# Support Vector Machines (SVM)

F. Panloup
LAREMA-Université d'Angers

—

**Cours : Apprentissage Statistique en Grande Dimension**

—

October 17, 2019

# Cadre Général

# Généralités

1. SVM : Machines à Vecteurs Supports ou Classifieurs à Vaste Séparateur de Marge (Apprentissage Supervisé)

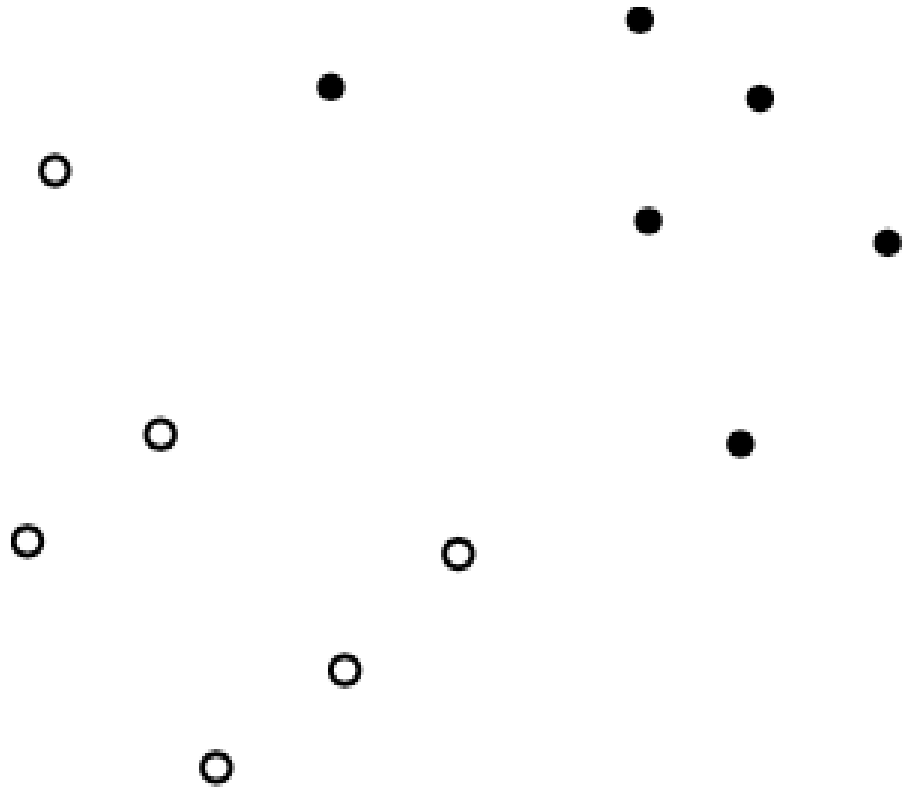
2. Famille de modèles de classification.

3. Cadre usuel : $Y$ variable binaire à valeurs dans $\{-1, 1\}$ (extensible au cadre multi-classes mais aussi au cadre quantitatif) et $\mathbf{X} = (X_1, \ldots, X_p)$, $p$ variables (explicatives).

4. Objectif : Fabriquer une règle de décision "robuste" par séparation de l'espace (en deux parties dans le cadre binaire).

