le cnam

Apprentissage, réseaux de neurones et modèles graphiques

Chapitre 7 : Machines à vecteurs support

Michel Crucianu
avec des contributions de Jean-Philippe Tarel
http://cedric.cnam.fr/~crucianm/ml.html

le cnam

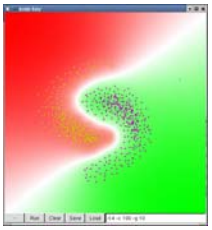
Contenu du chapitre

- SVM pour la discrimination
  - Maximisation de la marge
  - Astuce des noyaux (*kernel trick*)
- Estimation du support d'une distribution
- SVM pour la régression
- Méthodes à noyaux et analyse des données
- Ingénierie des noyaux
  - Noyaux valides et « bons » noyaux
  - Noyaux pour différents types de données
  - Combinaison de noyaux, noyaux hybrides
- Exemples d'utilisation des SVM

le cnam

SVM pour la discrimination

- *Support Vector Machines*
  1. Séparateurs à Vastes Marges
     - SVM ne servent pas seulement à séparer (aussi à la régression, …)
     - SVM pas les seuls séparateurs à vastes marges (autre ex. : *boosting*)
  2. Machines à vecteurs support

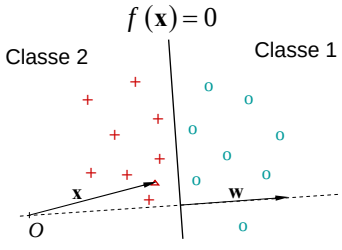


Exemple simple avec noyau angulaire
Intensité de la couleur proportionnelle à l'éloignement de la frontière
(outil employé : version maison de svm-toy)

le cnam

Classes linéairement séparables

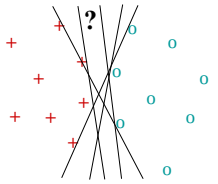
- Données d'apprentissage $D_n = \{(\mathbf{x}_i, y_i)\}_{1 \le i \le n},\ y_i \in \{-1, +1\}$
- On cherche une fonction de décision $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$, basée sur l'hyperplan $\mathbf{w}^T \mathbf{x} + b = 0$, $\mathbf{w}$ étant un vecteur normal à l'hyperplan
  - Affectation à la classe 1 si $f(\mathbf{x}) > 0$
  - Affectation à la classe 2 si $f(\mathbf{x}) < 0$
- Séparabilité linéaire : $\exists\, \mathbf{w}, b$ tels que $y_i\ f(\mathbf{x}_i) > 0$ pour $1 \le i \le n$
- Remarque : si la condition est valable pour $\mathbf{w}, b$, alors elle est valable pour $k\mathbf{w}, kb, \forall\, k > 0$



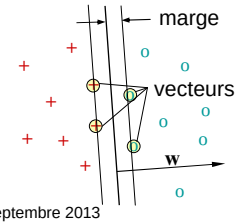
le cnam

Discrimination linéaire et marge

- Quelle séparation choisir quand plusieurs (éventuellement une infinité) sont possibles ?



- Une possibilité : choisir la séparation qui maximise la marge (distance minimale entre un exemple et la surface de séparation)



en fonction de $\mathbf{w}$: marge $= \left| \mathbf{x}_s^T \mathbf{w} + b \right| / \| \mathbf{w} \|$

normalisation : $\left| \mathbf{x}_s^T \mathbf{w} + b \right| = 1$ pour tout vecteur support $\mathbf{x}_s$

donc marge $= 1/\| \mathbf{w} \|$

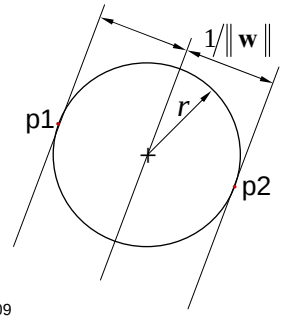
le cnam

Pourquoi s'intéresser à la marge

1. Marge et VC-dimension [Vap98] : pour des données qui se situent à l'intérieur d'une (hyper)sphère de rayon r, la VC-dimension h de la famille de (hyper)plans de marge $1/\| \mathbf{w} \| \geq m$ est bornée,

$$h \leq (r/m)^2 + 1$$

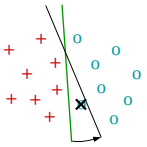
Exemple : dans $\mathbb{R}^2$ la VC-dimension des droites est $h = 3$, mais la VC-dimension des droites de marge $1/\| \mathbf{w} \| \geq r$ est $h = 2$ pour des données du disque de rayon r



le cnam

Pourquoi s'intéresser à la marge (2)

