

Apprentissage, réseaux de neurones et  
modèles graphiques (RCP209)  
Algorithmes à noyaux. Applications.

Marin FERECATU & Michel Crucianu  
(prenom.nom@cnam.fr)  
<http://cedric.cnam.fr/vertigo/Cours/ml2/>

Département Informatique  
Conservatoire National des Arts & Métiers, Paris, France

## Plan du cours

2 Objectifs et contenu de l'enseignement

3 Estimation du support d'une densité

4 SVM pour la régression

5 Applications

## Objectif

*“La raison d’être des statistiques, c’est de vous donner raison.” — Abe Burrows*

Algorithmes à noyaux :

- One class SVM (estimation du support d’une densité)
- SVM pour la régression
- Kernel PCA (KPCA)

Applications :

- Classes d’images
- Détection des objets

## Plan du cours

2 Objectifs et contenu de l’enseignement

3 Estimation du support d’une densité

4 SVM pour la régression

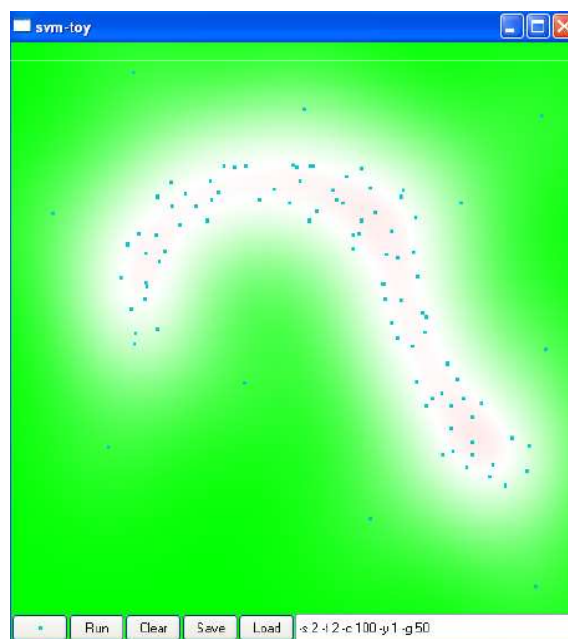
5 Applications

## Estimation du support d'une densité

Estimation du support d'une densité :

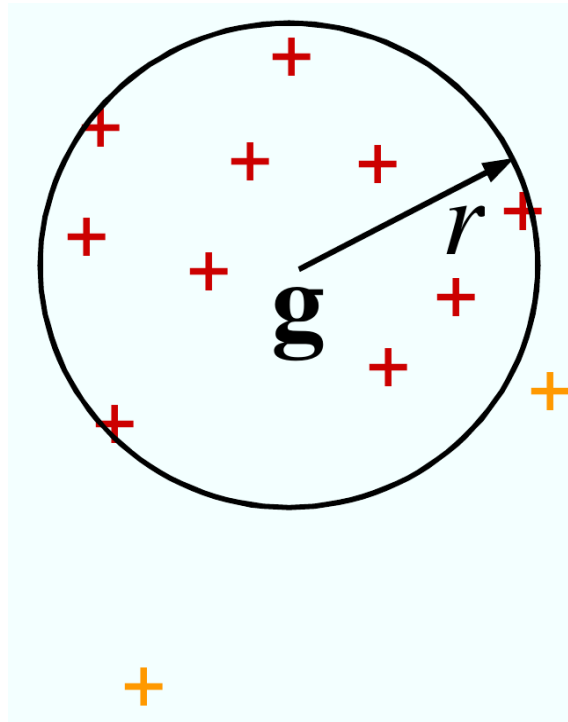
- Les données d'apprentissage  $D = \{x_1, x_2, \dots, x_n \in \mathcal{X}\}$ , issues de variables i.i.d. suivant la densité de probabilité  $p(x)$  inconnue.
- Pas d'étiquettes de classe  $y_i$
- Le problème consiste à décider si une nouvelle observation  $x$  est proche ou non de cet ensemble  $T$ , c.t.d. s'il est tiré de la même distribution.
- On cherche donc à estimer le support de cette densité ← moins de difficultés que pour l'estimation de la densité

## Estimation du support d'une densité



Exemple avec noyau RBF : intensité de la couleur proportionnelle à l'éloignement de la frontière

