Chapitre 1: Machines à Supports Vecteur ou Séparateurs à Vastes Marges

Adnène Arbi

adnene.arbi@enstab.ucar.tn

February 8, 2024

Plan de cours

Qu'est ce que l'apprentissage?

En psychologie

Toute acquisition d'un nouveau comportement à la suite d'un entraînement: habituation, conditionnement...

En neurobiologie

Modifications synaptiques dans des circuits neuronaux: règle de Hebb, règle de Rescorla et Wagner...

Apprentissage automatique

- Construire un modèle général à partir des données particulières.
- But:
  Prédire un comportement face à une nouvelle donnée.
  Approximer une fonction ou une densité de probabilité.

Formalisation

- Soit un ensemble d'apprentissage $S = (x_i, y_i)_{i=1,\ldots,n}$ dont les éléments obéissent à la loi jointe $P(x, y) = P(x)P(y/x)$
- On cherche à approcher une loi sous-jacente $f(x)$ telle que $y_i = f(x_i)$ par une hypothèse $h_\alpha(x)$ aussi proche que possible.
- Les α sont les paramètres du système d'apprentissage.

Remarque importante

- Si $f(.)$ est discrète on parle de **classification**.
- Si $f(.)$ est une fonction continue on parle alors de **régression**.

Mais que veut-on dire par

"aussi proche que possible" ?

Calcul du risque

Pour mesurer la qualité d'une hypothèse h_α on va considérer une fonction de coût $Q(z = (x, y), \alpha) \in [a, b]$ que l'on cherche à minimiser

Exemples de fonction coût

- Coût 0/1: vaut 0 lorsque les étiquettes prévues et observées coincident, 1 sinon: utilisé en classification.
- Erreur quadratique: $(f(x) - y)^2$ utilisé en régression.
- On cherche à minimiser: $\mathcal{R}(\alpha) = \int Q(z, \alpha) dP(z)$

Comme on ne peut pas accéder à cette valeur, on construit le risque empirique qui mesure les erreurs réalisées par le modèle:

$$\mathcal{R}_{emp}(\alpha) = \frac{1}{m} \sum_{i=1}^{n} Q(z_i, \alpha)$$

Mais quel est la relation entre $\mathcal{R}_{emp}(\alpha)$ et $\mathcal{R}(\alpha)$

Théorie de l'apprentissage de Vapnik (1995)

Vapnik a pu montrer l'expression suivante $\forall m$ avec une probabilité au moins égale à $1 - \eta$

$$\mathcal{R}(\alpha_m) \leq \mathcal{R}_{emp}(\alpha_m) + (b - a)\sqrt{\frac{d_{VC}(\ln(\frac{2m}{d_{VC}}) + 1) - \ln(\frac{\eta}{4})}{m}}$$

La minimisation du risque dépend

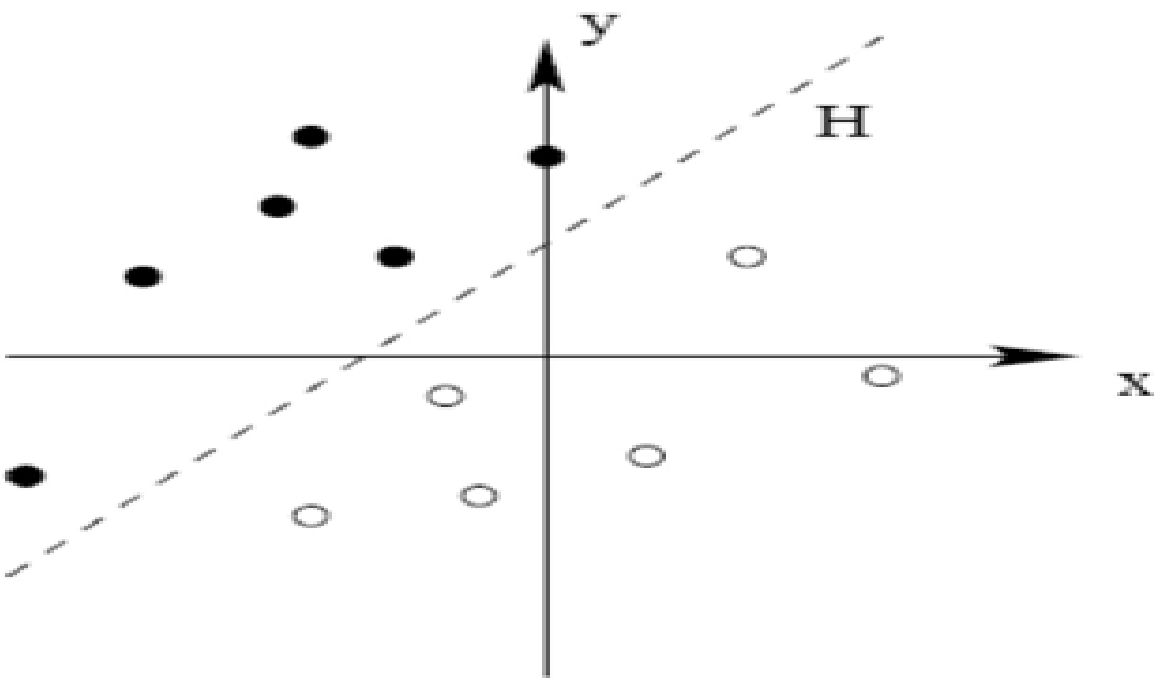
- **du risque empirique**.
- **un risque structurel** lié au terme d_{VC} qui dépend de la complexité du modèle h choisi (VC-dimension: Dimension de Vapnik et Chervonenkis).

