Introduction aux "Support Vector Machines" (SVM)

Olivier Bousquet
Centre de Mathématiques Appliquées
Ecole Polytechnique, Palaiseau

Orsay, 15 Novembre 2001

But de l'exposé

- Présenter les SVM
- Encourager leur utilisation
- Encourager leur étude

Plan

→ Quel est le contexte ?

- Qu'est-ce que c'est ?
- Comment cela marche-t-il ?
- Pourquoi est-ce utile ?
- Pourquoi est-ce que cela marche ?

Problème d'apprentissage

On s'intéresse à un phénomène f (éventuellement non-déterministe) qui,

- à partir d'un certain jeu d'entrées $\boldsymbol{x}$,
- produit une sortie $y = f(\boldsymbol{x})$.

$\rightarrow$ Le but est de retrouver f à partir de la seule observation d'un certain nombre de couples entrée-sortie $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$

Formalisation

On considère un couple (X, Y) de variables aléatoires à valeurs dans $\mathcal{X} \times \mathcal{Y}$. Seul le cas $\mathcal{Y} = \{-1, 1\}$ (classification) nous intéresse ici (on peut facilement étendre au cas $|\mathcal{Y}| = m > 2$ et au cas $\mathcal{Y} = \mathbb{R}$).

La distribution jointe de (X, Y) est inconnue.

- Données: on observe un échantillon $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ de n copies indépendantes de (X, Y).
- But: construire une fonction $h : \mathcal{X} \to \mathcal{Y}$ telle que $P(h(X) \neq Y)$ soit minimale.

Exemples de problèmes d'apprentissage

Reconnaissance de formes

- Reconnaissance de chiffres manuscrits (après segmentation: codes postaux)
- Reconnaissance de visages

Entrées: image bidimensionnelle en couleur ou en niveaux de gris

Sortie: classe (chiffre, personne)

Exemples de problèmes d'apprentissage

Catégorisation de textes

- Classification d'e-mails
- Classification de pages web

Entrées: document (texte ou html)

Sortie: catégorie (thème, spam/non-spam)

Exemples de problèmes d'apprentissage

Diagnostic médical

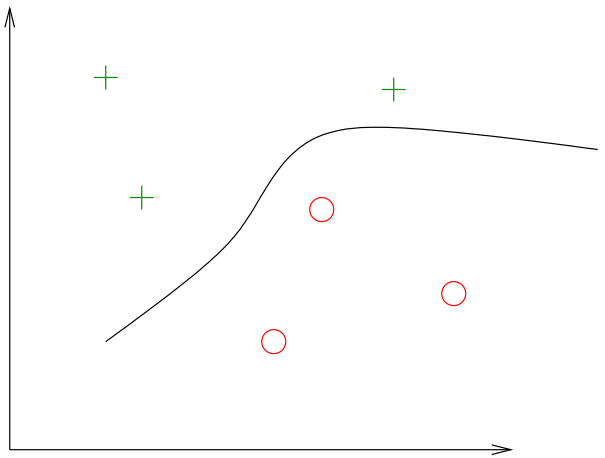
- Evaluation des risques de cancer
- Detection d'arythmie cardiaque

Entrées: etat du patient (sexe, age, bilan sanguin, génome...)

Sortie: classe (à risque ou non)

Illustration

Trouver une frontière de décision qui sépare l'espace en deux régions (pas forcément connexes).



→ Problème d'optimisation

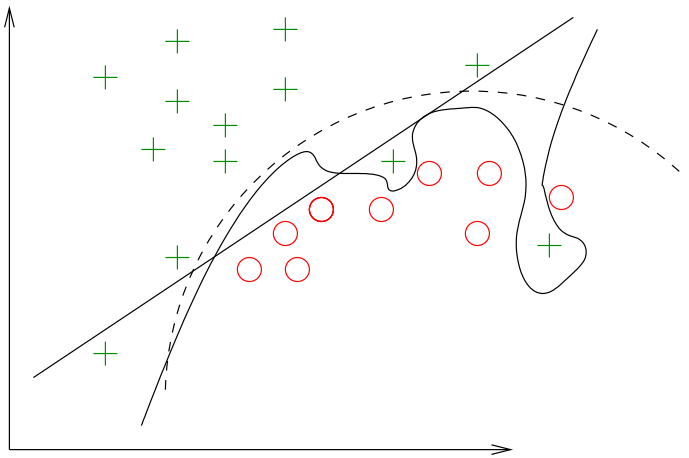
Critère: erreur constatée sur les données (Erreur empirique)

Espace de recherche: ensemble paramétré de fonctions par exemple

→ Problème mal posé (solution non unique)
→ Garanties ?