## Estimation du support d'une densité



Approche SVDD (Support Vector Data Description, Tax & Duin 2004) : trouver dans l'espace d'arrivée  $\mathcal{H}$  la plus petite hypersphère englobant les données

## Support Vector Data Description (SVDD)

Support Vector Data Description (SVDD) :

$$\left\{ \begin{array}{l} \min_{R,g} R^2 + C \sum_{i=1}^n \xi_i \\ \text{avec :} \\ \|x_i - g\|^2 \leq R^2 + \xi_i, i = 1, \dots, n \\ \xi_i \geq 0, i = 1, \dots, n \end{array} \right.$$

- $g$  le centre
- $R$  le rayon
- $C = \frac{1}{\nu n}$  permet de régler la proportion  $\nu$  de points que l'on désire maintenir en dehors de la boule (outliers).

## Support Vector Data Description (SVDD)

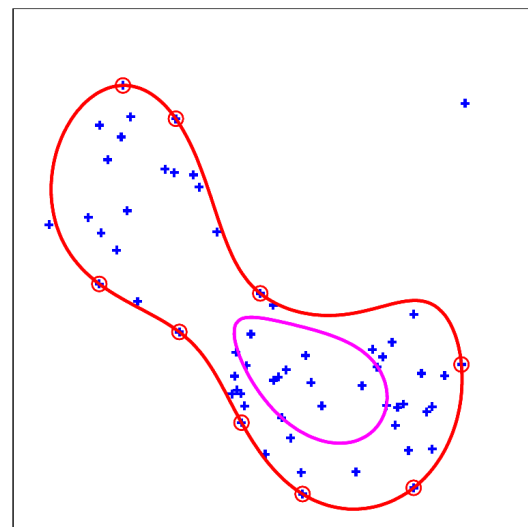
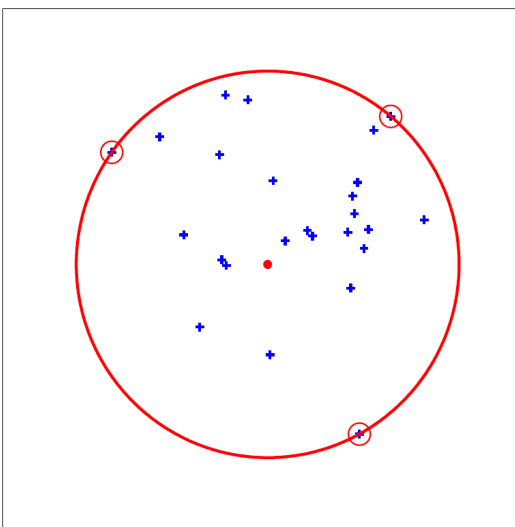
SVDD : Le problème dual

$$\left\{ \begin{array}{l} \min_{\alpha} \frac{1}{2} \alpha^T K \alpha - \frac{1}{2} \alpha^T \text{diag}(K) \\ \text{avec :} \\ e^t \alpha = 1 \\ 0 \leq \alpha_i \leq C, i = 1, \dots, n \end{array} \right.$$

- $K$  est la matrice de Gramm  $K_{ij} = k(x_i, x_j)$
- $g = \sum_{i=1}^n \alpha_i \phi(x_i)$
- Un nouveau point  $x$  appartient au support si  $\|\phi(x) - g\| \leq R^2$ , ou :

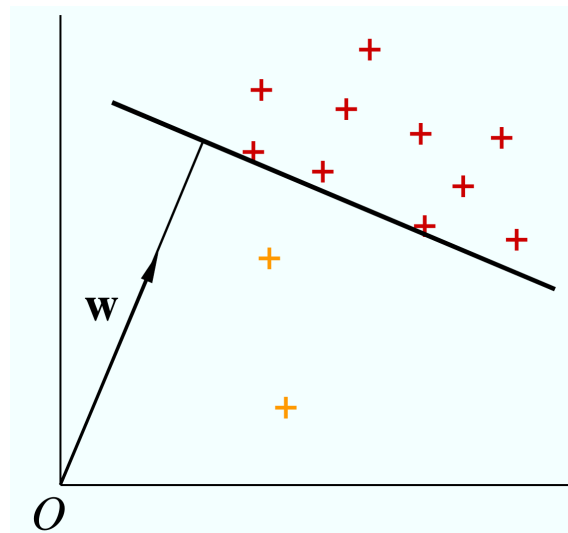
$$K(x, x) - 2 \sum_{i=1}^n \alpha_i K(x_i, x) + \sum_{i,j=1}^n \alpha_i \alpha_j K(x_i, x_j) \leq R^2$$