2. Vecteurs support et *leave one out cross-validation* [SS02]
   - Validation croisée *leave one out* : on retire un exemple de D_n et on apprend sur les $n-1$ restants, on évalue l'erreur sur l'exemple retiré, on répète la procédure pour chaque exemple de D_n ; la moyenne des erreurs ainsi obtenues constitue une estimation du risque espéré
   - Le risque espéré a comme borne supérieure $E\left[s_{(n)}\right]/n$, où $E\left[s_{(n)}\right]$ est l'espérance du nombre de vecteurs support obtenus en maximisant la marge sur tous les choix possibles d'ensembles d'apprentissage de taille n



25 septembre 2013 RCP209 7

le cnam

Problème d'optimisation

- Maximiser la marge : optimisation sous contraintes d'inégalité (problème primal)

$$\begin{cases} \min\limits_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \\ y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1, \quad 1 \leq i \leq n \end{cases}$$

- En introduisant les multiplicateurs de Lagrange $\boldsymbol{\alpha}$

$$\mathrm{L}(\mathbf{w},b,\boldsymbol{\alpha}) = -\frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n}\alpha_i\left[y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1\right]$$

- Des conditions nécessaires :

$$\frac{\partial \mathrm{L}}{\partial b}\left(\mathbf{w}^*,b^*,\boldsymbol{\alpha}^*\right) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n}\alpha_i^* y_i = 0$$

$$\frac{\partial \mathrm{L}}{\partial \mathbf{w}}\left(\mathbf{w}^*,b^*,\boldsymbol{\alpha}^*\right) = 0 \quad \Rightarrow \quad \mathbf{w}^* = \sum_{i=1}^{n}\alpha_i^* y_i \mathbf{x}_i$$

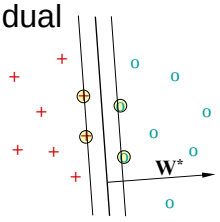
25 septembre 2013 RCP209 8

le cnam

Résolution du problème d'optimisation

- Par substitution on obtient le problème dual

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}} \sum\limits_{i=1}^{n} \alpha_i - \sum\limits_{i,j=1}^{n} \alpha_i \alpha_j y_i y_i \mathbf{x}_i^T \mathbf{x}_j \\ \alpha_i \geq 0, \quad 1 \leq i \leq n \\ \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$



- Les vecteurs support sont ceux pour lesquels $\alpha_i \neq 0$
- b^* est obtenu à partir de la condition $\left| \mathbf{x}_s^T \mathbf{w}^* + b^* \right| = 1$ valable pour les vecteurs support
- La fonction de décision permettant de classer une nouvelle observation $\mathbf{x}$ sera $f^*(\mathbf{x}) = \sum\limits_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i^T \mathbf{x} + b^*$

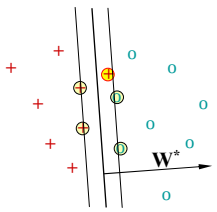
25 septembre 2013 RCP209 9

le cnam

Cas des données non séparables

- Si les données d'apprentissage ne sont pas linéairement séparables, on emploie les contraintes assouplies $y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i$ pour $1 \leq i \leq n$ et avec $\xi_i \geq 0$
- Le nouveau problème d'optimisation doit inclure une pénalité pour les erreurs ($\xi_i > 0$) faites sur l'ensemble d'apprentissage

$$\begin{cases} \min\limits_{\mathbf{w},b} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{n} \xi_i \\ y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, 1 \leq i \leq n, \xi_i \geq 0 \end{cases}$$



25 septembre 2013 RCP209 10

le cnam

Cas des données non séparables (2)

- Le problème dual résultant est

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}} \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ C \geq \alpha_i \geq 0, \quad 1 \leq i \leq n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

- C est une constante de régularisation (la régularisation est d'autant plus forte que C est proche de 0 !)
- Pour une meilleure stabilité numérique, b^* est obtenu à partir de la moyenne sur l'ensemble I des vecteurs pour lesquels $0 < \alpha_i < C$

$$b^* = \left(1/|I|\right) \sum_{i \in I} \left(y_i - \sum_{j=1}^{n} \alpha_j^* y_j \mathbf{x}_i^T \mathbf{x}_j \right)$$