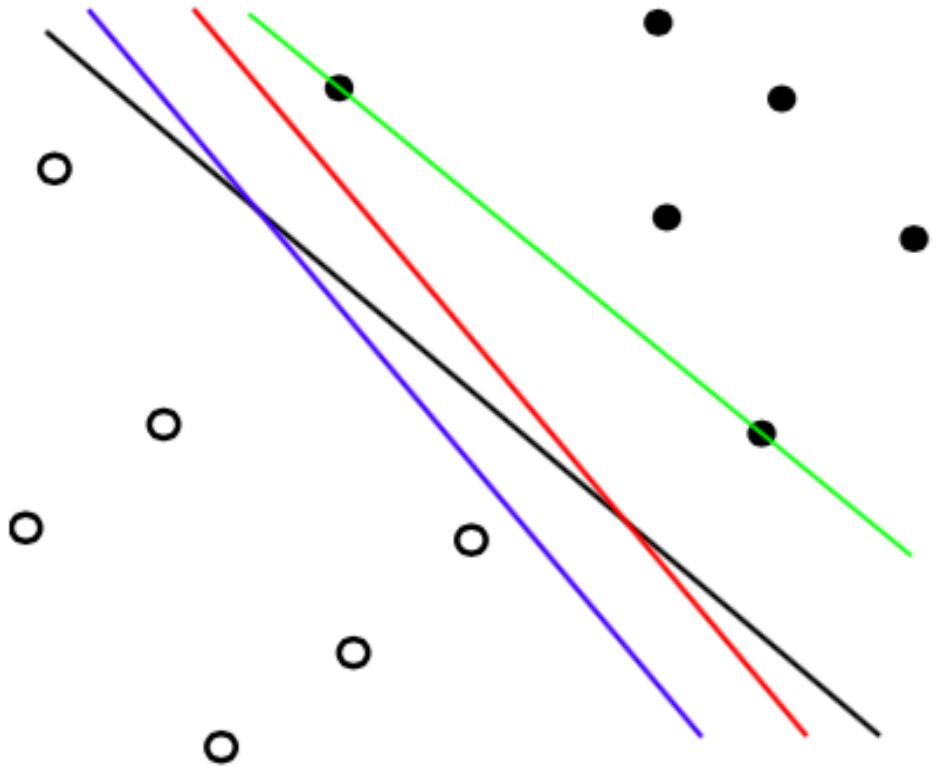
# Données linéairement séparables

**Question** : Intuitivement, comment sépareriez-vous ces points ?



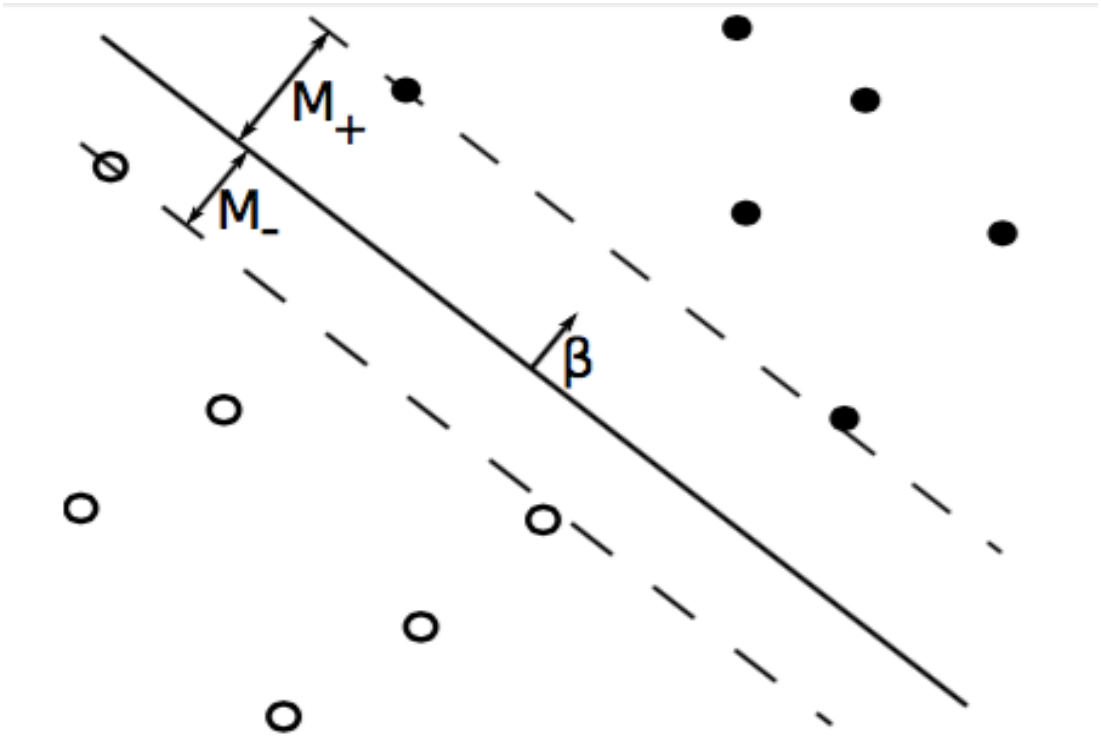
# Données linéairement séparables

**Réponse 1** : Par une droite (**séparation linéaire**). Oui, mais laquelle ?



# Données linéairement séparables

**Principe** : Dans ce cadre linéairement séparable, le principe des SVM est de fabriquer l'**hyperplan** maximisant la **marge**, *i.e.* la distance aux observations les plus proches :