## Support Vector Data Description (SVDD)



Exemple SVDD : linéaire (à gauche) et noyau gaussien (à droite). A droite, le calcul a été fait pour deux valeurs de  $C$ . Le point en haut à droite est un outlier (il est placé en dehors de l'enveloppe calculée).

## Estimation du support d'une densité



Approche OCSVM (One Class SVM, Schölkopf et al. 2001) trouver dans l'espace d'arrivée  $\mathcal{H}$  l'hyperplan le plus éloigné de l'origine, qui sépare les données de l'origine

## One Class SVM (OCSVM)

One Class SVM (OCSVM) :

$$\left\{ \begin{array}{l} \min_{w, \xi_i, \rho} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i - \rho \\ \text{avec :} \\ w \cdot x_i \geq \rho - \xi_i, i = 1, \dots, n \\ \xi_i \geq 0, i = 1, \dots, n \end{array} \right.$$

- Fonction de décision :  $f(x) = \text{sign}(\langle w, \phi(x) \rangle - \rho)$
- $\rho$  : distance à l'origine
- $C = \frac{1}{\nu n}$  paramètre de régularisation qui permet de contrôler le nombre de outliers.

## One Class SVM (OCSVM)

Le dual est le même que celui des SVDD avec le terme linéaire de la fonction cout en moins :

$$\left\{ \begin{array}{l} \min_{\alpha} \frac{1}{2} \alpha^T K \alpha \\ \text{avec :} \\ e^t \alpha = 1 \\ 0 \leq \alpha_i \leq C, i = 1, \dots, n \end{array} \right.$$

- $K$  est la matrice de Gramm  $K_{ij} = k(x_i, x_j)$
- Fonction de décision :

$$f(x) = \text{sign}(\langle w, \phi(x) \rangle - \rho) = \text{sign}\left(\sum_{i=1}^n \alpha_i K(x_i, x) - \rho\right)$$

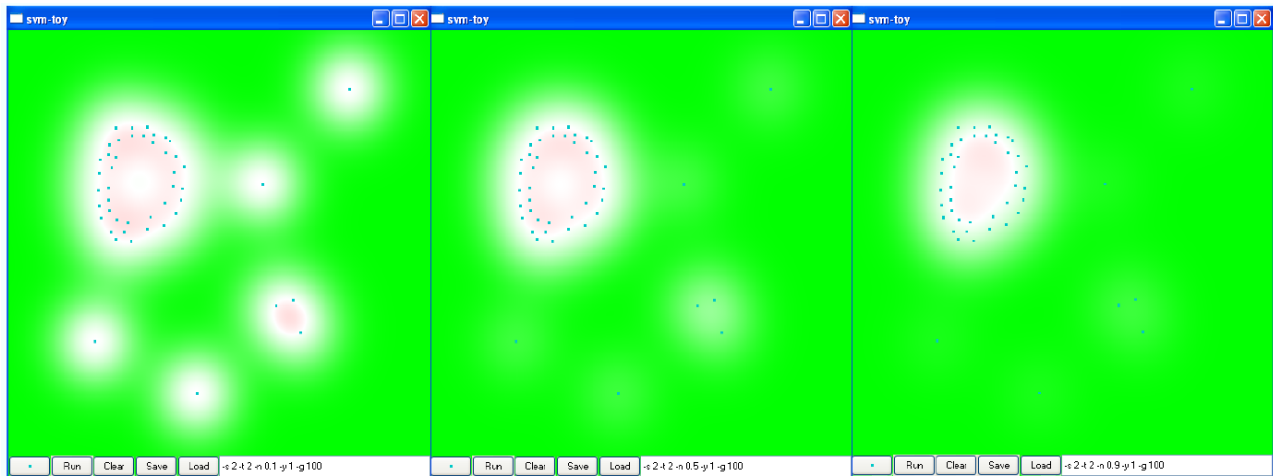
avec  $\rho = \langle w, \phi(x_s) \rangle = \sum_{i=1}^n \alpha_i K(x_i, x_s)$

- $C = \frac{1}{\nu n}$  paramètre de régularisation qui permet de contrôler le nombre de outliers.