- La fonction de décision permettant de classer une nouvelle observation $\mathbf{x}$ est toujours $f^*(\mathbf{x}) = \sum\limits_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i^T \mathbf{x} + b^*$

le cnam

Conditions KKT

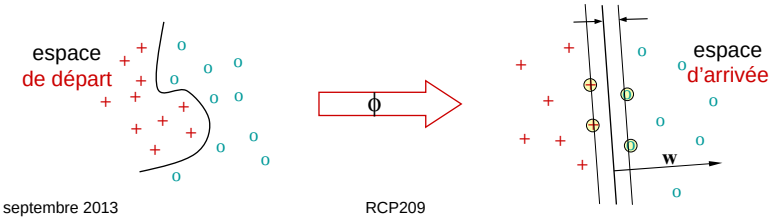
- Conditions nécessaires et parfois suffisantes d'optimalité (Karush-Kuhn-Tucker) pour le problème dual :

$$\begin{aligned} \alpha_i = 0 &\Rightarrow y_i f(\mathbf{x}_i) > 1 \quad \text{et} \quad \xi_i = 0 \quad \leftarrow \text{hors marge} \\ 0 < \alpha_i < C &\Rightarrow y_i f(\mathbf{x}_i) = 1 \quad \text{et} \quad \xi_i = 0 \quad \leftarrow \text{sur la marge} \\ \alpha_i = C &\Rightarrow y_i f(\mathbf{x}_i) < 1 \quad \text{et} \quad \xi_i \geq 0 \quad \leftarrow \text{dans la marge ou mal classés} \end{aligned}$$

le cnam

Astuce des noyaux (*kernel trick*)

- Comment étendre ces résultats à des séparateurs non linéaires ?
- Principe : transposer les données dans un autre espace (en général de plus grande dimension) dans lequel elles sont linéairement séparables (ou presque)
- Transformation $\phi : \mathbb{R}^d \rightarrow \mathcal{H},\ \mathbf{x} \rightarrow \phi(\mathbf{x})$, $\mathcal{H}$ espace de Hilbert
- Sous certaines conditions, l'existence de ϕ et de $\mathcal{H}$ est garantie et ϕ est associée à une fonction noyau $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, telle que $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = K(\mathbf{x}_i, \mathbf{x}_j) \Rightarrow$ calculs faits dans l'espace de départ !

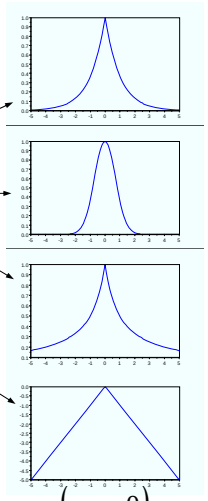


25 septembre 2013 RCP209 13

le cnam

Exemples de noyaux

- Linéaire : $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$
- Exponentiel : $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \| \mathbf{x}_i - \mathbf{x}_j \|\right)$
- Gaussien (RBF) : $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \| \mathbf{x}_i - \mathbf{x}_j \|^2\right)$
- Hyperbolique : $K(\mathbf{x}_i, \mathbf{x}_j) = 1 / \left(\varepsilon + \gamma \| \mathbf{x}_i - \mathbf{x}_j \|\right)$
- Angulaire : $K(\mathbf{x}_i, \mathbf{x}_j) = -\| \mathbf{x}_i - \mathbf{x}_j \|$
- Puissance : $K(\mathbf{x}_i, \mathbf{x}_j) = -\| \mathbf{x}_i - \mathbf{x}_j \|^{\beta}$ $(0 < \beta \leq 2)$



25 septembre 2013 RCP209 14

le cnam

Exemple de noyau polynomial

- Noyau polynomial de degré 2 : $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i^T \mathbf{x}_j)^2$
- Dans un espace 2D, $\mathbf{x}_i = (x_i, y_i)$
- Donc, en développant,

$$K(\mathbf{x}_i, \mathbf{x}_j) = \begin{pmatrix} 1 \\ \sqrt{2}x_i \\ \sqrt{2}y_i \\ x_i^2 \\ \sqrt{2}x_i y_i \\ y_i^2 \end{pmatrix}^T \begin{pmatrix} 1 \\ \sqrt{2}x_j \\ \sqrt{2}y_j \\ x_j^2 \\ \sqrt{2}x_j y_j \\ y_j^2 \end{pmatrix}$$

→ Produit scalaire en dimension 6 !