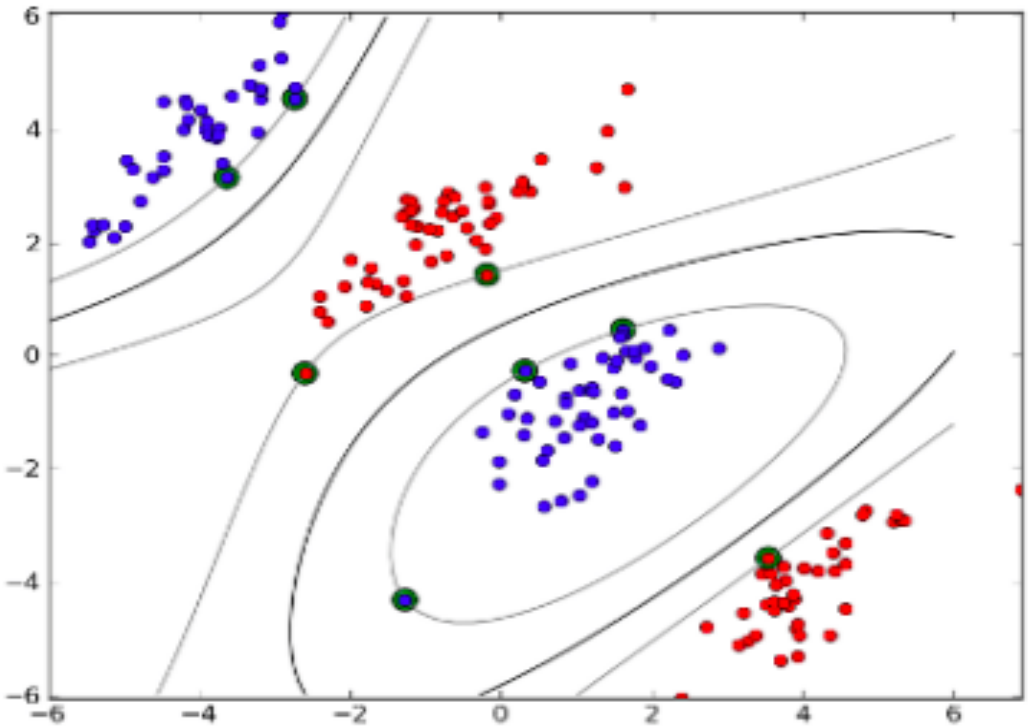


# Plan de la suite de ce chapitre

- Présentation en détail du cadre "Données Linéairement Séparables"
- Extension au cadre des données "presque linéairement séparables" (*i.e.* où la séparation par un hyperplan reste un choix efficace)
- Présentation des machines à noyau permettant d'étendre largement le principe à des données non linéairement séparables.

# Un exemple de SVM "non linéaire"

On reviendra sur ce type d'exemple plus tard mais afin de donner un aperçu de ces méthodes, voici ci-dessous un exemple de classification (hautement) non linéaire traité par SVM (**à noyau**) :

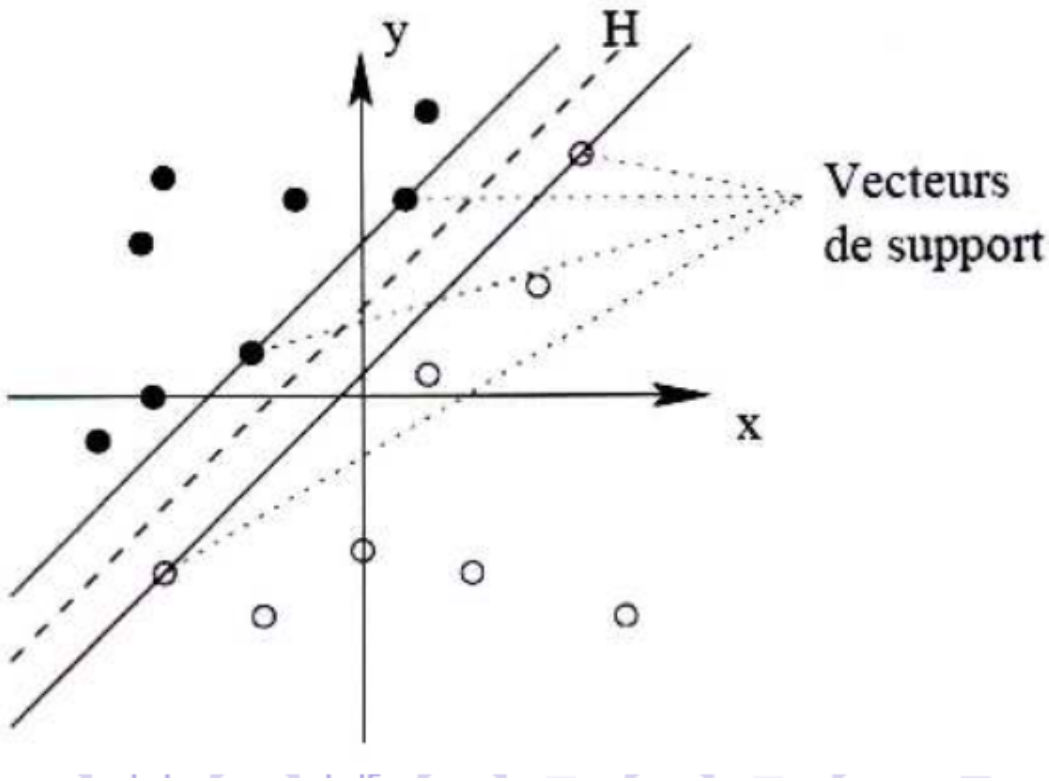


# Classifieurs linéaires à Vaste Séparateur de Marge

# Construction du SVM linéaire

**But** : Trouver un classifieur linéaire (hyperplan) qui va séparer les données et maximiser la distances entre ces 2 classes.