## One Class SVM (OCSVM)

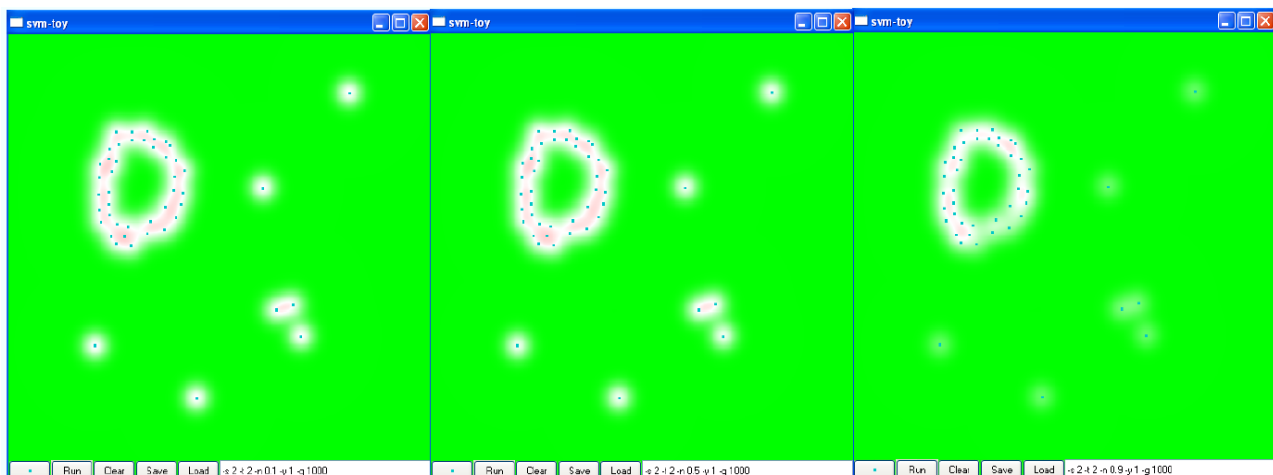
- Dans les deux formulations  $\nu \in (0, 1]$  et  $\nu n = 1/C$  est :
  - Une borne supérieure pour la fraction de outliers
  - Une borne inférieure pour la fraction de vecteurs de support
- Bornes de généralisation : la probabilité pour que de nouveaux exemples (tirages i.i.d. suivant la densité  $p(x)$ ) soient en dehors d'une région un peu plus grande que le support déterminé ne sera pas supérieure de beaucoup à la fraction de outliers dans les données d'apprentissage