le cnam

Résolution du problème d'optimisation

- Dans la formulation du problème d'optimisation, les produits scalaires $\mathbf{x}_i^T \mathbf{x}_j$ sont remplacés par $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = K(\mathbf{x}_i, \mathbf{x}_j)$, donc pour des données qui ne sont pas complètement séparables dans l'espace $\mathcal{H}$ le problème primal est

$$\begin{cases} \min_{\mathbf{w},b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \\ y_i \left(\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle + b \right) \geq 1 - \xi_i, 1 \leq i \leq n, \xi_i \geq 0 \end{cases}$$

et le problème dual correspondant :

$$\begin{cases} \max_{\boldsymbol{\alpha}} \sum_{i=1}^n \alpha_i \\ \quad - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ C \geq \alpha_i \geq 0, \quad 1 \leq i \leq n \\ \sum_{i=1}^n \alpha_i y_i = 0 \end{cases}$$

avec b^* donné par

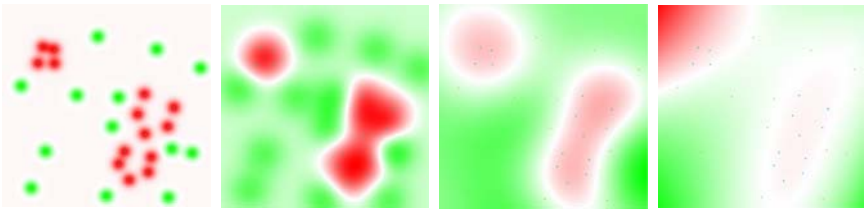
$$b^* = \frac{1}{|I|} \sum_{i \in I} \left(y_i - \sum_{j=1}^n \alpha_j^* y_j K(\mathbf{x}_i, \mathbf{x}_j) \right)$$

et la fonction de décision

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

le cnam

Effet de l'échelle pour noyau RBF



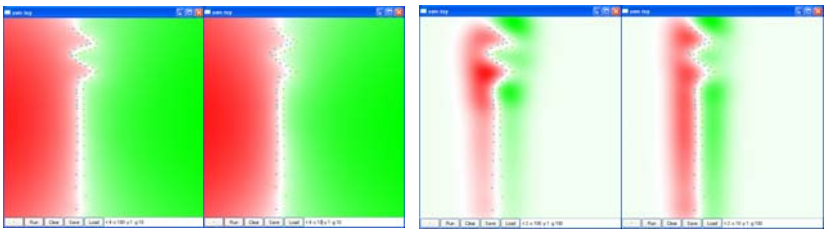
1/γ=0.001, 27 SV | 1/γ=0.01, 26 SV | 1/γ=0.1, 12 SV | 1/γ=1.0, 10 SV

Sur-apprentissage : mauvaise généralisation, apprentissage du bruit

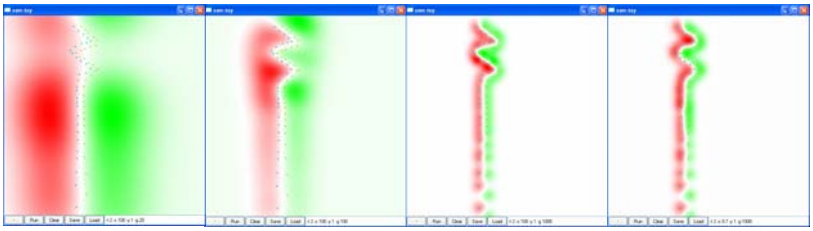
Sous-apprentissage : mauvaise généralisation, frontière imprécise

le cnam

Effet de l'échelle et de C pour RBF



Noyau angulaire : C=100 (gauche) et 10 (droite) Noyau RBF γ=100 : C=100 (gauche) et 10 (droite)



Noyau RBF avec γ=20 & C=100, puis γ=100 & C=100, puis γ=1000 & C=100, puis γ=1000 & C=0,7

le cnam

Formulation alternative : ν-SVM

- La signification de la constante C n'est pas très intuitive, on re-formule le problème d'optimisation de la façon suivante

$$\begin{cases} \min\limits_{\boldsymbol{\alpha}} \sum\limits_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ 1/n \geq \alpha_i \geq 0, \quad 1 \leq i \leq n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \sum_{i=1}^{n} \alpha_i \geq \nu \end{cases}$$