SVM

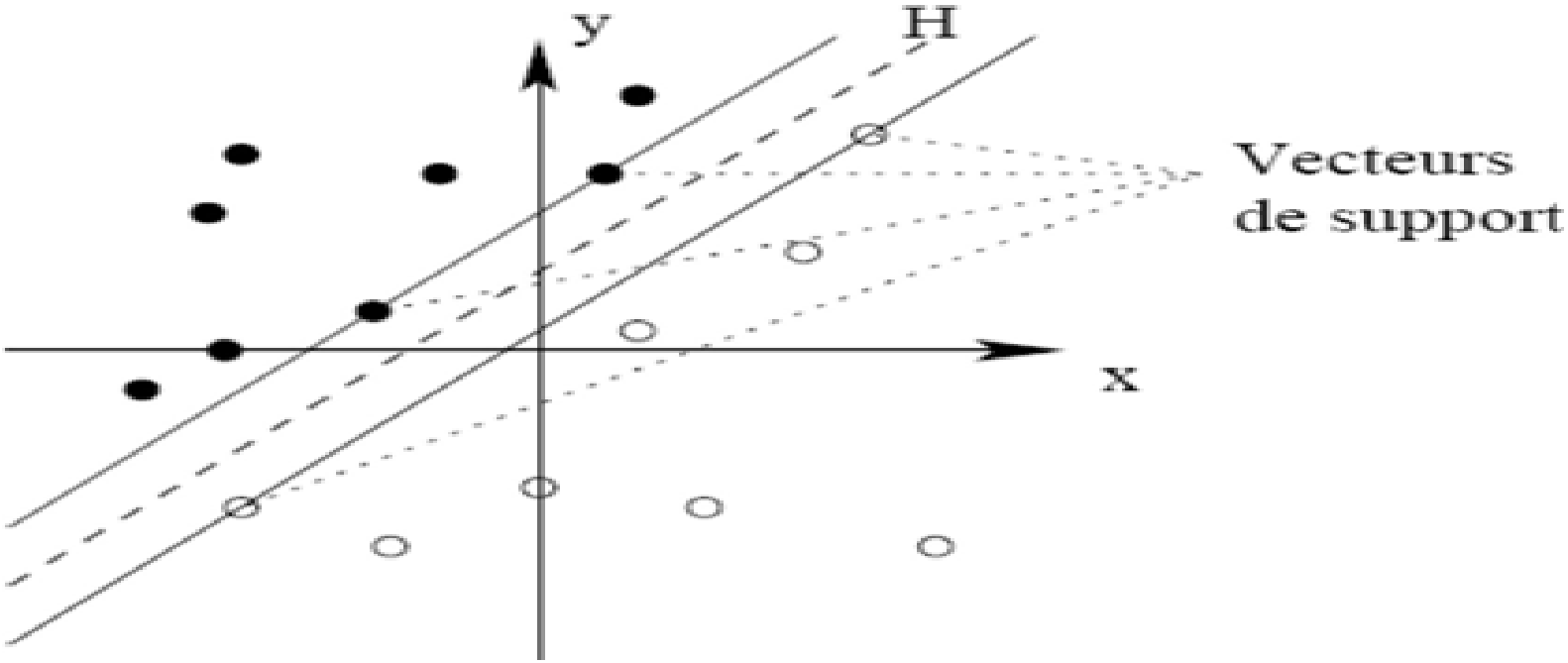
- Parmi les méthodes à noyaux, inspirées de la théorie statistique de l'apprentissage de Vladimir Vapnik, les SVM présentent la forme la plus connue.
- SVM est une méthode de classification binaire par apprentissage supervisé.
- SVM a été introduit par Vapnik en 1995.
- Cette méthode repose sur l'existence d'un classificateur linéaire dans un espace approprié.
- Puisque c'est un problème de classification à deux classes, cette méthode fait appel à un jeu de données d'apprentissage pour apprendre les paramètres du modèle.
- Elle est basée sur l'utilisation de fonction dites noyau (kernel) qui permettent une séparation optimale des données.

