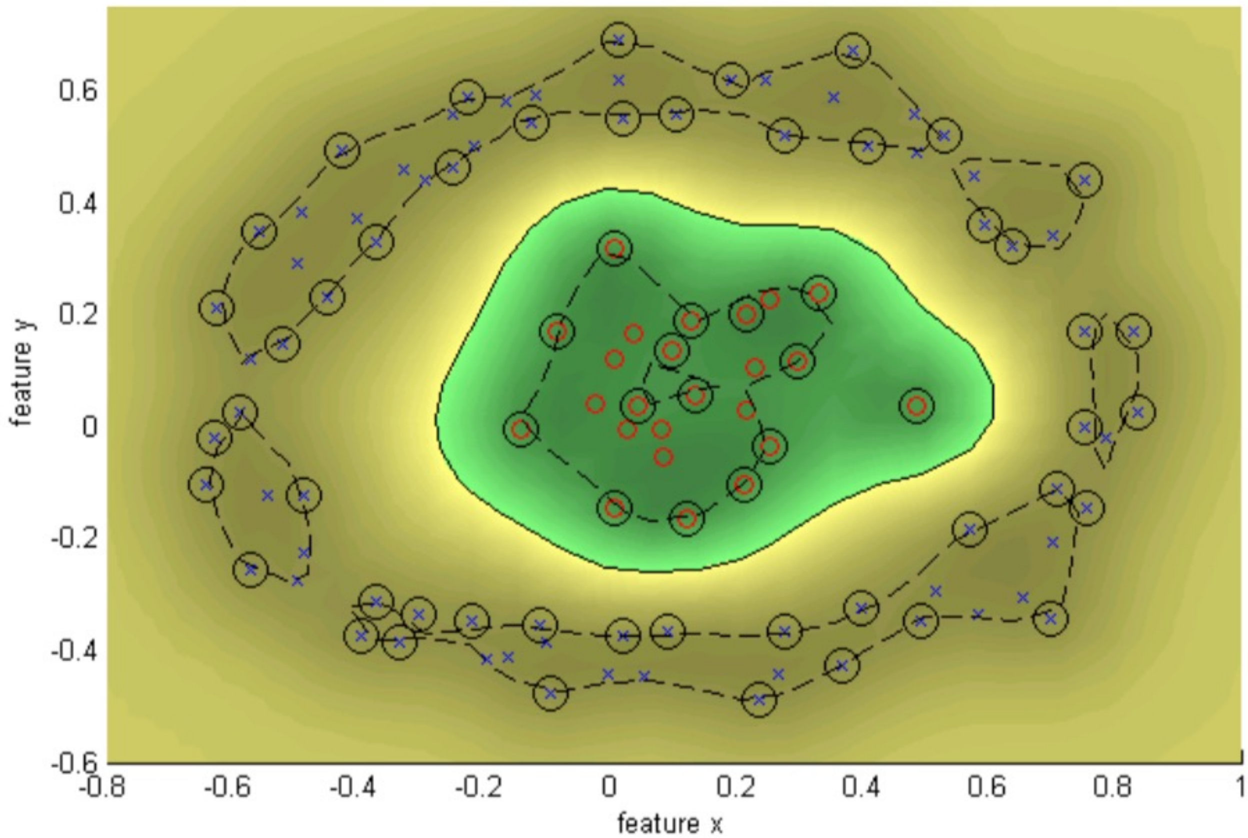
$$\boldsymbol{\sigma = 1.0 \text{ et } C = \infty}$$

Etudions pour finir l'impact de $\boldsymbol{\sigma}$ et C dans le cas d'un noyaux gaussien dans le cas ci-dessous :