Illustration

Sur et sous-apprentissage



- Si les données sont générées par un modèle quadratique
- Le modèle linéaire est en situation de sous-apprentissage
- Le modèle de haut degré est en situation de sur-apprentissage (apprentissage par coeur)

$\rightarrow$ Un compromis est à trouver entre adéquation aux données et complexité pour pouvoir généraliser.

Plan

- Quel est le contexte ?

→ Qu'est-ce que c'est ?

- Comment cela marche-t-il ?

- Pourquoi est-ce utile ?

- Pourquoi est-ce que cela marche ?

Qu'est-ce que c'est ?

Les Support Vector Machines sont une classe d'algorithmes d'apprentissage.

Principe général

- Construction d'un classifieur à valeurs réelles
- Découpage du problème en deux sous-problèmes
  1. Transformation non-linéaire des entrées
  2. Choix d'une séparation linéaire 'optimale'

Classification à valeurs rélles

- Plutôt que de construire directement $h : \mathcal{X} \to \{-1, 1\}$, on construit

$$f : \mathcal{X} \to \mathbb{R} \, .$$

- La classe est donnée par le signe de f

$$h = \mathrm{sgn}(f) \, .$$

- L'erreur se calcule avec

$$P(h(X) \neq Y) = P(Y f(X) \leq 0) \, .$$

$\to$ Donne une certaine idée de la confiance dans la classification. Idéalement, $|Y f(X)|$ est 'proportionnel' à $P(Y|X)$.

$\to$ $Y f(X)$ est la marge de f en (X, Y).

Transformation des entrées

- $\mathcal{X}$ est un espace quelconque d'objets.

- On transforme les entrées en vecteurs dans un espace $\mathcal{F}$ (feature space).

$$\Phi : \mathcal{X} \rightarrow \mathcal{F} .$$

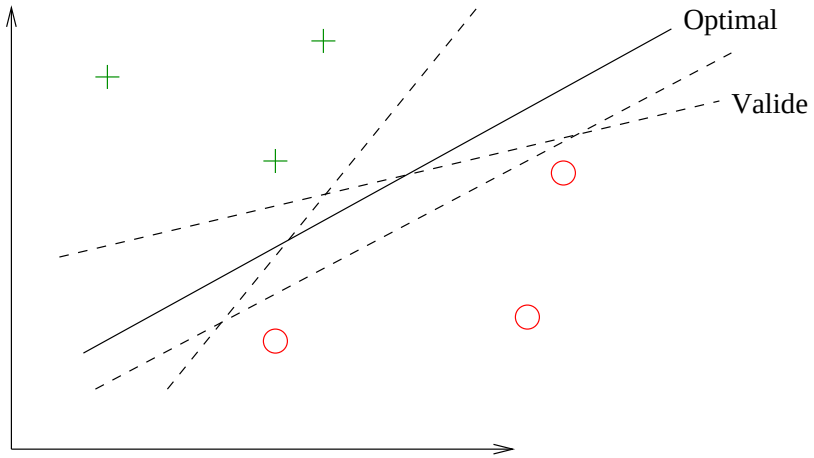
- F n'est pas nécessairement de dimension finie mais dispose d'un produit scalaire (espace de Hilbert)

- La non-linéarité est traitée dans cette transformation, on peut donc choisir une séparation linéaire.

Choix d'une séparation

Hyperplan optimal

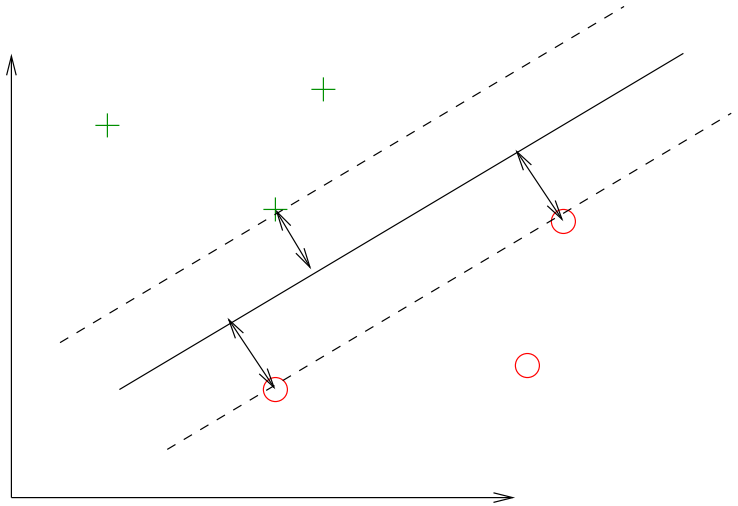
Choix de la séparation: hyperplan qui classifie correctement les données (lorsque c'est possible) et qui se trouve "le plus loin possible de tous les exemples".