Notion de base: Hyperplan, marge et support vecteur

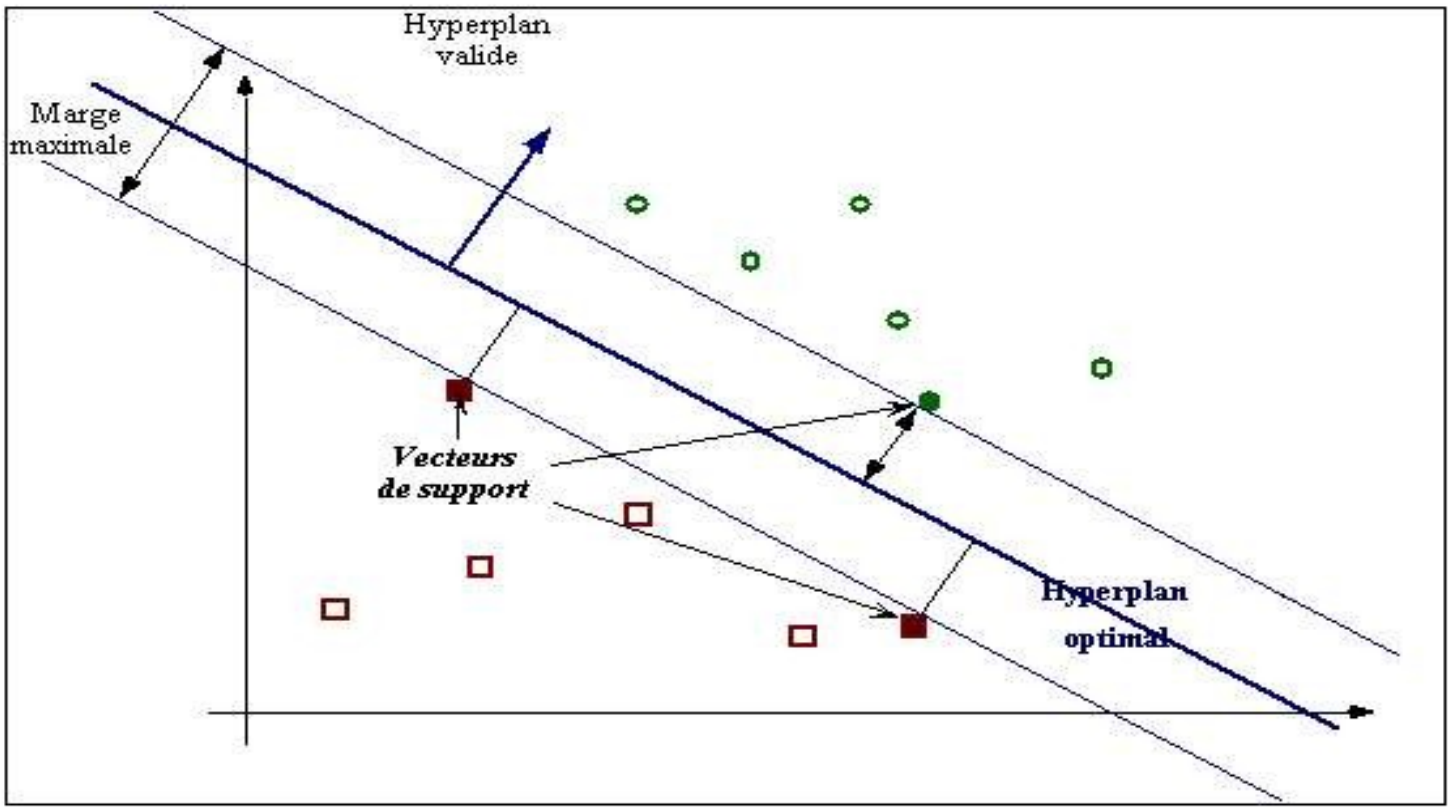
- Pour deux classes d'exemples de donnés, le but de SVM est de trouver un classificateur qui va séparer les données et maximiser la distance entre ces deux classe.
- Avec SVM, ce classificateur est un classificateur linéaire appelé hyperplan.



- Les points les plus proches, qui seuls sont utilisés pour la détermination de l'hyperplan, sont appelés vecteurs de support.

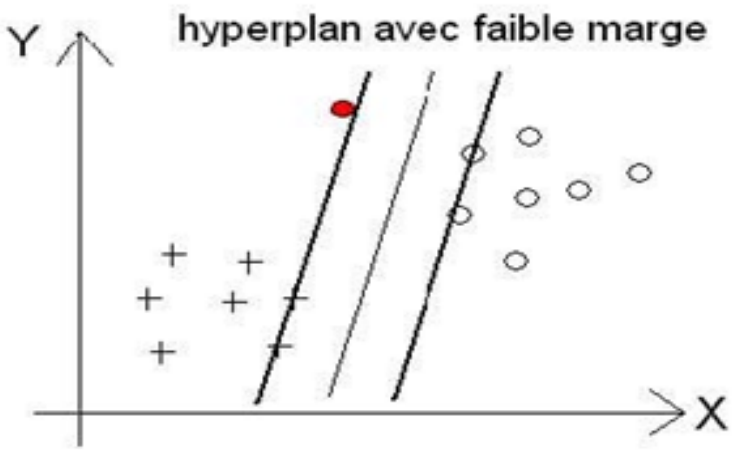


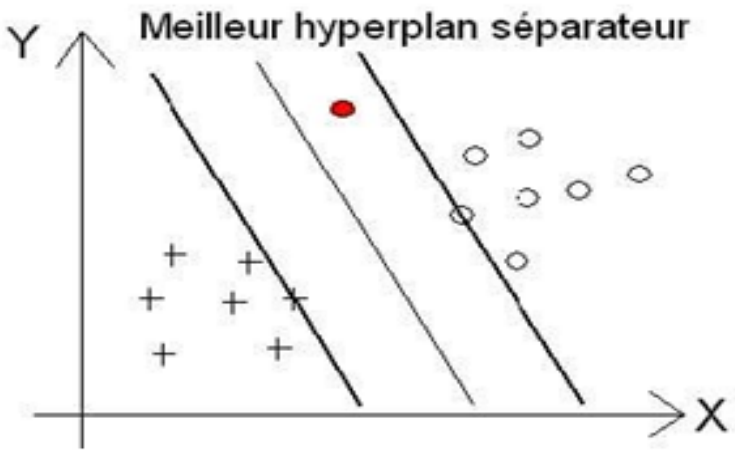
- Il est évident qu'il existe plusieurs hyperplan valide mais la propriété remarquable des SVM est que cet hyperplan doit être optimal.
- Nous allons donc en plus chercher parmi les hyperplans valides, celui qui passe au milieu des points des deux classes d'exemples.