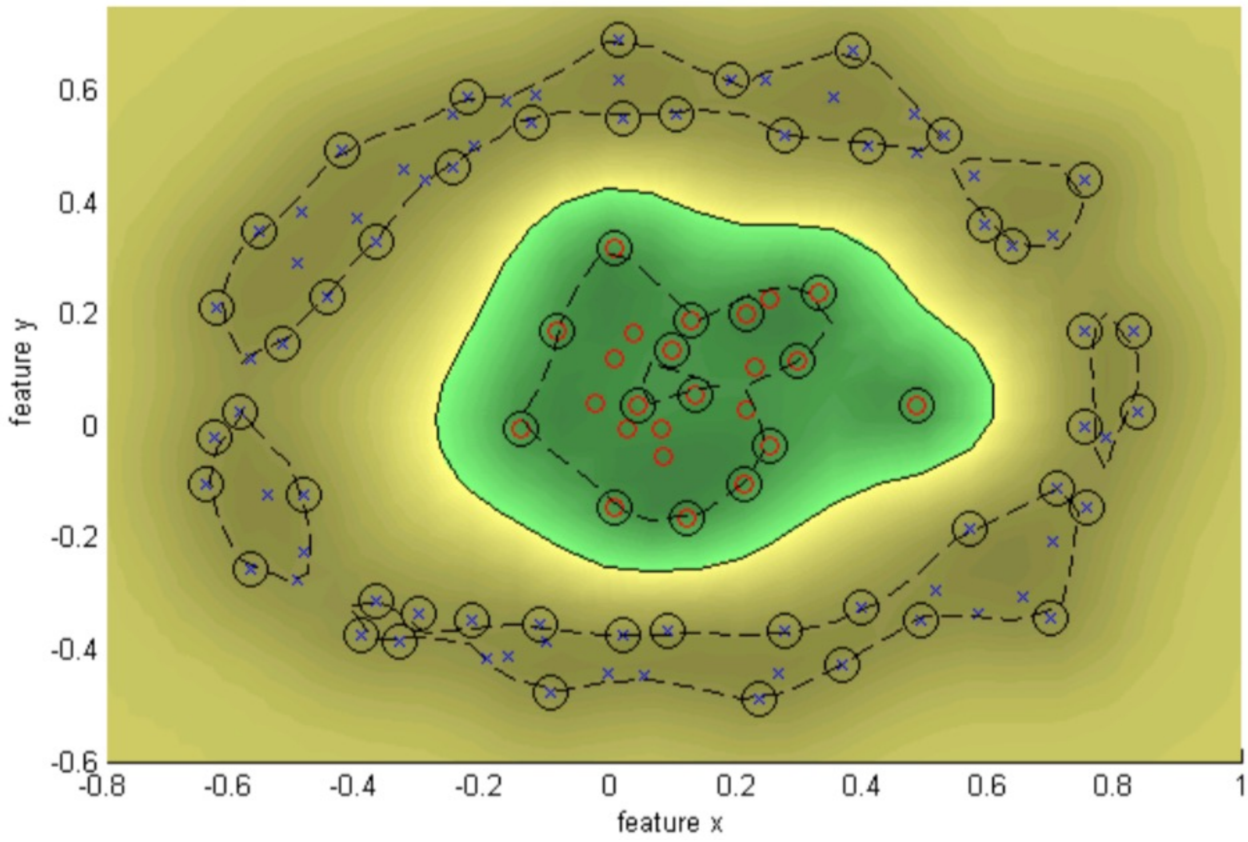
$$\boldsymbol{\sigma = 0.25 \text{ et } C = \infty}$$

Etudions pour finir l'impact de $\boldsymbol{\sigma}$ et C dans le cas d'un noyaux gaussien dans le cas ci-dessous :



$$\boldsymbol{\sigma = 0.1 \text{ et } C = \infty}$$

Etudions pour finir l'impact de $\boldsymbol{\sigma}$ et C dans le cas d'un noyaux gaussien dans le cas ci-dessous :



$$\boldsymbol{\sigma = 0.1} \textbf{ et } \boldsymbol{C = \infty}$$

Un paramétrage raisonnable est nécessaire pour ne pas sur- ou sous-apprendre…
Clairement efficace par ailleurs !!!

MERCI