Maximisation de la marge

Définition géométrique

Marge = distance du point le plus proche à l'hyperplan



Plan

- Quel est le contexte ?
- Qu'est-ce que c'est ?

→ Comment cela marche-t-il ?

- Pourquoi est-ce utile ?
- Pourquoi est-ce que cela marche ?

Maximisation de la marge

Propriétés

Modèle linéaire

$$f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b$$

Définition de l'hyperplan (frontière de décision)

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$$

La distance d'un point au plan est donnée par

$$d(\boldsymbol{x}) = \frac{|\boldsymbol{w} \cdot \boldsymbol{x} + b|}{\|\boldsymbol{w}\|}$$

$\rightarrow$ Maximiser la marge revient à minimiser $\|\boldsymbol{w}\|$ sous contraintes

Problème primal

Un point $(\boldsymbol{x}_i, y)$ est bien classé si et seulement si

$$yf(\boldsymbol{x}) > 0$$

Comme le couple $(\boldsymbol{w}, b)$ est défini à un coefficient multiplicatif près, on s'impose

$$yf(\boldsymbol{x}) \geq 1$$

Rappelons que

$$d(\boldsymbol{x}) = \frac{|f(\boldsymbol{x})|}{\|w\|}$$

On obtient le problème de minimisation sous contraintes

$$\begin{cases} \min \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \forall i,\ \ y_i(\boldsymbol{w}.\boldsymbol{x}_i + b) \geq 1 \end{cases}$$

Problème dual

On passe au problème dual en introduisant des multiplicateurs de Lagrange pour chaque contrainte.

Ici on a une contrainte par exemple d'apprentissage

$$\begin{cases} \max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j \\ \forall i, \alpha_i \geq 0 \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

Problème de programmation quadratique de dimension n (nombre d'exemples).

Matrice hessienne: $(\boldsymbol{x}_i \cdot \boldsymbol{x}_j)_{i,j}$.

Propriétés

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \boldsymbol{x}_i$$

Seuls les α_i correspondant aux points les plus proches sont non-nuls. On parle de **vecteurs de support**.

Fonction de décision

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^* y_i \boldsymbol{x}_i \cdot \boldsymbol{x} + b$$

Cas non-séparable

Traitement des erreurs

On introduit des variables 'ressort' pour assouplir les contraintes.

$$\begin{cases} \min \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n} \xi_i \\ \forall i,\ y_i(\boldsymbol{w}.\boldsymbol{x}_i + b) \geq 1 - \xi_i \end{cases}$$

On pénalise par le dépassement de la contrainte.

Cas non-séparable

Problème dual

Le problème dual a la même forme que dans le cas séparable:

$$\begin{cases} \max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j \\ \forall i, 0 \leq \alpha_i \leq C \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

La seule différence est la borne supérieure sur les α.

Fonction noyau

Espace intermédiaire

Au lieu de chercher un hyperplan dan l'espace des entrées, on passe d'abord dans un espace de représentation intermédiaire (*feature space*) de grande dimension.

$$\begin{aligned}\Phi \; : \; & \mathbb{R}^d \rightarrow \mathcal{F} \\ & \boldsymbol{x} \mapsto \Phi(\boldsymbol{x})\end{aligned}$$

On doit donc résoudre

$$\begin{cases} \max \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}_j) \\ \forall i, 0 \leq \alpha_i \leq C \\ \sum_{i=1}^n \alpha_i y_i = 0 \end{cases}$$

et la solution a la forme

$$f(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i^* y_i \Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}) + b$$

Fonction noyau

Propriétés

Le problème et sa solution ne dépendent que des produits scalaires $\Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{x}')$.

$\rightarrow$ Plutôt que de choisir la transformation non-linéaire $\Phi : \mathcal{X} \rightarrow \mathcal{F}$, on choisit une fonction $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ appelée fonction noyau.

- Elle représente un produit scalaire dans l'espace de représentation intermédiaire. Elle traduit donc la répartition des exemples dans cet espace.

$$k(\boldsymbol{x}, \boldsymbol{x}') = \Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{x}')$$

- Lorsque k est bien choisie, on n'a pas besoin de calculer la représentation des exemples dans cet espace pour calculer cette fonction.
- Permet d'utiliser des représentations non-vectorielles

$\rightarrow$ Le noyau matérialise une notion de proximité adaptée au problème.