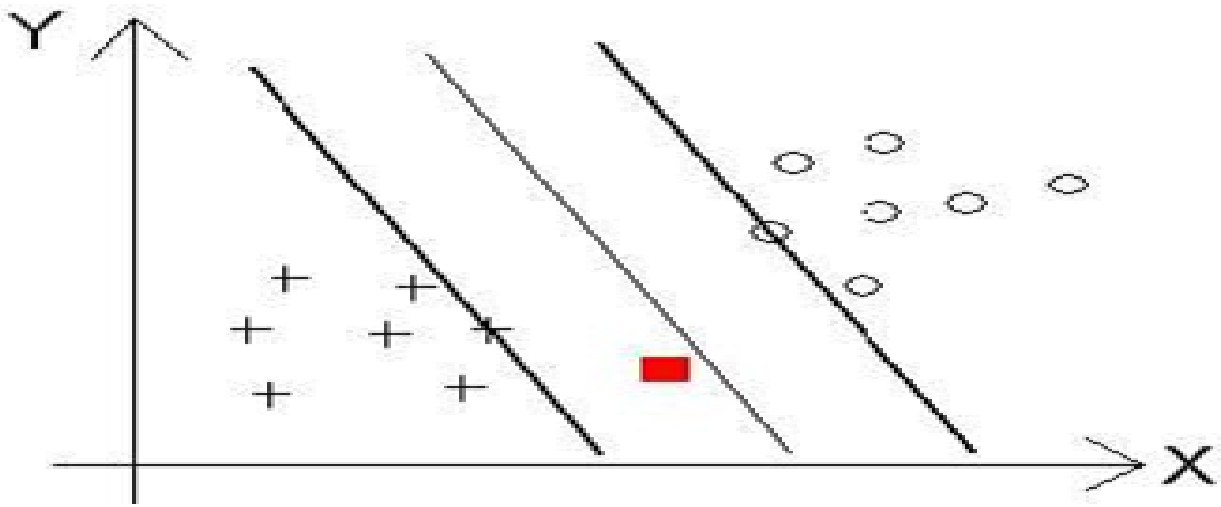
Pourquoi maximiser la marge?

- Intuitivement, le fait d'avoir une marge plus large procure plus de sécurité lorsque l'on classe un nouvel exemple.
- De plus, si l'on trouve le classificateur qui se comporte le mieux vis-à-vis des données d'apprentissage, il est clair qu'il sera aussi celui qui permettra au mieux de classer les nouveaux exemples.
- Dans le shéma qui suit, la partie droite nous montre qu'avec un hyperplan optimal, un nouvel exemple reste bien classé alors qu'il tombe dans la marge. On constate sur la partie gauche qu'avec une plus petite marge, l'exemple se voit mal classé.





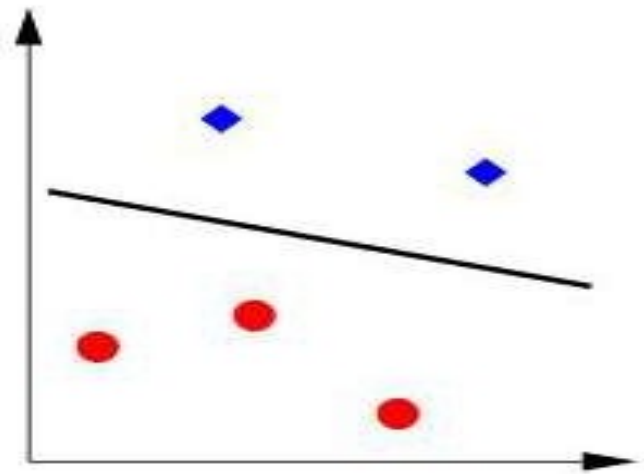
- En général, la classification d'un nouvel exemple inconnu est donnée par sa position par rapport à l'hyperplan optimal.
- Dans le shéma suivant, le nouvel élément sera classé dans la catégorie des "+".



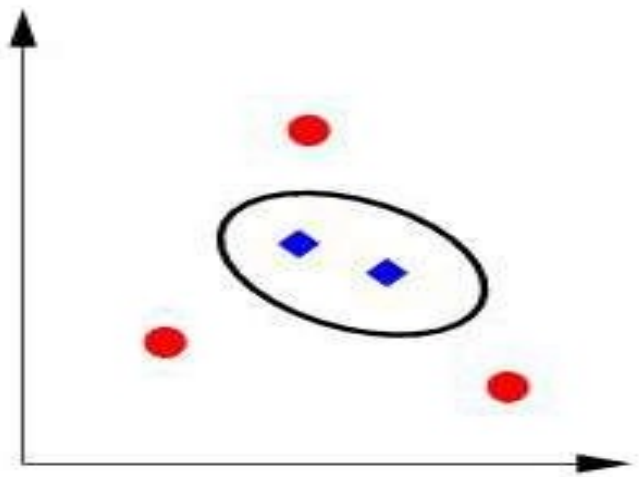
Linéarité et non-linéarité

- Parmi les données traitées via SVM, on constate le cas linéairement séparable et les cas non linéairement séparable.
- Les premiers sont les plus simple de SVM car il permettent de trouver facilement le clasificateur linéaire.