**Vecteurs Supports** : la détermination de l'hyperplan est réalisée (uniquement) via les vecteurs supports (points de chaque classe les plus proches).



# Hyperplans

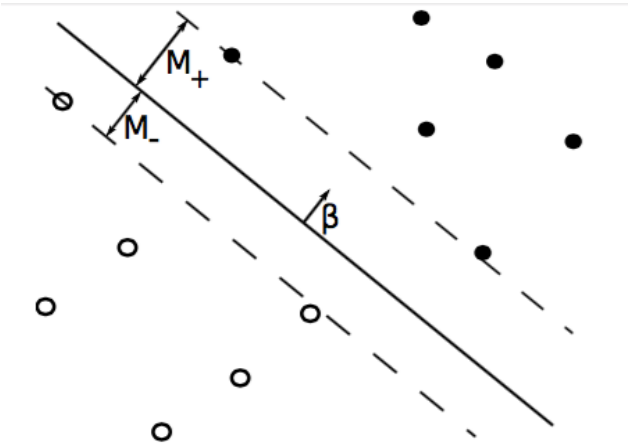
- Un hyperplan affine $\mathcal{H}$ de $\mathbb{R}^p$ est un sous-espace affine de dimension $p-1$.
- Ainsi, il est caractérisé par un point $M_0$ appartenant à $\mathcal{H}$ et un vecteur normal $\beta$ :
$$\mathcal{H} = \{M \in \mathbb{R}^p, \langle \beta, \overrightarrow{M_0 M} \rangle = 0\}$$
- Son équation s'écrit :
$$\langle \beta, x \rangle + \beta_0 = 0.$$
- L'hyperplan sépare l'espace en deux parties $\mathcal{C}_+$ et $\mathcal{C}_-$ telles que $\mathcal{C}_+ = \{x, \langle \beta, x \rangle + \beta_0 > 0\}$ et $\mathcal{C}_- = \{x, \langle \beta, x \rangle + \beta_0 < 0\}$

# Hyperplans de séparation

Supposons que $\{(x_1, y_1), \dots, (x_n, y_n)\}$ est un échantillon tel que $(x_i, y_i) \in \mathbb{R}^p \times \{-1, 1\}$. On dit que $\mathcal{H}$ d'équation $\langle \beta, x \rangle + \beta_0 = 0$ est un **hyperplan de séparation** pour l'échantillon si

$$\forall i \in \{1, \dots, N\}, \quad \langle \beta, x_i \rangle + \beta_0 \begin{cases} > 0 & \text{si } y_i = 1 \\ < 0 & \text{si } y_i = -1 \end{cases}$$

*i.e.* $\forall i \in \{1, \dots, N\}$, $y_i \left( \langle \beta, x_i \rangle + \beta_0 \right) > 0$.



# Définition du SVM (dans le cas linéairement séparable)

**Principe** : Parmi les hyperplans de séparation de l'échantillon d'apprentissage, on veut choisir celui qui a la plus **vaste marge**, i.e. telle que $\min_{i=1}^{N} d(x_i, \mathcal{H})$ est maximal. On fait ici l'hypothèse que celui-ci existe.
**Remarque** : Lorsque $\|\beta\| = 1$, $d(x_i, \mathcal{H}) = |\langle \beta, x_i \rangle + \beta_0| = y_i(\langle \beta, x_i \rangle + \beta_0)$ (Exercice). Ainsi, le problème de détermination du meilleur hyperplan pour la marge se définit comme suit :
**Hyperplan de séparation optimal** : solution du problème

$$\begin{cases} \max_{\beta, \beta_0} M \text{ sous la contrainte} \\ \forall i \in \{1, \dots, N\}, \quad y_i(\langle \beta, x_i \rangle + \beta_0) \geq M \text{ et } \|\beta\| = 1. \end{cases} \tag{1}$$