Fonction noyau

Exemple

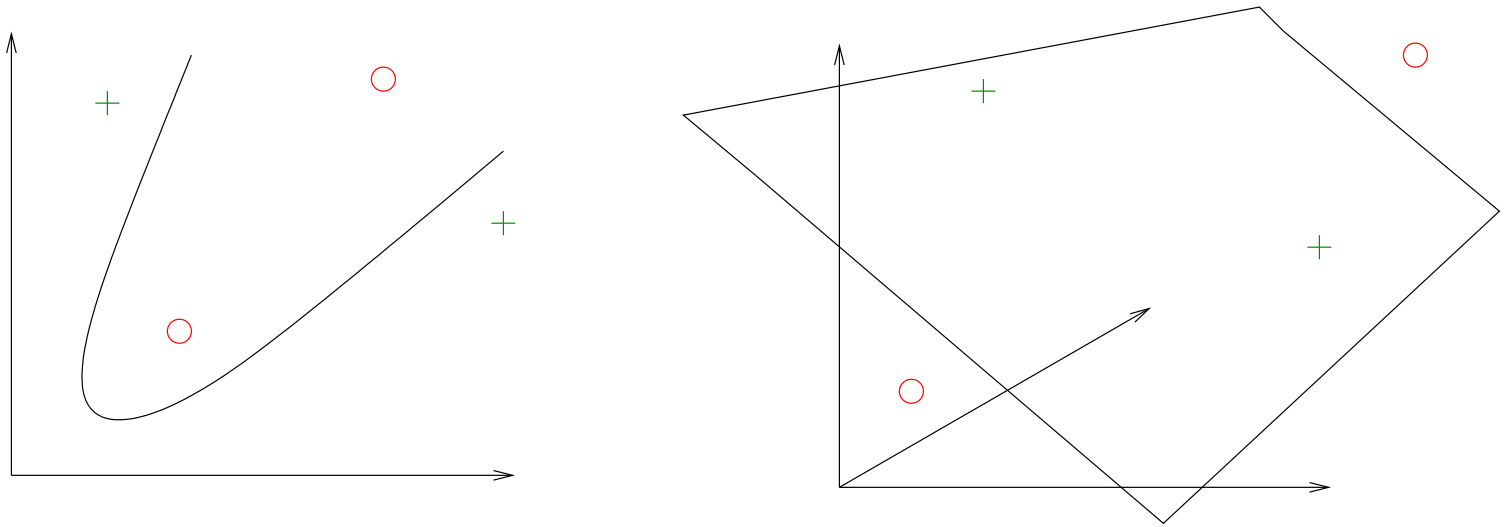
Soit $\boldsymbol{x} = (x_1, x_2)$ et $\Phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$. Dans l'espace intermédiaire, le produit scalaire donne

$$\begin{aligned}\Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{x}') &= x_1^2 {x'_1}^2 + 2x_1x_2x'_1x'_2 + x_2^2 {x'_2}^2 \\ &= (x_1x'_1 + x_2x'_2)^2 \\ &= (\boldsymbol{x} \cdot \boldsymbol{x}')^2\end{aligned}$$

On peut donc calculer $\Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{x}')$ sans calculer Φ.

Fonction noyau

Exemple



$\rightarrow$ Le passage dans $\mathcal{F} = \mathbb{R}^3$ rend possible la séparation linéaire des données.

Fonction noyau

Conditions de Mercer

Une fonction $k(.,.)$ symétrique est un noyau si pour tous les $\boldsymbol{x}_i$ possibles, $(k(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j}$ est une matrice définie positive.

Dans ce cas, il existe un espace $\mathcal{F}$ et une fonction Φ tels que

$$k(\boldsymbol{x}, \boldsymbol{x}') = \Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{x}')$$

- Difficile à vérifier
- Ne donne pas d'indication pour la construction de noyaux
- Ne permet pas de savoir comment est Φ

$\rightarrow$ En pratique, on combine des noyaux simples pour en obtenir de plus complexes.

Fonction noyau

Exemples de noyaux

- Linéaire

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x} \cdot \boldsymbol{x}'$$

- Polynomial

$$k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x} \cdot \boldsymbol{x}')^d \text{ ou } (c + \boldsymbol{x} \cdot \boldsymbol{x}')^d$$

- Gaussien

$$k(\boldsymbol{x}, \boldsymbol{x}') = e^{-\|\boldsymbol{x}-\boldsymbol{x}'\|^2/\sigma}$$

- Laplacien

$$k(\boldsymbol{x}, \boldsymbol{x}') = e^{-\|\boldsymbol{x}-\boldsymbol{x}'\|_1/\sigma}$$

Plan

- Quel est le contexte ?
- Qu'est-ce que c'est ?
- Comment cela marche-t-il ?