Cas linéairement séparable

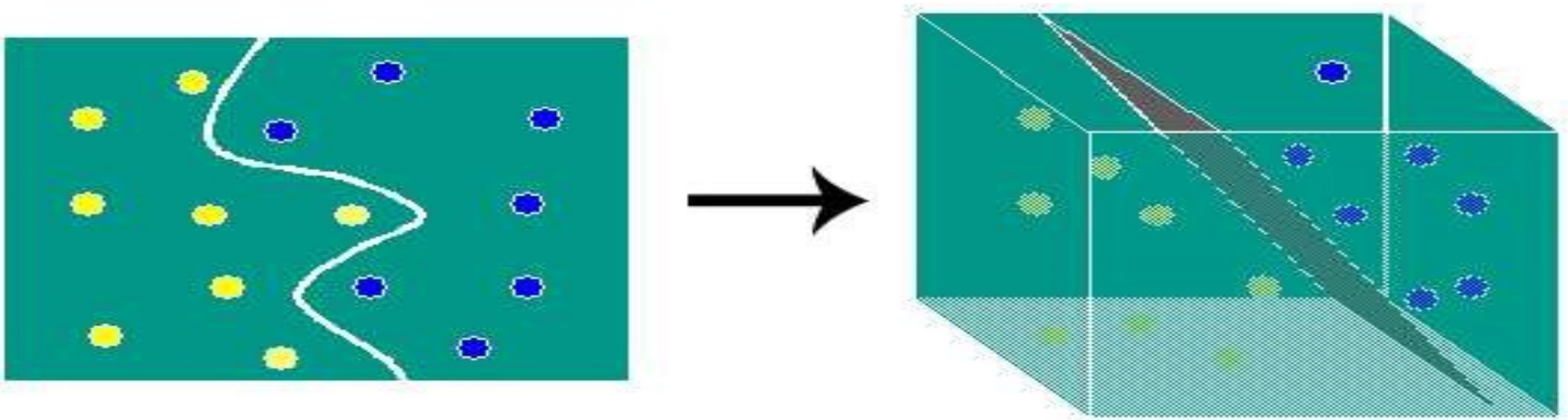


Cas non linéairement séparable



SVM non linéaires: espace de caractéristiques

- L'idée des SVM est de changer l'espace des données.
- La transformation non linéaire des données peut permettre une séparation linéaire des exemples dans un nouvel espace.
- Intuitivement, plus la dimension de l'espace de re-description est grande, plus la probabilité de pouvoir trouver un hyperplan séparateur entre les exemples est élevée.

