# Machines à vecteurs de support.

Master 2 MMA - 2022/2023

29 novembre 2022

# Discrimination linéaire.

- On se place dans le cas de la classification supervisée, binaire : $Y_i \in \{-1, 1\}$, et avec des données quantitatives : $X_i \in \mathbb{R}^P$.
- Beaucoup de méthode de classification commencent par définir une **fonction discriminante** $f : \mathcal{X} \to \mathbb{R}$, pour ensuite définir le classifieur

$$t(X) = \begin{cases} -1 \text{ si } f(X) < 0 \\ 1 \text{ si } f(X) \geq 0 \end{cases}$$

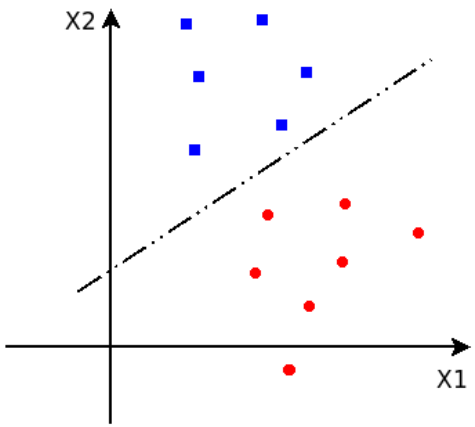
- Un cas classique est celui des **classifieurs linéaires** : le fonction $f$ est alors de la forme

$$f(X) = w_0 + \sum_{p=1}^{P} w_p X^p = w_0 + \langle w, X \rangle$$

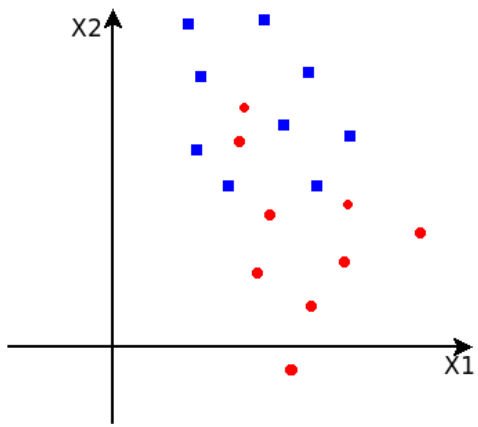
c'est-à-dire que la fonction discriminante $f$ est une fonction affine (et pas linéaire !). Le choix des coefficients $w_p$, $0 \leq p \leq P$ définit alors la fonction $f$ et donc le classifieur.

# Données linéairement séparables.

- Utiliser un classifieur linéaire est approprié lorsque les données $X_i$ sont **linéairement séparables**. Géométriquement, ceci signifie que les $X_i$ se trouvent de part et d'autre d'un **hyperplan séparateur** (une droite si $P = 2$, un plan si $P = 3$, etc.).




cas séparable

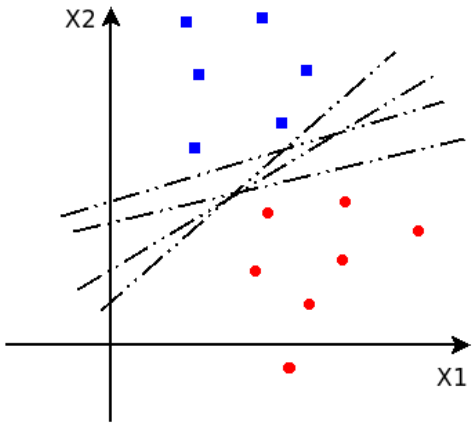


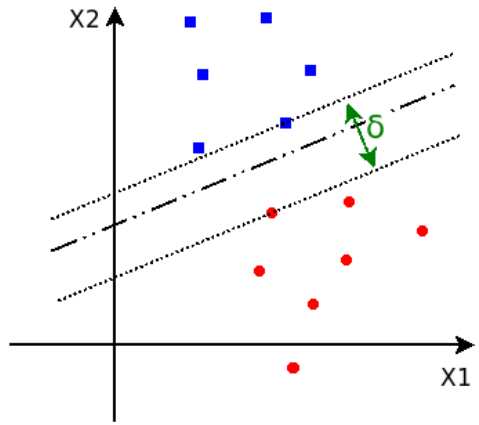

cas non séparable

- Pour des données linéairement séparables, il est logique de chercher une fonction discriminante affine $f$ qui sépare les données, c'est-à-dire telle que $f(X) = 0$ soit l'équation d'un hyperplan séparateur.

# Choix de l'hyperplan séparateur.

- Lorsque des données sont linéairement séparables, il y a en général une infinité d'hyperplans séparateurs.