# Reformulation du problème d'optimisation

Si l'on ne contraint plus $\beta$ à être de norme 1, dans ce cas :

$$d(x_i, \mathcal{H}) = \frac{y_i(\langle \beta, x_i \rangle + \beta_0)}{\|\beta\|}$$

de sorte que la contrainte s'écrit

$$y_i(\langle \beta, x_i \rangle + \beta_0) \geq M\|\beta\|.$$

**Remarque** : On peut choisir $C = \|\beta\|$ en fonction de $M$ ! Prenons $C = 1/M$. Le problème d'optimisation peut alors se reformuler :

$$\begin{cases} \max_{\beta, \beta_0} M \text{ sous la contrainte} \\ \forall i \in \{1, \dots, N\}, \quad y_i(\langle \beta, x_i \rangle + \beta_0) \geq 1 \text{ et } \|\beta\| = \frac{1}{M}. \end{cases} \tag{2}$$

Mais dans ce cas,

$$\max_{\beta, \beta_0} M = \max_{\beta, \beta_0} \frac{1}{\|\beta\|} = \min_{\beta, \beta_0} \|\beta\|$$

# Reformulation du problème d'optimisation

de sorte que l'on peut finalement écrire le problème sous la forme :

$$\begin{cases} \min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 \text{ sous la contrainte} \\ \forall i \in \{1,\dots,N\}, \quad y_i(\langle \beta, x_i\rangle + \beta_0) \geq 1. \end{cases} \tag{3}$$

Il s'agit d'un *problème d'optimisation quadratique*. Pour le résoudre, on préfère généralement passer au "problème dual" en introduisant le lagrangien :

$$L(\beta,\beta_0,\alpha) = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{N} \alpha_i \left(y_i(\langle \beta, x_i\rangle + \beta_0) - 1\right).$$

Plus précisément, le premier problème est dit *primal* tandis que sa reformulation via le passage au Lagrangien est dite *duale*. Les $\alpha_i$ sont appelées *variables duales*.

# Problème Dual

Le problème dual consiste à déterminer $(\beta^*, \beta_0^*, \alpha^*)$ tel que

$$L(\beta^*, \beta_0^*, \alpha^*) = \max_{\alpha_1, \ldots, \alpha_N \geq 0} \min_{\beta, \beta_0} L(\beta, \beta_0, \alpha).$$

Pour cela, on minimise d'abord en $(\beta, \beta_0)$ puis on maximise en $\alpha$.

- Points critiques de $(\beta, \beta_0) \mapsto L(\beta, \beta_0, \alpha)$ solutions de $\partial_\beta L = 0$ et $\partial_{\beta_0} L = 0$ ce qui donne :

$$\begin{cases} \beta(\alpha) = \sum_{i=1}^N \alpha_i y_i x_i \\ \sum_{i=1}^N \alpha_i y_i = 0. \end{cases}$$

- Si on réinjecte ces conditions dans $L$, on trouve (voir TD pour détails) :

$$\tilde{L}(\alpha) = L(\beta(\alpha), \beta_0(\alpha), \alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle.$$

# Problème Dual

Le problème dual s'écrit alors

$$\begin{cases} \min_{\alpha \in (\mathbb{R}^+)^N} \frac{1}{2} \alpha^T G \alpha - \langle \alpha, \mathbf{e} \rangle \\ \text{avec } \langle \mathbf{y}, \alpha \rangle = 0. \end{cases}$$

où $G$ est la matrice définie par $G_{i,j} = y_i y_j \langle x_i, x_j \rangle$, $\mathbf{e} = (1, \ldots, 1)^T$ et $\mathbf{y} = (y_1, \ldots, y_N)^T$. On a donc deux formulations (primale et duale) du problème d'optimisation qui du point de vue numérique sont concurrentes (selon que $p < n$ ou $p \geq n$ principalement), voir cours d'Optimisation de M1 pour plus de détails sur les problèmes d'optimisation quadratique.

## $\beta^*$, $\beta_0^*$

On suppose maintenant que l'on a réussi à approcher $\alpha^*$ défini par

$$\alpha^* = (\alpha_1^*, \dots, \alpha_N^*) = \text{Argmax}_{\alpha \in \mathbb{R}^N} \tilde{L}(\alpha).$$

Comment retrouver ensuite l'équation de l'hyperplan de séparation optimal ?