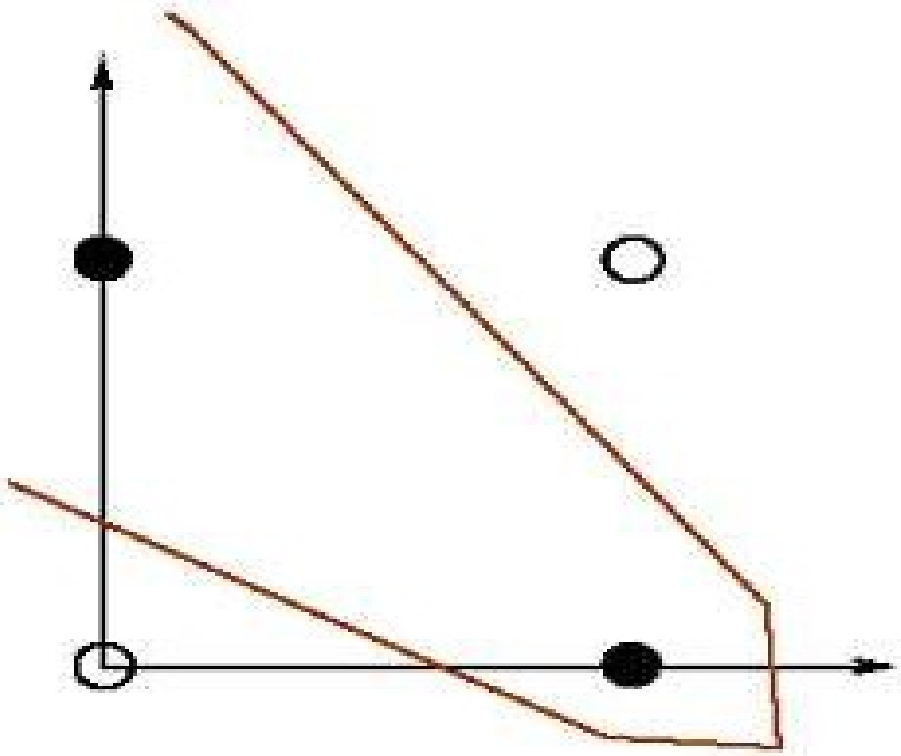


Idée générale

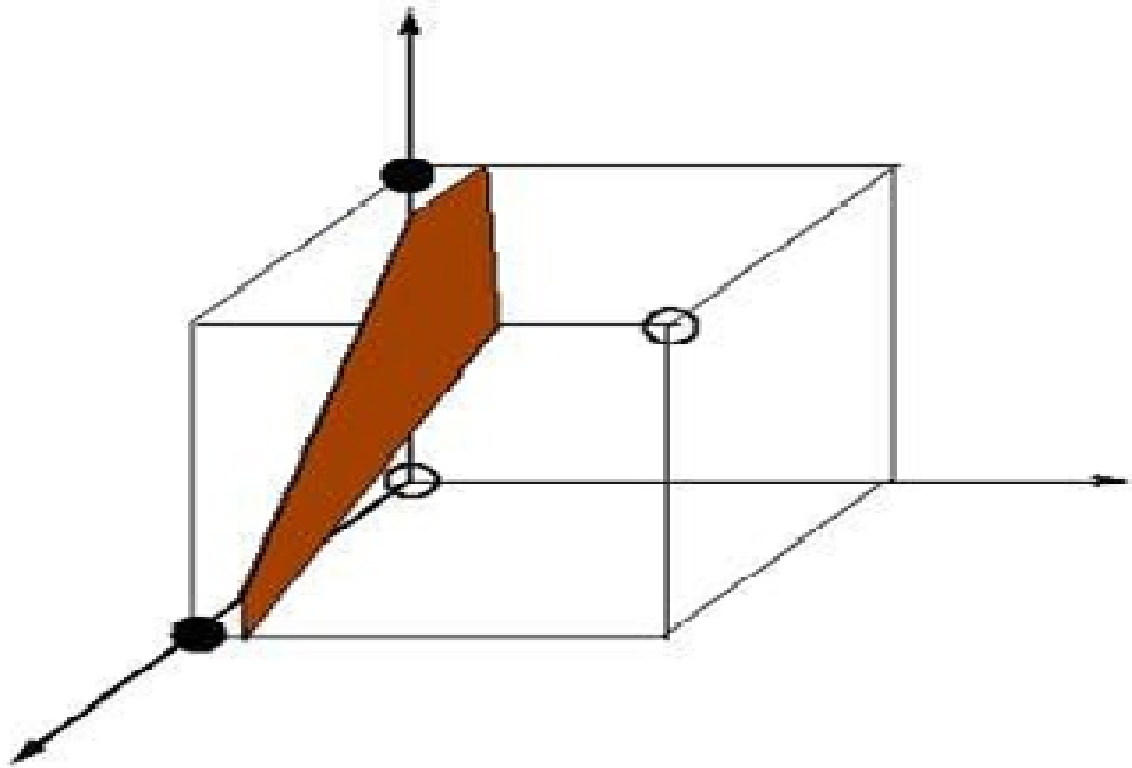
L'espace des données peut toujours être plongé dans un espace de plus grande dimension dans lequel les données peuvent être séparées linéairement.

Exemple: le cas de XOR

- Le cas de XOR n'est pas linéairement séparable, si on place les points dans un plan à deux dimension, on obtient la figure suivante: Coordonnées des points: (0,0); (0,1); (1,0); (1,1).



- Si on prend une fonction polynomiale $(x, y) \rightarrow (x, y, x.y)$ qui fait passer d'un espace de dimension 2 à un espace de dimension 3, on obtient un problème en trois dimensions linéairement séparable.

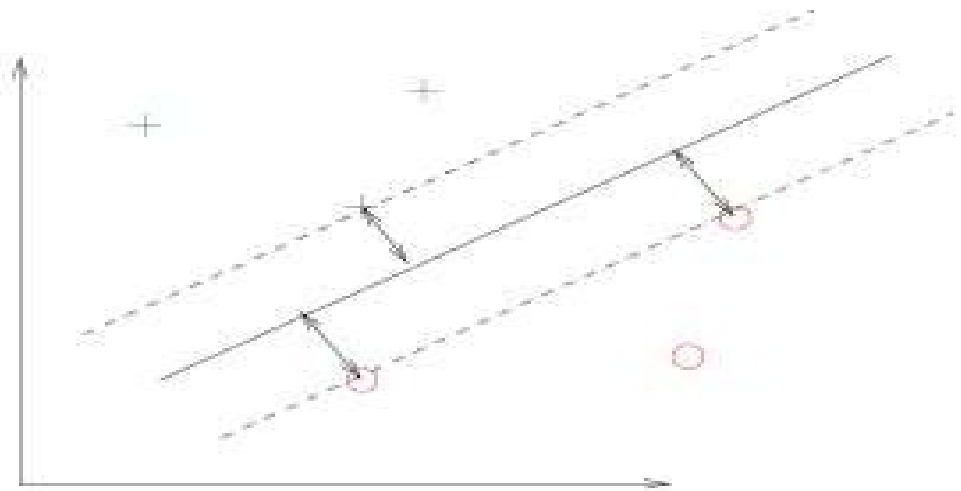


Maximisation de la marge

On appelle marge d la distance entre les 2 classes

C'est cette distance d qu'on souhaiterait maximiser.

- La marge est la distance du point le plus proche à l'hyperplan.
- Dans un modèle linéairement séparable, on a $f(x) = wx + b$.
- L'hyperplan séparateur a donc pour équation $wx + b = 0$.



Formulation du problème

- La distance d'un point au plan est donnée par $\frac{|wx+b|}{\|w\|}$.
- L'hyperplan optimal est celui pour lequel la distance aux points les plus proches (marge) est maximale.
- Soit x_1 et x_2 deux points de classes différentes $f(x_1) = 1$ et $f(x_2) = -1$.
- Donc $w(x_1 - x_2) = 2$, d'où $\frac{w}{\|w\|}(x_1 - x_2) = \frac{2}{\|w\|}$.
- On peut donc en déduire que maximiser la marge revient à minimiser $\|w\|$ sous certaines contraintes.

Problème Primal

Théorème 1

Un point (x, y) est bien classé si et seulement si $yf(x) > 0$.