## One Class SVM (OCSVM)



Noyau RBF avec  $\gamma=100$  :  $\nu=0,1$ ,  $\nu=0,5$ ,  $\nu=0,9$

## One Class SVM (OCSVM)



Noyau RBF avec  $\gamma=1000$  :  $\nu=0,1$ ,  $\nu=0,5$ ,  $\nu=0,9$



## Plan du cours

2 Objectifs et contenu de l'enseignement

3 Estimation du support d'une densité

4 SVM pour la régression

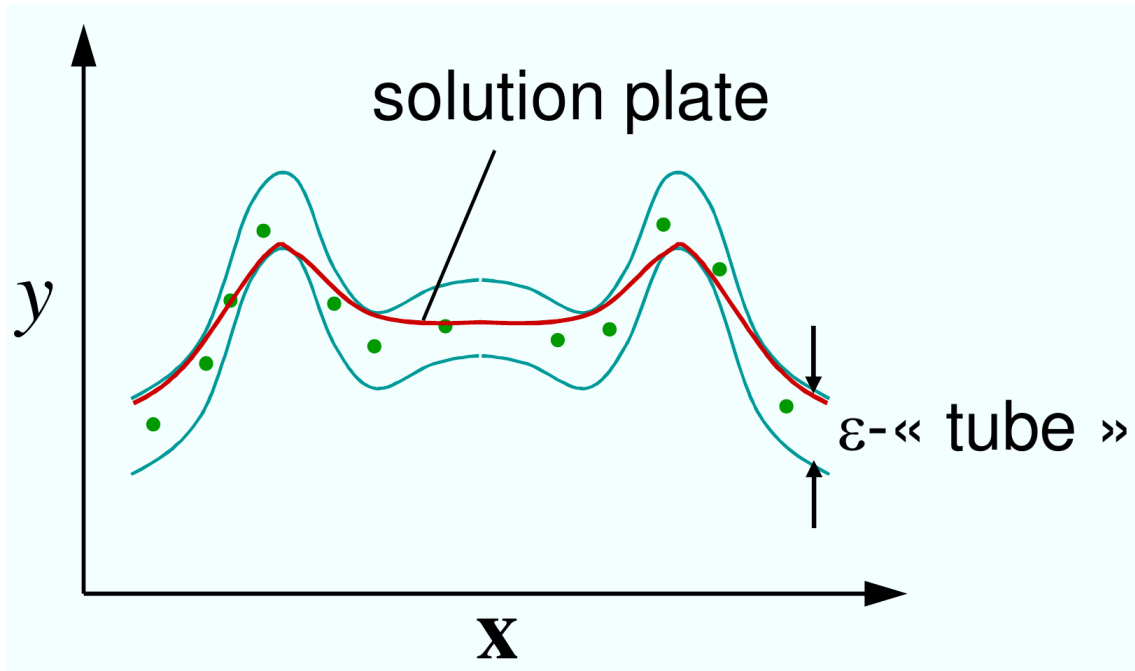
5 Applications

## SVM pour la régression (SVR)

### SVM pour la régression (SVR) :

- Données d'apprentissage  $\mathcal{D} = \{(x_i, y_i); i = 1, \dots, n\}$
- $x_i \in \mathcal{X}, y_i \in \mathcal{R}$
- En **régression  $\epsilon$ -SV** on cherche une fonction  $f : \mathcal{X} \rightarrow \mathcal{R}$  aussi "plate" que possible et  $|f(x_i) - y_i| < \epsilon$
- On cherchera des solutions de la forme  $f(x) = \langle w, \phi(x) \rangle + b$  dans l'espace  $\mathcal{H}$  d'arrivée.
- La condition d'aplatissement se traduit par la minimisation de  $\|w\|^2 = \langle w, w \rangle$

## SVM pour la régression

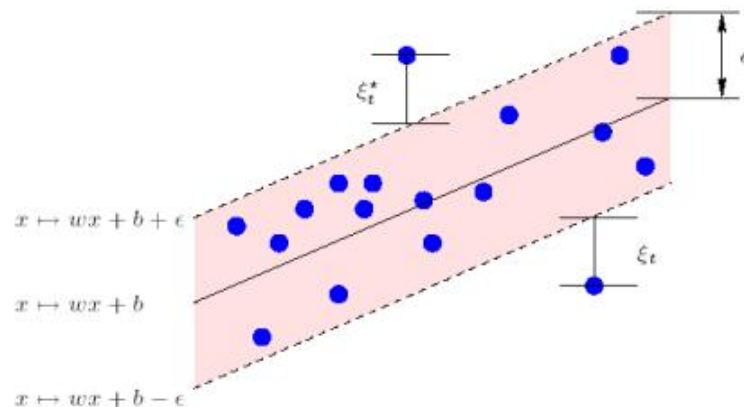


SVM pour la régression : on cherche une solution aussi plate que possible sans s'éloigner trop des points d'apprentissage (en vert).

## SVM pour la régression

La régression  $\epsilon$ -SV correspond à l'utilisation de la fonction de coût  $\epsilon$ -insensible :

$$|\xi|_{\epsilon} = \begin{cases} 0 & \text{si } |\xi| < \epsilon \\ |\xi| - \epsilon & \text{sinon} \end{cases}$$



## SVM pour la régression

Comme en discrimination, on accepte quelques erreurs au-delà de  $\epsilon$  et on introduit les « variables d'assouplissement  $\xi_i, \xi_i^*$

Le problème d'optimisation sera :

$$\left\{ \begin{array}{l} \min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \\ \text{avec :} \\ y_i - \langle w, \phi(x_i) \rangle - b \leq \epsilon + \xi_i, i = 1, \dots, n \\ \langle w, \phi(x_i) \rangle + b - y_i \leq \epsilon + \xi_i^*, i = 1, \dots, n \\ \xi_i, \xi_i^* \geq 0, i = 1, \dots, n \end{array} \right.$$