- Sous certaines conditions, $\nu \in (0,\ 1]$ est
  - Une borne supérieure pour la fraction d' « erreurs de marge »
  - Une borne inférieure pour la fraction de vecteurs de support

le cnam

Types de bases

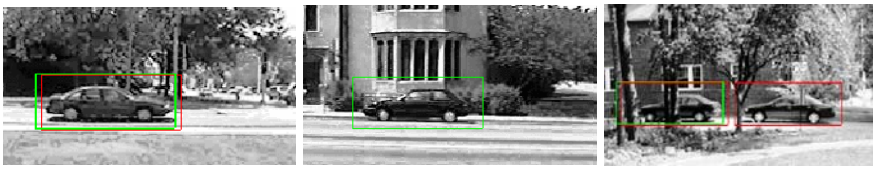
- Base d'**apprentissage** (*training database*) : données étiquetées permettant de construire le modèle de l'objet d'intérêt, échantillonnées (en général) pour être représentatives
- Base de **validation** (*validation database*) : données étiquetées différentes de la base d'apprentissage, permettant d'optimiser le choix des paramètres (C, noyau, etc.)
- Base de **test** (*test database*) : données étiquetées différentes des bases précédentes, permettant d'évaluer les performances du modèle appris

le cnam

Exemple applicatif : détection véhicule

- Base d'apprentissage

- Résultats sur base de test



le cnam

Courbes ROC : sensitivité, spécificité

- Évaluation de la discrimination entre présence d'une classe et absence de la classe (ex. : détection d'objets)
- Table de vérité

Détecteur \ Vérité terrain	Classe présente	Classe absente
Classe détectée	Vrai Positif	Faux Positif
Classe non détectée	Faux Négatif	Vrai Négatif

- Sensitivité et spécificité pour un seuil de décision fixé

$$\text{TPR} = \text{Sensibilité} = \frac{VP}{\text{Total positif}} = \frac{VP}{VP + FN}$$

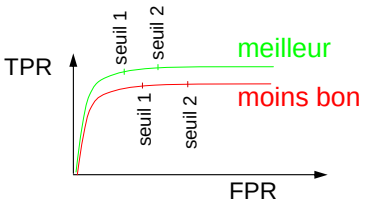
$$\text{FPR} = 1 - \text{Spécificité} = \frac{FP}{\text{Total négatif}} = \frac{FP}{VN + FP} = 1 - \frac{VN}{VN + FP}$$

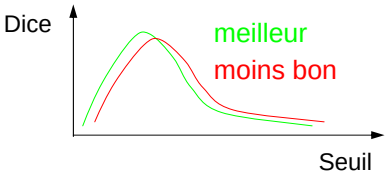
le cnam

Évaluation comparative

- Courbe ROC (*Receiver Operating Characteristics*) : sensibilité vs. 1-spécificité pour différents seuils de décision
- En cas d'intersection des courbes ROC choisir un point sur la courbe ROC par exemple avec l'indice Dice maximum

$$\text{Dice} = \frac{2\,VP}{2\,VP + FP + FN}$$

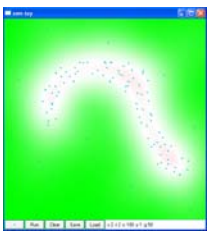




le cnam

Support d'une distribution

- Soit $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ un ensemble d'observations dans $\mathcal{X}$, issues de variables i.i.d. suivant la densité de probabilité $p(\mathbf{x})$ (inconnue)
- On cherche à estimer le support de cette densité ← moins de difficultés que pour l'estimation de la densité

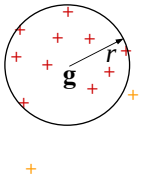


Exemple simple avec noyau RBF
Intensité de la couleur proportionnelle à l'éloignement de la frontière
(outil employé : version maison de svm-toy)

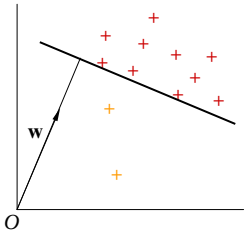
le cnam

One-class SVM : alternatives

1. Trouver dans l'espace d'arrivée $\mathcal{H}$ la plus petite hyper-sphère englobant les données
2. Trouver dans l'espace d'arrivée $\mathcal{H}$ l'hyperplan le plus éloigné de l'origine, qui sépare les données de l'origine



Formulation 1 [TD99] : hyper-sphère englobant les données

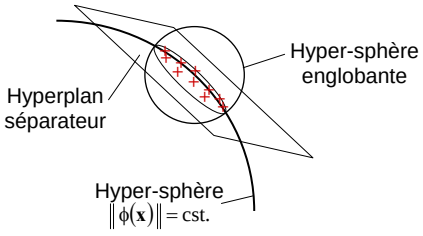


Formulation 2 [SPS99] : hyperplan séparant de l'origine

le cnam

Condition d'équivalence

- Les deux formulations sont équivalentes si le noyau employé satisfait la condition $K(\mathbf{x},\mathbf{x}) = R^2$(const.), $\forall\, \mathbf{x} \in \mathcal{X}$
  - On a alors $\|\phi(\mathbf{x})\| = R$, donc l'image de l'espace de départ dans l'espace d'arrivée se situe sur l'hyper-sphère de rayon R
  - L'hyper-sphère de rayon r minimal ou l'hyperplan le plus éloigné de l'origine sépareront le même segment de l'hyper-sphère de rayon R :