Remarque

Comme le couple (w, b) est défini à un coefficient près, on s'impose $yf(x) \geq 1$.

Le problème primal

On en déduit le problème de minimisation sous contraintes suivantes:

$$\begin{cases} \min \frac{1}{2}\|w\|^2, \\ y_i(wx_i + b) \geq 1. \ \ i = 1, ..., n \end{cases}$$

Remarque

Il est plus facile de minimiser $\|w\|^2$ plutot que directement $\|w\|$.

Du primal au dual

Cette minimisation est possible sous les conditions dites de "Karush-Kuhn-Tucker (KKT)"

Soit le lagrangien $\mathcal{L}$:

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(w.x_i + b) - 1]$$

Les conditions de KKT sont alors:

$$\frac{\partial \mathcal{L}}{\partial w} = 0, \quad \frac{\partial \mathcal{L}}{\partial b} = 0, \quad \frac{\partial \mathcal{L}}{\partial \alpha_i} = 0, \quad \alpha_i \geq 0$$

$$\alpha_i [y_i(w.x_i + b) - 1] = 0$$

Par ailleurs la dernière condition implique que pour tout point ne vérifiant pas $y_i(w.x_i + b) = 1$ le α_i est nul.

Les vecteurs supports

Remarque

Les points qui vérifient $y_i(w.x_i + b) = 1$, sont appelés "vecteurs supports". Ce sont les points les plus près de la marge. Ils sont sensés être peu nombreux par rapport à l'ensemble des exemples.

Problème dual

Le problème dual

On passe du problème primal au problème dual en introduisant des multiplicateurs de Lagrange pour chaque contrainte:

$$\begin{cases} \max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i . x_j, \\ \alpha_i \geq 0, \quad i = 1, ..., n \\ \sum_i \alpha_i y_i = 0. \end{cases}$$

C'est un problème de programmation quadratique de dimension n.

- Si les α_i^* sont les solutions du problème dual, alors on a:

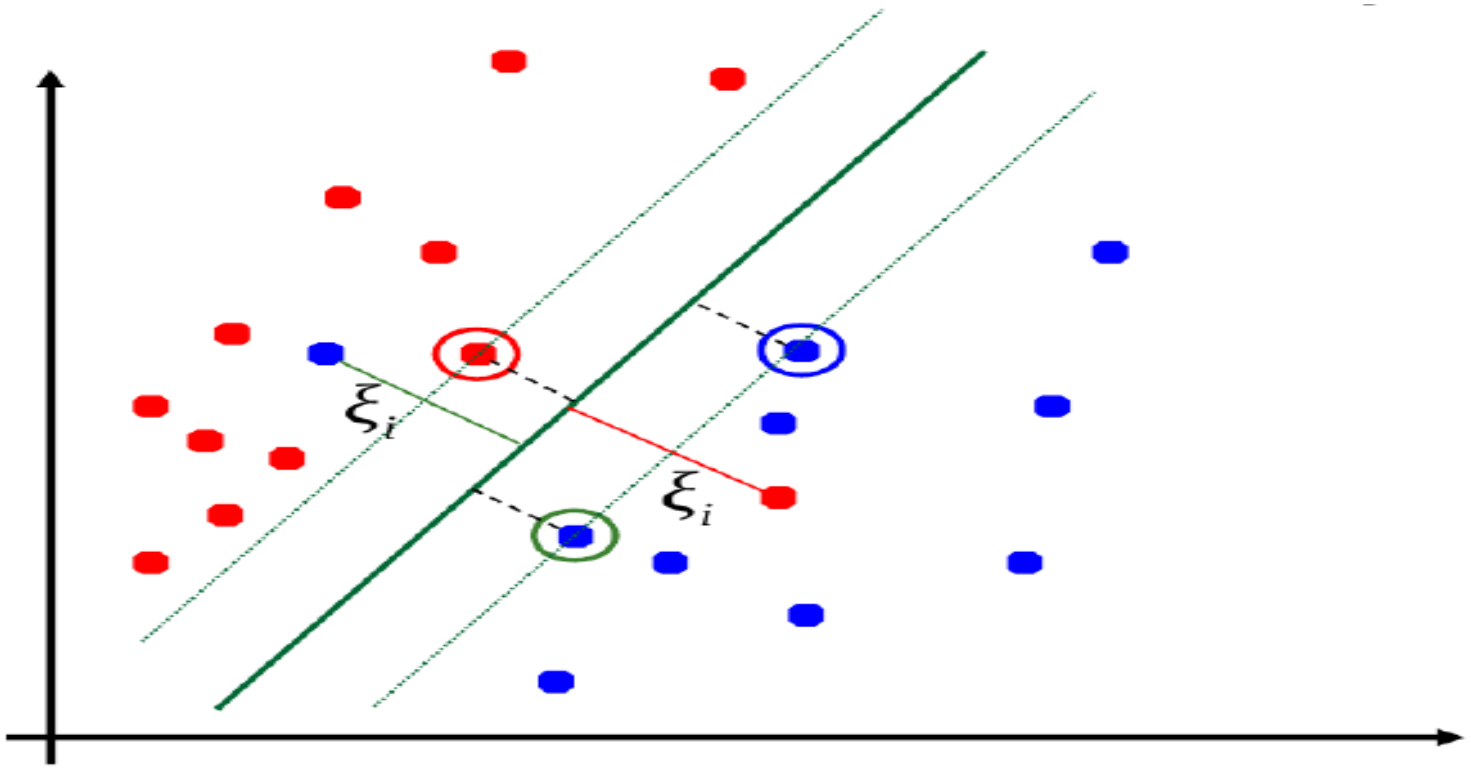
$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

- Seul les α_i correspondant aux points les plus proches sont non-nuls. On parle de vecteurs de support.
- La fonction de décision associée est donc

$$f(x) = \sum_{i=1}^{n} \alpha_i^* y_i x_i . x + b$$

Classification à marge souple

- L'idée est d'ajouter des variables d'ajustement ξ_i dans la formulation pour prendre en compte les erreurs de classification ou le bruit.