- Par construction, $\beta^* = \sum_{i=1}^N \alpha_i^* y_i x_i$.
- Pour $\beta_0^*$, on s'appuie sur les *conditions complémentaires de KKT* qui garantissent que

$$\forall i \in \{1, \dots, N\}, \quad \alpha_i^* (y_i (\langle \beta^*, x_i \rangle + \beta_0^*) - 1) = 0.$$

En d'autres termes, soit le coefficient $\alpha_i^*$ est nul, soit le point considéré atteint la frontière. Dans ce cas, c'est un vecteur support. On peut donc retrouver $\beta_0^*$ en considérant un $(x_i, y_i)$ vecteur support (il en existe au moins 2). On a alors

$$\beta_0^* = y_i - \langle \beta^*, x_i \rangle.$$

**Remarque :** En prenant un vecteur support "positif" et un "négatif" (voir TD), on peut aussi obtenir la formule :

$$\beta_0^* = -\frac{1}{2} \left( \min_{y_i=1} \langle \beta^*, x_i \rangle + \min_{y_i=-1} \langle \beta^*, x_i \rangle \right).$$

# Règle de Décision

A l'issue de ces calculs, on a donc fabriqué un **prédicteur** $\hat{f}$ de la classe d'un point $x$ :

$$\hat{f}(x) = \mathrm{sgn}(\langle \beta^*, x \rangle + \beta_0^*) = \mathrm{sgn}(\sum_{i=1}^{N} y_i \alpha_i^* \langle x_i, x \rangle + \beta_0^*).$$

**Remarques**

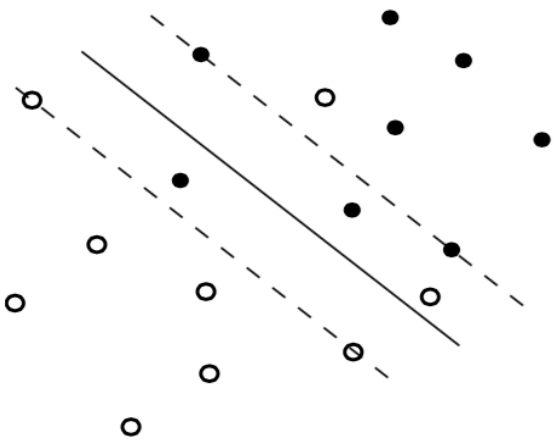
- Marge ? Pour rappel, l'idée était de fabriquer l'hyperplan de séparation ayant la marge maximale. Si l'on revient à la construction du problème d'optimisation, on constate que la marge $M^*$ associée à l'hyperplan optimal est donnée par :

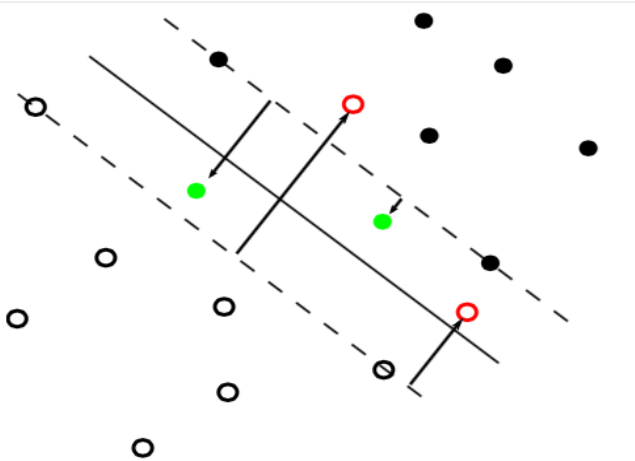
$$M^* = \frac{1}{\|\beta^*\|}.$$

- Comme d'habitude, si l'on veut étudier l'erreur de classification de cet algorithme de prévision, on fabriquera l'hyperplan de séparation sur un sous-ensemble d'entraînement et on calculera l'erreur de classification sur un échantillon test.
- En pratique, les données sont rarement linéairement séparables. On va donc dans la suite voir comment injecter de la "souplesse " dans cet algorithme.

# SVM linéaires à marge flexible

# SVM linéaire pour données non linéairement séparables ?





**Question** : Comment fabriquer un classifieur linéaire basé sur la même approche lorsqu'il n'existe pas d'hyperplan de séparation ? (ou lorsqu'il en existe un mais qu'il est clairement trop dépendant de données *outliers*)

**Réponse** : Fabriquer un SVM linéaire à marge souple : La condition $y_i \left( \langle \beta, x_i \rangle + \beta_0 \right) \geq 1$ est remplacée par $y_i \left( \langle \beta, x_i \rangle + \beta_0 \right) \geq 1 - \xi_i$ avec $\xi_i \geq 0$.

- Si $\xi_i \in ]0, 1]$, le point est bien classé mais "sous" la marge du SVM précédent.
- Si $\xi_i > 1]$, le point est mal classé (voir figure en haut à droite).

# Problème d'optimisation associé (aux marges flexibles)

- Si l'on fait trop souvent recours à $\xi_i > 0$ voir grand, cela signifie que le classifieur qui serait créé n'est pas très bon.