→ Pourquoi est-ce utile ?

- Pourquoi est-ce que cela marche ?

Pourquoi est-ce utile ?

- Flexibilité des noyaux

  Noyau $\equiv$ mesure de similitude

- Temps de calcul raisonnable

  Trouver un hyperplan qui minimise l'erreur est NP-complet.

- Vecteurs de support

  Représentation parcimonieuse et éventuellement interprétable.

Exemples

Si les données sont sous forme vectorielle, on peut utiliser un noyau classique.

On peut aussi contruire un noyau spécifique qui travaille plus directement sur la représentation initiale (structure).

- Images 2D: choix d'une résolution et d'une discrétisation des couleurs
  - $\longrightarrow$ vecteur des valeurs des pixels
  - $\longrightarrow$ coefficients d'une transformée en ondelettes
  - $\longrightarrow$ histogramme des couleurs
- Textes: choix d'un dictionnaire (suppression des mots simples)
  - $\longrightarrow$ Sac de mots: vecteur des occurences
  - $\longrightarrow$ Autres attributs (liens, position des mots...)

Exemples

- Séquences d'ADN
  - Fenêtres de taille fixe (sous-séquence) et représentation binaire (un bit par valeur possible)

...	A	T	A	G	C	A	...
	1	0	1	0	0	1	
	0	1	0	0	0	0	
	0	0	0	1	0	0	
	0	0	0	0	1	0	

$\rightarrow$ (1,0,0,0,0,1,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,1,0,0,0)

  - Coefficients d'un modèle probabiliste générateur

Temps de calcul

Complexité

d = dimension des entrées

n = nombre d'exemples d'apprentissage

$$dn^2 \leq \text{Complexite} \leq dn^3$$

Taille de la matrice hessienne: n^2

$\rightarrow$ Méthodes de décomposition

Plan

- Quel est le contexte ?
- Qu'est-ce que c'est ?
- Comment cela marche-t-il ?
- Pourquoi est-ce utile ?

$\rightarrow$ Pourquoi est-ce que cela marche ?

Pourquoi cela marche-t-il ?

Eléments de réponse

Plusieurs réponses possibles

- Maximisation de la marge

  L'ensemble des hyperplans de marge donnée M a une VC dimension bornée par

  $$\frac{R^2}{M^2},$$

  si les X sont dans une boule de rayon R.

- Parcimonie de la représentation

  L'erreur leave-one-out est bornée en moyenne par le nombre de vecteurs support.

En pratique cela donne des bornes relativement prédictives mais très pessimistes.

Régularisation

On peut voir l'algorithme SVM comme la minimisation d'une fonctionnelle régularisée:

$$\min_f \sum_{i=1}^{n} c(f, X_i, Y_i) + \lambda \| f \|_H^2 .$$

1. Choix du paramètre de régularisation
2. Pénalisation linéaire vs quadratique
3. Lien entre parcimonie et choix de la pénalisation

Questions Ouvertes

Noyaux

1. Choix automatique des paramètres
2. Construction de noyaux exotiques (sur des objets structurés)
3. Approche non-paramétrique pour la construction de noyau
4. Classes étendues de noyaux (conditionnellement positifs)

Questions Ouvertes

Généralisation

1. Notion de complexité (covering numbers, fat-shattering dimension)
2. Structure du feature space
3. Structure de l'espace des fonctions possibles (qui dépend des données)
4. Propriétés de la matrice de Gram

Questions Ouvertes

Marge - Parcimonie

1. Bornes de type perceptron (Novikoff)
2. Lien avec la fat-shattering dimension
3. Rôle de la classification a valeurs réelles
4. Lien avec la marge du boosting

Conclusion

1. La technique SVM est d'une grande flexibilité grâce aux noyaux
2. La maximisation de la marge semble être une bonne heurisitque
3. De nombreuses questions restent ouvertes
   - Applications: construction de noyaux, adaptation des paramètres, implémentation efficace, incorporation d'invariances...
   - Théorie: comprendre la généralisation, la complexité, la marge, la parcimonie...

Essayez les SVM !

Etudiez les SVM !

Venez au séminaire/groupe de travail !

Pour en savoir plus

Page du séminaire/GT:

`http://www.math.u-psud.fr/~blanchard/gtsvm/index.html`

→ Références bibliographiques

Un lien utile:

`http://www.kernel-machines.org`

→ Programmes, articles en ligne, tutoriels...