Classification à marge souple: formulation

Problème original

$$\begin{cases} \min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i, \\ y_i(w^*.x_i + b) \geq 1 - \xi_i. \;\; i = 1, ..., n \\ \xi_i \geq 0 \end{cases}$$

Problème dual

$$\begin{cases} \max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i.x_j, \\ 0 \leq \alpha_i \leq C, \;\; i = 1, ..., n \\ \sum_i \alpha_i y_i = 0. \end{cases}$$

Le noyau "Kernel"

La résolution des SVM ne s'appuient que sur le produit scalaire $< x_i, x_j >$ entre les vecteurs d'entrée.
Si les données d'apprentissage sont plongées dans un espace de plus grande dimension via la transformation $\phi : x \longrightarrow \phi(x)$, le produit scalaire devient

$$K(x_i, x_j) =< \phi(x_i), \phi(x_j) > .$$

($K(x_i, x_j)$ est appelée fonction noyau).

Remarque

Pour faire apprendre le SVM seul le noyau est important, sans qu'il ne soit nécessaire d'effectuer la tensformée $\phi(.)$

SVM non linéaire: Formulation

Problème original

$$\begin{cases} \min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i, \\ y_i(w^*\phi(x_i)+b) \geq 1-\xi_i. \;\; i=1,...,n \end{cases}$$

Problème dual

$$\begin{cases} \max \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \phi(x_i).\phi(x_j), \\ 0 \leq \alpha_i \leq C, \;\; i=1,...,n \\ \sum_i \alpha_i y_i = 0. \end{cases}$$

La frontière de décision

La solution est:

$$f(x) = \sum_{i=1}^{n}\alpha_i^* y_i \phi(x_i).\phi(x) + b.$$

Exemples de noyaux

Noyau polynôme de degré 2 à 2 variables

Transformée non-linéaire:

$$x = (x_1, x_2)$$

$$\phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

Le noyau est alors:

$$\phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

$$\phi(y) = (1, \sqrt{2}y_1, \sqrt{2}y_2, y_1^2, y_2^2, \sqrt{2}y_1y_2)$$

$$K(x, y) = \phi(x).\phi(y) = (1 + x.y)^2$$

Comment savoir si K est un noyau?

Noyau de Mercer

- On appel noyau de Mercer une fonction continue, symétrique, semi-définie positive $K(x, y)$.

Matrice de Gram

Matrice des termes $< x_i, x_j >$. Elle est symétrique et semi-définie positive pour un noyau de Mercer.

Théorème de Moore-Aronszajn (1950)

- Toute fonction semi-définie positive $K(x, y)$ est un noyau, et réciproquement. Elle peut s'exprimer comme un produit scalaire dans un espace de grande dimension.
- $K(x_i, x_j) =< \phi(x_i), \phi(x_j) >$.

Exemples de noyaux

- Noyau linéaire: $K(x, x') = x.x'$
- Noyau polynomial de degré d: $K(x, x') = (c + x.x')^d$
- Noyau Gaussien: $\exp \frac{-\|x-x'\|^2}{2\sigma^2}$
  Cette formulation est équivalente aux réseaux de neurones à base radiale.
- Perceptron à deux couches: $K(x, x') = \tanh \alpha x.x' + \beta$.

Cas multi-classes

Les séparateurs à vaste marge ont été développés pour traiter des problèmes binaires mais ils peuvent être adaptés pour traiter les problèmes multi-classes.

Stratégie un contre tous

- L'idée consiste simplement à transformer le problème à k classes en k classifieurs binaires.
- Le classement est donné par le classifieur qui répond le mieux.

Stratégie un contre un

- Cette fois le problème est transformé en $\frac{k(k-1)}{2}$ classifieurs binaires: chaque classe i étant en effet comparée à chaque classe j.
- Le classement est donné par le vote majoritaire ou un graphe acyclique de décision.

Les domaines d'applications

- SVM est une méthode de classification qui montre de bonnes performances dans la résolution de problèmes variés.
- Cette méthode a montré son efficacité dans de nombreux domaines d'applications tels que le traitement d'image, la catégorisation des textes ou le diagnostique médicales et ce même sur des ensembles de données de très grandes dimensions.
- Classification des données biologiques / Physiques.
- Classification de documents numériques.
- Reconnaissance de la parole.
- Classification d'expressions faciales.
- Classification des textures.

Conclusion

- Les avantages théoriques (minimisation de l'erreur empirique ou structurelle) et pratiques (algorithmes optimisés) des SVM en ont fait un outil très prisé dans de nombreux problèmes pratiques de classification.
- Dans bien de cas, il s'agit de construire un noyau (donc une mesure de similarité) adaptés aux données à traiter.

Merci pour votre attention