le cnam

One-class SVM : formulation 1

- En notant par $\mathbf{g}$ le centre et par r le rayon de l'hyper-sphère et en faisant appel aux « variables d'assouplissement » $\xi_i \geq 0$, le problème d'optimisation à résoudre est

$$\begin{cases} \min\limits_{\mathbf{g},r,\xi} \left(r^2 + (1/\nu n) \sum_{i=1}^{n} \xi_i \right) \\ \left\| \phi(\mathbf{x}_i) - \mathbf{g} \right\|^2 \leq r^2 + \xi_i, 1 \leq i \leq n, \xi_i \geq 0 \end{cases}$$

- Avec les multiplicateurs de Lagrange on obtient le problème dual

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i,j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\ (1/\nu n) \geq \alpha_i \geq 0, \quad 1 \leq i \leq n \\ \sum_{i=1}^{n} \alpha_i = 1 \end{cases} \quad \text{avec} \quad \mathbf{g} = \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x}_i)$$

Un nouveau point $\mathbf{x}$ appartient au support si $\left\| \phi(\mathbf{x}) - \mathbf{g} \right\|^2 \leq r^2$, ou

$$K(\mathbf{x}, \mathbf{x}) - 2 \sum_{i=1}^{n} \alpha_i K(\mathbf{x}, \mathbf{x}_i) - \sum_{i,j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \leq r^2$$

25 septembre 2013 RCP209 27

le cnam

One-class SVM : formulation 2

- Avec la fonction de décision $f(\mathbf{x}) = \operatorname{sgn}\left(\langle \mathbf{w}, \phi(\mathbf{x}) \rangle - \rho \right)$ en faisant appel aux « variables d'assouplissement » $\xi_i \geq 0$, avec $\nu \in (0, 1]$, le problème d'optimisation à résoudre est

$$\begin{cases} \min\limits_{\mathbf{w},\xi,\rho} \frac{1}{2} \left\| \mathbf{w} \right\|^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i - \rho \\ \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle \geq \rho - \xi_i, 1 \leq i \leq n, \xi_i \geq 0 \end{cases}$$

- Avec les multiplicateurs de Lagrange on obtient le problème dual

$$\begin{cases} \min\limits_{\boldsymbol{\alpha}} \sum_{i,j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\ (1/\nu n) \geq \alpha_i \geq 0, \quad 1 \leq i \leq n \\ \sum_{i=1}^{n} \alpha_i = 1 \end{cases}$$

avec $\rho = \langle \mathbf{w}, \phi(\mathbf{x}_s) \rangle = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x}_s)$
pour tout $\mathbf{x}_s$ tel que $(1/\nu n) > \alpha_i > 0$
Un nouveau point $\mathbf{x}$ appartient au support si $\operatorname{sgn}\left(\langle \mathbf{w}, \phi(\mathbf{x}) \rangle - \rho \right) = 1$

25 septembre 2013 RCP209 28

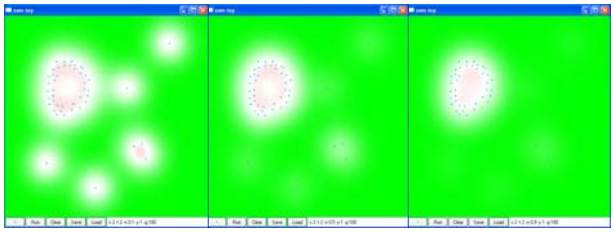
le cnam

One-class SVM : formulation 2 (2)

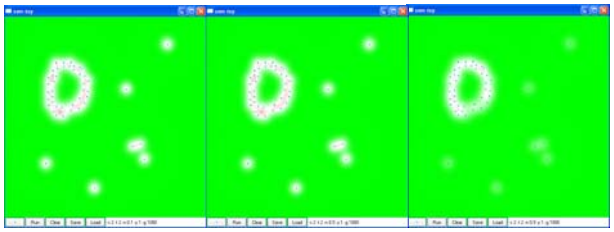
- Dans les deux formulations, si pour la solution $\rho \neq 0$, alors $\nu \in (0,1]$ est
  - Une borne supérieure pour la fraction de *outliers*
  - Une borne inférieure pour la fraction de vecteurs de support
- Bornes de généralisation (formulation 2, [SPS99]) : la probabilité pour que de nouveaux exemples (tirages i.i.d. suivant la densité $p(\mathbf{x})$) 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 $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$