- Ainsi, l'idée est de pénaliser la variable $\xi_i$ de la manière suivante : on résout ici

$$\begin{cases} \min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^n \xi_i \text{ sous les contraintes} \\ \forall i \in \{1,\dots,N\}, \quad y_i(\langle \beta, x_i\rangle + \beta_0) \geq 1 - \xi_i \\ \xi_i \geq 0. \end{cases} \tag{4}$$

où $C$ est un paramètre à régler (*tuning parameter*). Il règle le taux de souplesse. Plus $C$ est grand, moins l'on "tolère" les erreurs de classification. On est donc dans un paradigme "Biais-Variance" : il faut trouver un compromis entre le surapprentissage qui guette lorsque $C$ est petit et le fort biais lorsque $C$ est grand (Choix du $C$ par Validation Croisée ?)

# Résolution du problème aux marges flexibles

Là encore, l'idée est de passer à la formulation duale. Le Lagrangien s'écrit ici :

$$L(\beta, \beta_0, \xi, \alpha, \mu) = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i \left(y_i(\langle \beta, x_i \rangle + \beta_0) - (1 - \xi_i)\right) - \sum_{i=1}^{N} \mu_i \xi_i.$$

Le deuxième terme permet de gérer la contrainte $\xi_i \geq 0$, $i \in \{1, \ldots, N\}$. A nouveau, on peut écrire les conditions KKT :

$$\text{Stationnarité} : \begin{cases} \partial_\beta L = 0 \implies \beta = \sum_{i=1}^{N} \alpha_i y_i x_i, \\ \partial_{\beta_0} L = 0 \implies \sum_{i=1}^{N} \alpha_i y_i = 0, \\ \partial_\xi L = 0 \implies \alpha_i = C - \mu_i, \end{cases} \tag{5}$$

de sorte qu'en réinjectant dans $L$, on a encore à maximiser

$$\tilde{L}(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

sous contrainte $\sum_{i=1}^{N} \alpha_i y_i = 0$ et $0 \leq \alpha_i \leq C$.

# Vecteurs supports

Notons $(\alpha^*, \beta^*, \beta_0^*, \xi^*, \mu_i^*)$ la solution du problème. Les conditions complémentaires de KKT donnent

$$\forall i \in \{1, \dots, N\}, \begin{cases} \alpha_i^* \left( \langle \beta^*, x_i \rangle + \beta_0^* \right) - (1 - \xi_i^*)) = 0, \\ \mu_i^* \xi_i^* = 0 \iff \xi_i^* (\alpha_i^* - C) = 0. \end{cases}$$

Les vecteurs supports sont à nouveau les $x_i$ pour lesquels $\alpha_i^* > 0$. Néanmoins, on a deux classes :

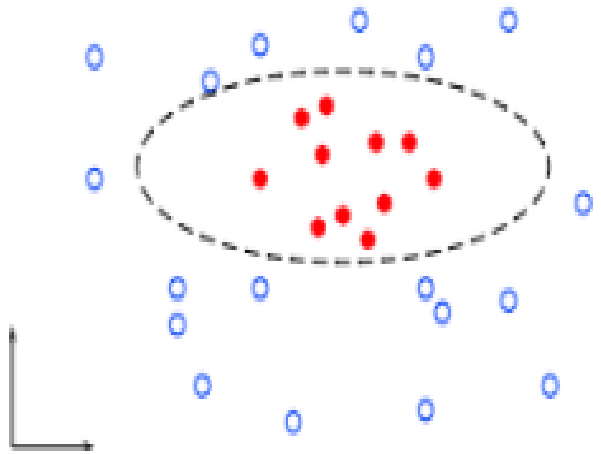
1. Ceux pour lesquels $\xi_i^* = 0$ (on ne fait pas usage de la flexibilité : vecteur supports du précédent modèle).
2. Ceux pour lesquels $\xi_i^* > 0$. Dans ce cas, $\alpha_i^* = C$. Ces points sont les nouveaux vecteurs supports issus de la flexibilisation du modèle. On peut remarquer qu'ils sont plus "suspects", *i.e.* leur présence dans la construction du prédicteur est la plus susceptible d'être remise en cause.

# Noyaux

**Constat :** Les SVMs linéaires constituent des classifieurs linéaires robustes. Néanmoins, la contrainte de linéarité est clairement restreignante. Dans un cadre où les points ont une répartition dans l'espace ne permettant absolument pas une séparation par un hyperplan, que faire ?
**Question :** Peut-on adapter les approches précédentes pour fabriquer des *variétés* de dimension $p-1$ permettant de séparer les points de la même manière que les SVMs linéaires ?
**Réponse :** Oui, en introduisant des noyaux. L'idée est de déformer l'espace ou plus précisément les vecteurs d'entrée $x_i$ dans un nouvel espace appelé **espace de représentation** (**feature space**). Cette idée est due à Boser, Guyon et Vapnik (1992).