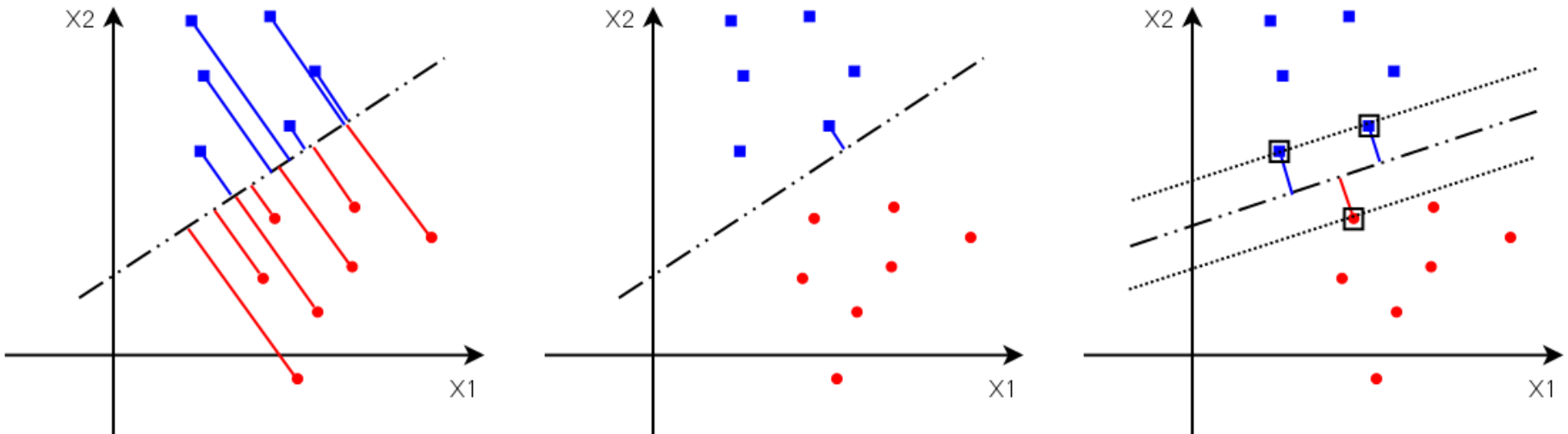

- La **marge** $\delta$ d'un hyperplan séparateur est la largeur de la bande maximale parallèle à l'hyperplan et ne contenant aucune observation.
- La méthode des **machines à vecteurs de support**, dans le cas linéaire, consiste à trouver l'hyperplan séparateur qui **maximise la marge**.

# SVM linéaire : formulation mathématique.

- remarque : $f$ discriminante, donc $Y_i\, f(X_i) \geq 0$.
- Pour calculer la marge d'un hyperplan, on doit calculer la distance d'une observation $X$ à l'hyperplan $H_w$ défini par les coefficients $w_p$ :

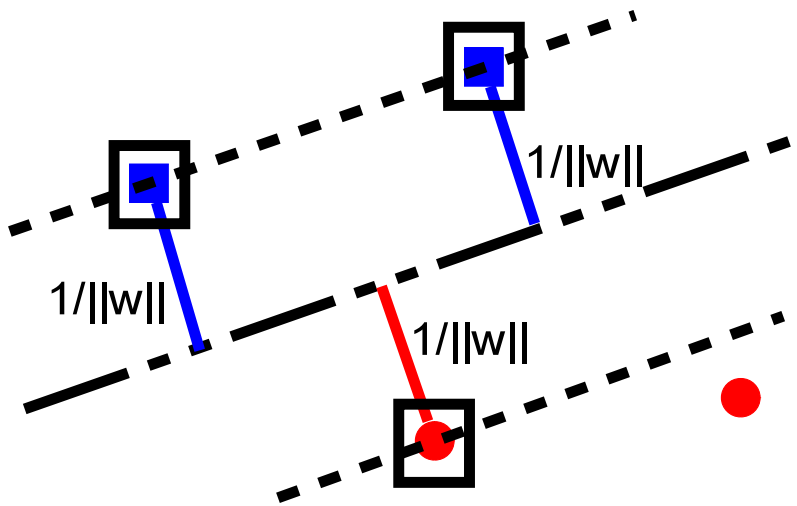
$$d(X, H_w) = \frac{\left| w_0 + \sum_{p=1}^{P} w_p X^p \right|}{\sqrt{\sum_{p=1}^{P} w_p^2}} = \frac{Y_i(w_0 + \langle w, X \rangle)}{\|w\|}$$

- En maximisant la distance minimale des observations à l'hyperplan, on obtient l'hyperplan de marge maximale $\hat{\delta} = 2 \max_w \min_i d(X_i, H_w)$.



# SVM linéaire : formulation mathématique.

- Les coefficients $w_p$ optimaux sont définis à une constante multiplicative près. On peut choisir cette constante de telle manière que la marge optimale soit égale à $\hat{\delta} = 2/\|\hat{w}\|$, ou autrement dit, telle que les distances minimales de l'hyperplan aux observations soient égales à $1/\|\hat{w}\|$.