le cnam

One-class SVM : exemples



Noyau RBF avec γ=100 : ν=0,1, ν=0,5, ν=0,9

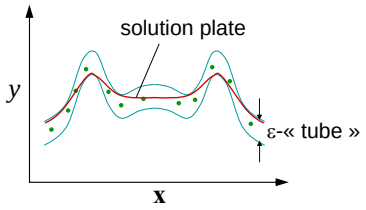


Noyau RBF avec γ=1000 : ν=0,1, ν=0,5, ν=0,9

le cnam

SVM pour la régression

- Données d'apprentissage $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{1 \le i \le n}$, $\mathbf{x} \in \mathcal{X}$, $y \in \mathbb{R}$
- En régression ε-SV on cherche une fonction $f : \mathcal{X} \to \mathbb{R}$, aussi « plate » que possible, et telle que $| f(\mathbf{x}_i) - y_i | \le \varepsilon$
- On cherchera des solutions de la forme $f(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b$ dans l'espace d'arrivée, la condition d'aplatissement se traduisant par la minimisation de $\| \mathbf{w} \|^2 = \langle \mathbf{w}, \mathbf{w} \rangle$



Remarque : la régression ε-SV correspond à l'utilisation de la fonction de coût ε-insensible

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

le cnam

Problème d'optimisation

- Comme en discrimination, on accepte quelques erreurs au-delà de ε et on introduit les « variables d'assouplissement » $\xi_i, \xi_i^* \ge 0$
- Le problème d'optimisation résultant sera

$$\begin{cases} \min\limits_{\mathbf{w},\xi,b} \dfrac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ y_i - \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle - b \le \varepsilon + \xi_i \\ \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, \ 1 \le i \le n \end{cases}$$

- 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 ε

le cnam

Résolution du problème d'optimisation

- Avec les multiplicateurs de Lagrange on obtient le problème dual

$$\begin{cases} \min\limits_{\boldsymbol{\alpha}} \frac{1}{2}\sum\limits_{i,j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(\mathbf{x}_i,\mathbf{x}_j)+\varepsilon\sum\limits_{i=1}^{n}(\alpha_i+\alpha_i^*)-\sum\limits_{i=1}^{n}y_i(\alpha_i-\alpha_i^*) \\ C\geq\alpha_i,\alpha_i^*\geq 0, \quad 1\leq i\leq n \\ \sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0 \end{cases}$$

- Comme $\mathbf{w}=\sum\limits_{i=1}^{n}(\alpha_i-\alpha_i^*)\phi(\mathbf{x})$, la fonction recherchée sera

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

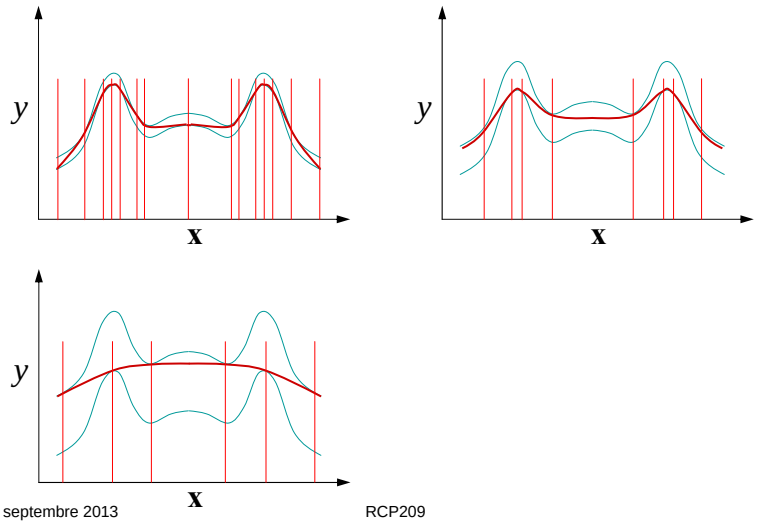
- Les conditions KKT impliquent $\alpha_i\,\alpha_i^*=0$ et

$$\varepsilon-y_i+\langle\mathbf{w},\phi(\mathbf{x}_i)\rangle+b\geq 0 \text{ et } \xi_i=0 \text{ si } \alpha_i<C$$
$$\varepsilon-y_i+\langle\mathbf{w},\phi(\mathbf{x}_i)\rangle+b\leq 0 \text{ si } \alpha_i>C$$