# Exemple

Considérons $p = \mathbb{R}^2$ et l'ensemble de points ci-dessous.



L'ensemble des points ci-dessus n'est clairement pas linéairement séparable. Néanmoins, si l'on note $\phi(x) = \phi(a, b) = (a^2, b^2, a, b)$, alors, l'ensemble des points $\{\phi(x_i), y_i, i = 1, \ldots, N\}$ l'est dans $E = \mathbb{R}^4$.

# Remarques Importantes

Considérons $\phi : \mathbb{R}^p \mapsto E$ où $E$ est l'espace de représentation et supposons que $E$ est muni d'un produit scalaire $\langle\ ,\ \rangle_E$. On remarque qu'en suivant le raisonnement précédent (dans le cas flexible), le problème d'optimisation

$$\tilde{L}(\alpha) = \sum_{i=1}^{N} \alpha_i - \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle_E$$

sous les contraintes $\sum_{i=1}^{N} \alpha_i y_i = 0$ et $0 \leq \alpha_i \leq C$, peut encore être résolu par la même méthode et mène au meilleur hyperplan dans $E$ pour les points $\{\phi(x_i), y_i, i = 1, \ldots, N\}$. La règle de décision associée est :

$$\hat{f}(x) = \operatorname{sgn}\left( \sum_{i=1}^{N} y_i \alpha_i^* \langle \phi(x_i), \phi(x) \rangle_E + \beta_0^* \right).$$

**Remarque** : On n'a pas besoin de fabriquer $\phi$ explicitement puisqu'on ne fait usage que de $\langle \phi(x), \phi(x') \rangle$. On a simplement besoin de $k(x, x') = \langle \phi(x), \phi(x') \rangle$.

# Noyaux

La fonction $k$ est appelé un noyau :

## Definition

Une fonction $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ telle que $k(x, x') = \langle \phi(x), \phi(x') \rangle_E$ pour une fonction $\phi : \mathcal{X} \to E$ est appelée un noyau.

Par exemple, $k(x, x') = \langle x, x' \rangle^2_{\mathbb{R}^2}$ est un noyau. Un espace $E$ associé est $\mathbb{R}^3$ avec la fonction $\phi$ définie par $\phi(x) = \phi(a, b) = (a^2, \sqrt{2}ab, b^2)$.

## Proposition

*(Condition de Mercer). Si la fonction $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est continue symétrique et si pour tous points $x_1, \ldots, x_k$, $k(x_i, x_j)_{1 \leq i,j \leq k}$ est définie positive (sur $\mathbb{R}^k$), alors il existe un espace de Hilbert $E$ et une fonction $\phi : \mathcal{X} \to E$ tels que $k(x, x') = \langle \phi(x), \phi(x') \rangle_E$. (L'espace $E$ est appelé espace à noyau reproduisant.)*

# Exemples de Noyaux

Supposons que $\mathcal{X} = \mathbb{R}^p$. Donnons quelques exemples de noyaux usuels.

- Noyau polynomial de degré $r$ : $k(x, x') = (c_0 + \gamma\langle x, x'\rangle)^r$ (où $\langle x, x'\rangle$ est le produit scalaire usuel sur $\mathbb{R}^p$).
- Noyau Gaussien : $k(x, x') = \exp\left(-\gamma\|x - x'\|^2\right)$.
- Noyau de Laplace : $k(x, x') = \exp\left(-\gamma\|x - x'\|\right)$.
- Noyau Sigmoïde : $k(x, x') = \tanh(\gamma\langle x, x'\rangle + c_0)$.

**Remarques** : Les noyaux Gaussiens et Laplaciens sont dits radiaux (dépendent de la norme de la différence) tandis que les deux autres basés sur le produit scalaire sont dits projectifs.
L'espace $E$ n'est pas précisé. Dans le cas gaussien, l'espace $E$ est un espace de fonctions (en particulier de dimension infinie).

# Conclusion

- Objectif du cours : introduction aux SVM.
- Dans le cadre à noyau, l'idée est de plonger la variable $x$ dans un espace plus gros (beaucoup plus gros, voir de dimension infinie) dans lequel les points sont linéairement séparables puis d'utiliser les méthodes développées dans le cadre linéaire.
- Sujet très large. Nombreuses extensions/variations (selon les contextes), résultats théoriques sur le sujet.
- Parmi ces extensions, on peut en particulier noter que les SVM s'appliquent dans un cadre **multi-classes** et même dans le cadre de la **régression**.