- Les observations $X_i$ telles que $d(X_i, H_{\hat{w}}) = 1/\|\hat{w}\|$ sont appelées **vecteurs de support**.

# SVM linéaire : formulation mathématique.

- Avec cette convention de normalisation, les observations vérifient toujours $d(X_i, H_{\hat{w}}) \geq 1/\|\hat{w}\|$, soit encore $Y_i(w_0 + \langle w, X_i \rangle) \geq 1$.
- On obtient finalement la formulation suivante du problème de maximisation de la marge :

$$\begin{cases} \text{Maximiser } \delta(w_0, w) = 2/\|w\| \\ \text{sous les contraintes } Y_i(w_0 + \langle w, X_i \rangle) \geq 1 \end{cases}$$

- On préfère l'écrire sous la forme d'un problème de minimisation :

$$\begin{cases} \text{Minimiser } J(w_0, w) = \dfrac{\|w\|^2}{2} \\ \text{sous les contraintes } Y_i(w_0 + \langle w, X_i \rangle) \geq 1 \end{cases}$$

# Dualité en optimisation.

Théorie de la **dualité lagrangienne** en optimisation sous contraintes : sous certaines hypothèses sur la fonction coût $J(x)$ et les fonctions $g_i(x)$ définissant les contraintes, il y a équivalence entre les formulations **primale** et **duale** suivantes :

- Formulation primale : $\begin{cases} \text{Minimiser } J(x) \\ \text{sous les contraintes } g_i(x) \leq 0 \end{cases}$
- Formulation duale : $\begin{cases} \text{Maximiser } \psi(\mu) = \text{argmin}_x L(x, \mu) \\ \text{sous les contraintes } \mu_i \geq 0 \end{cases}$

  où $L$ est le **Lagrangien** du problème : $L(x, \mu) = J(x) + \sum_i \mu_i g_i(x)$
- Les coefficients $\hat{\mu}_i$ optimaux vérifient la propriété suivante : si $g_i(x) < 0$ (contrainte dite inactive), alors $\hat{\mu}_i = 0$

# SVM linéaire : formulation duale.

Dans le contexte des SVM linéaires on a :

- Formulation primale :

  Minimiser $J(w_0, w) = \dfrac{\|w\|^2}{2}$ sous les contraintes $Y_i\,(w_0 + \langle w, X_i\rangle) \geq 1$

- Formulation duale :

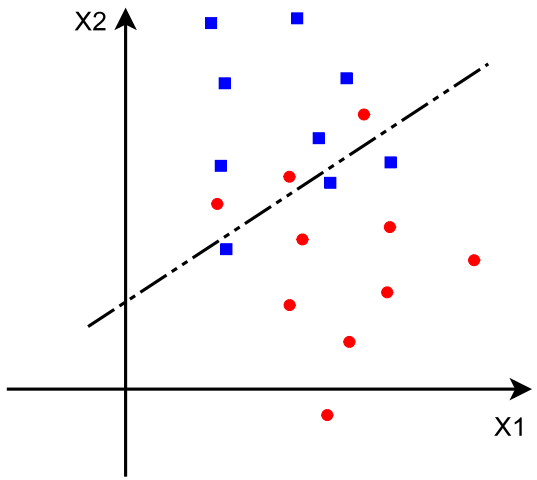
$$\begin{cases} \text{Maximiser } \psi(\mu) = \displaystyle\sum_{i=1}^{n} \mu_i - \frac{1}{2} \sum_{i,j} \mu_i \mu_j \; Y_i Y_j \; \langle X_i, X_j \rangle \\ \text{sous les contraintes } \mu_i \geq 0 \text{ et } \displaystyle\sum_{i=1}^{n} \mu_i Y_i = 0 \end{cases}$$

- Si on résout le problème dual, on retrouve le vecteur $\hat{w}$ des coefficients optimaux via la formule

$$\hat{w} = \sum_{i=1}^{n} \hat{\mu}_i Y_i X_i = \sum_{i, X_i \text{ support}} \hat{\mu}_i Y_i X_i$$