- La constante  $C > 0$  permet de choisir le point d'équilibre entre l'aplatissement de la solution et l'acceptation d'erreurs au-delà de  $\epsilon$

## SVM pour la régression

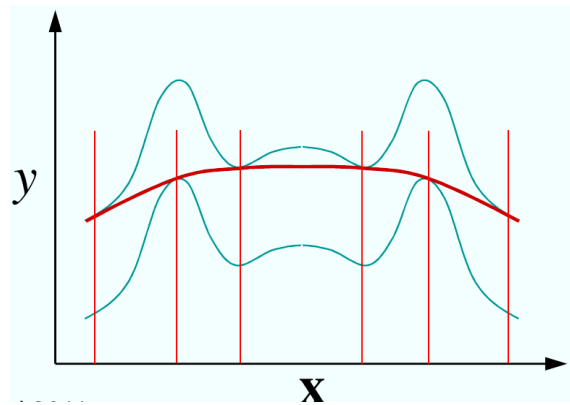
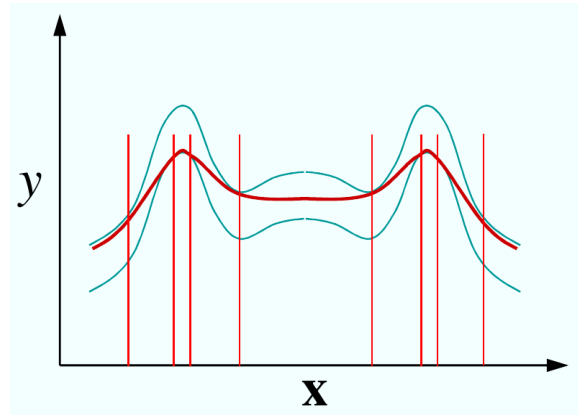
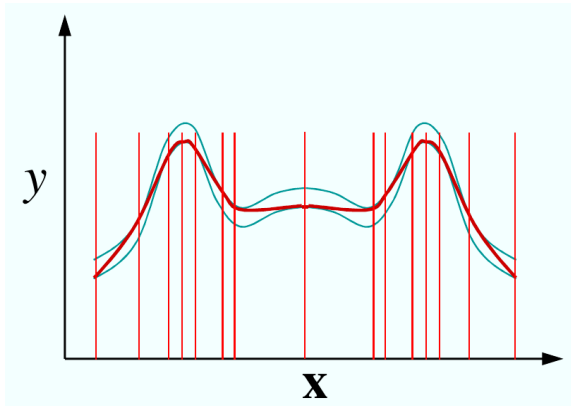
Avec les multiplicateurs de Lagrange on obtient le problème dual :

$$\left\{ \begin{array}{l} \min_{\alpha, \alpha^*} \frac{1}{2} \sum_{i,j=1}^n (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j) + \epsilon \sum_{i=1}^n (\alpha_i + \alpha_i^*) - \sum_{i=1}^n y_i (\alpha_i - \alpha_i^*) \\ \text{avec :} \\ 0 \leq \alpha_i, \alpha_j \leq C, i, j = 1, \dots, n \\ \sum_{i=1}^n (\alpha_i - \alpha_i^*) = 0 \end{array} \right.$$