# SVM linéaire : cas non séparable.

- cas non séparable : relaxation des contraintes

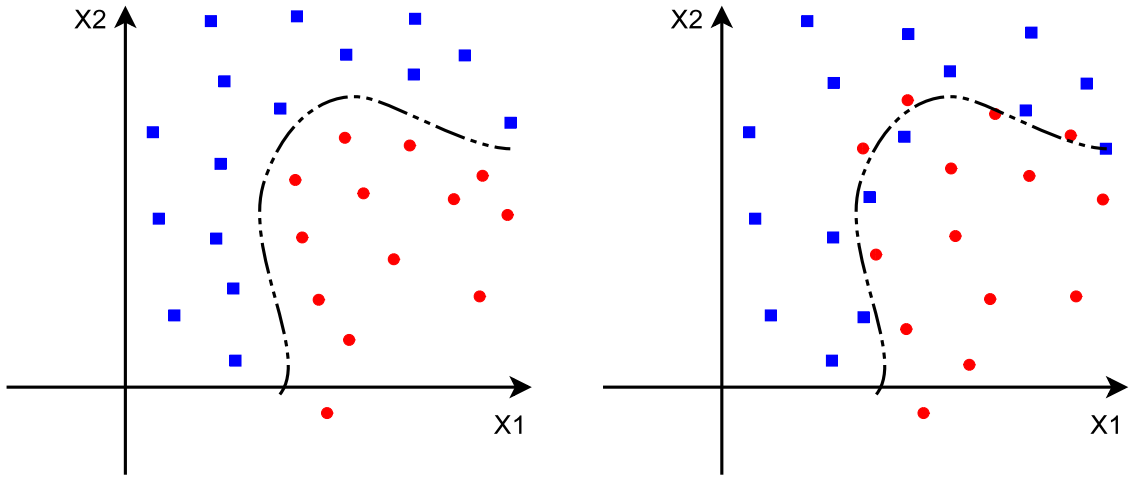



- On introduit des variables auxiliaires (*slack variables*) $\xi_i \geq 0$ :

$$\begin{cases} \text{Minimiser } J(w_0, w, \xi) = \dfrac{\|w\|^2}{2} + C \displaystyle\sum_{i=1}^{n} \xi_i \\ \text{sous les contraintes } \xi_i \geq 0 \text{ et } Y_i \left(w_0 + \langle w, X_i \rangle\right) \geq 1 - \xi_i \end{cases}$$

# SVM non linéaire : utilisation de noyaux.

- En général, la frontière de séparation idéale n'est pas linéaire




- Pour généraliser les SVM au cas non linéaire, on remplace les produits scalaires $\langle X_i, X_j \rangle$ dans la formulation duale par un **noyau** $k(X_i, X_j)$ :

$$\begin{cases} \text{Maximiser } \psi(\mu) = \sum_{i=1}^{n} \mu_i - \frac{1}{2} \sum_{i,j} \mu_i \mu_j \; Y_i Y_j \; k(X_i, X_j) \\ \text{sous les contraintes } \mu_i \geq 0 \text{ et } \sum_{i=1}^{n} \mu_i Y_i = 0 \end{cases}$$

# SVM non linéaire : utilisation de noyaux.

- la fonction noyau $k$ doit avoir certaines propriétés (voir slide suivante). Exemples :
  - noyau gaussien $k(X_i, X_j) = e^{\|X_i - X_j\|^2/\sigma^2}$
  - noyau polynomial $k(X_i, X_j) = (1 + \langle X_i, X_j \rangle)^d$,
  - noyau linéaire $k(X_i, X_j) = \langle X_i, X_j \rangle$, qui redonne le modèle linéaire.
- La fonction discriminante est alors

$$f(X) = w_0 + \sum_{i=1}^{n} \hat{\mu}_i \; Y_i \; k(X_i, X)$$

- La méthode *SVM* avec noyau $k$ correspond en fait à appliquer la méthode SVM linéaire pour des données transformées $\phi(X_i)$ (voir slide suivante). Cependant il est inutile de calculer les $\phi(X_i)$ ; tout ce qui compte est de bien choisir le noyau $k$ et de résoudre le problème dual.

# Interprétation théorique : espaces à noyaux reproduisants

- L'application $\phi$ est en fait la suivante : $\phi(X_i) = k(X_i, \cdot)$, c'est-à-dire qu'elle assigne à l'observation $X_i \in \mathbb{R}^P$ la fonction $X \mapsto k(X_i, X)$. Les données se retrouvent donc plongées dans un espace de fonctions (donc de dimension infinie en général).
- La quantité $k(X_i, X_j)$ correspond alors au produit scalaire pour les données transformées : $k(X_i, X_j) = \langle \phi(X_i), \phi(X_j) \rangle_k$.
- On a ainsi l'identité suivante, appelée **propriété reproduisante** du noyau :

$$k(X_i, X_j) = \langle k(X_i, \cdot), k(X_j, \cdot) \rangle_k.$$

- Pour que cette construction mathématique fonctionne, il faut que la fonction $k$ soit un **noyau de type positif** : quels que soient les $X_i \in \mathbb{R}^P$, la matrice $n \times n$ de coefficient général $k(X_i, X_j)$ doit être symétrique à valeurs propres positives. C'est le cas pour les noyaux linéaires, gaussiens et polynomiaux notamment.