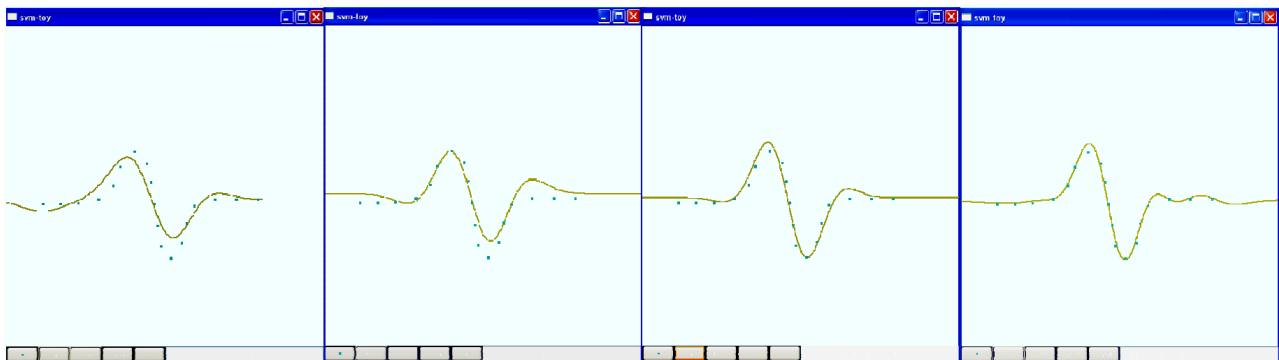
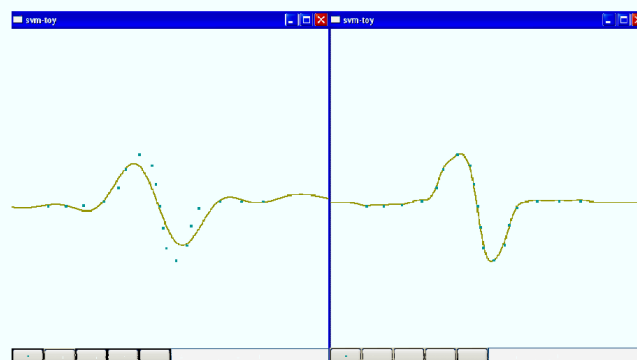
- Tous les points d'apprentissage à l'intérieur du  $\epsilon$  - tube ont  $\alpha_i = \alpha_i^* = 0$ . Les points qui ont  $\alpha_i, \alpha_i^* \neq 0$  sont appelés vecteurs de support.
- Comme  $w = \sum_{i=1}^n (\alpha_i - \alpha_i^*) \phi(x_i)$ , la fonction recherchée sera :

$$f(x) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

## SVM pour la régression



## SVM pour la régression

Noyau RBF avec  $\gamma=100$  :  $\nu=0,1$  et  $\epsilon=0,1$ ,  $\nu=0,9$  et  $\epsilon=0,1$ ,  $\nu=0,1$  et  $\epsilon=0,01$ ,  $\nu=0,9$  et  $\epsilon=0,01$ Noyau RBF avec  $\gamma=40$  (gauche) et avec  $\gamma=1000$  ( $\nu=0,9$  et  $\epsilon=0,01$  dans les deux cas)

## Algorithmes à noyaux

Algorithmes à noyaux :

- Kernel PCA (Principal Component Analysis) Scholkopf et al. 2001
- Kernel CCA (Cannonical Correlation Analysis) Hardoon et al. 2003
- Kernel FDA (Factorial Discriminant Analysis) Roth et al. 2000
- Tout algorithme qui utilise des produits scalaires entre les échantillons peut être non-linéarisé par le "truc à noyaux"

## Plan du cours

2 Objectifs et contenu de l'enseignement

3 Estimation du support d'une densité

4 SVM pour la régression

5 Applications

## Applications

- Finance (évolution des prix, valeurs en bourse, etc.)
- Structure des protéines (Protein Folding)
- Génomique (microarray gene expression data)
- Reconnaissance de visage
- Détections des catastrophes, forecasting
- Images satellite et surveillance
- Diagnostic médical (cancer du sein)
- En physique ; exemple : Particle and Quark-Flavour Identification in
- High Energy Physics (Classifying LEP Data with Support Vector Algorithms by Schölkopf et al. AIHENP'99)

## Références

Livres, articles, web :

- Steinwart, Christmann, *Support Vector Machines*, Springer 2008
- Scholkopf, Smola, *Learning with Kernels*, The MIT Press, 2001
- Hastie, Tibshirani, Friedman, *The elements of statistical learning : Data mining, inference, and prediction*, New York, Springer Verlag, 2006
- —, *Machines à vecteurs supports (WikiStat)*, <http://wikistat.fr>
- Tax and Duin, *Support Vector Data Description*, Machine Learning, 54(1), 2004
- Haroon, Szedmak, Shawe-Taylor, *Canonical correlation analysis ; an overview with application to learning methods*, Tech. Rep., University of London, 2003.
- Roth, Steinhage, *Nonlinear discriminant analysis using kernel functions*, Advances in Neural Information Processing Systems, 2000.