25 septembre 2013 RCP209 33

le cnam

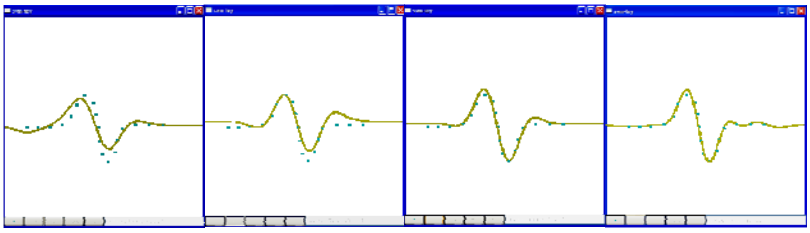
Régression : exemple 1



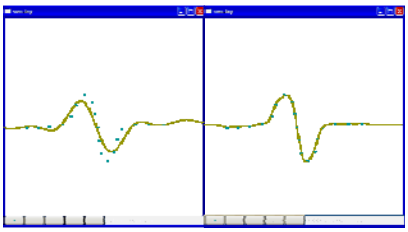
25 septembre 2013 RCP209 34

le cnam

Régression : exemple 2



Noyau RBF avec γ=100 : ν=0,1 et ε=0,1, ν=0,9 et ε=0,1, ν=0,1 et ε=0,01, ν=0,9 et ε=0,01



Noyau RBF avec γ=40 (gauche) et avec γ=1000 (ν=0,9 et ε=0,01 dans les deux cas)

le cnam

Méthodes à noyaux et analyse des données

le cnam

ACP à noyaux

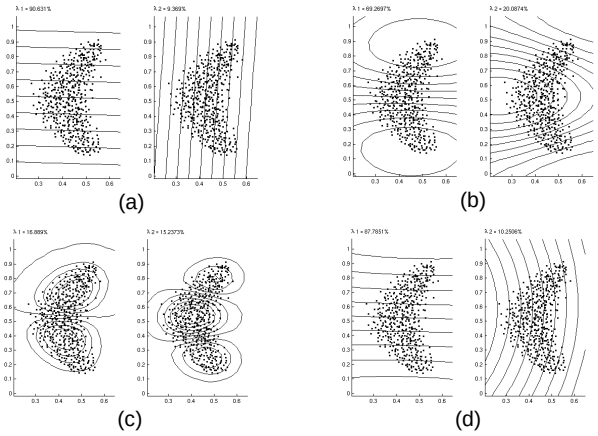
- L'analyse en composantes principales peut être effectuée dans l'espace d'arrivée, sur le nuage des points $\phi(\mathbf{x}_i)$
- Le problème de valeurs et vecteurs propres à résoudre est
  $\sum_{i=1}^{n} \langle \hat{\phi}(\mathbf{x}_i), \mathbf{u} \rangle \hat{\phi}(\mathbf{x}_i) = \lambda \mathbf{u}$
- En utilisant les vecteurs centrés $\hat{\phi}(\mathbf{x}_i) = \phi(\mathbf{x}_i) - \frac{1}{n}\sum_{j=1}^{n} \phi(\mathbf{x}_j)$
  et en définissant la matrice $\hat{\mathbf{K}}$ d'éléments $\langle \hat{\phi}(\mathbf{x}_i), \hat{\phi}(\mathbf{x}_j) \rangle$
  on obtient $\hat{\mathbf{K}}\,\boldsymbol{\beta} = \lambda\,\boldsymbol{\beta}$ où

$$\langle \hat{\phi}(\mathbf{x}_i), \hat{\phi}(\mathbf{x}_j) \rangle = K(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{n}\sum_{k=1}^{n} K(\mathbf{x}_i, \mathbf{x}_k) - \frac{1}{n}\sum_{k=1}^{n} K(\mathbf{x}_j, \mathbf{x}_k) + \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} K(\mathbf{x}_k, \mathbf{x}_l)$$

- La projection d'un nouveau point sur un vecteur propre sera
  $\langle \hat{\phi}(\mathbf{x}), \mathbf{u}_\alpha \rangle = \sum_{j=1}^{n} \beta_{\alpha j} K(\mathbf{x}, \mathbf{x}_j) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \beta_{\alpha i} K(\mathbf{x}_i, \mathbf{x}_j)$

le cnam

ACP à noyaux : exemple



(a) (b) (c) (d)

Courbes de niveau obtenues par ACP (a) linéaire, (b) à noyau RBF γ=5, (c) à noyau RBF γ=100 et (d) à noyau RBF γ=0,5 (figure de [CAB04])

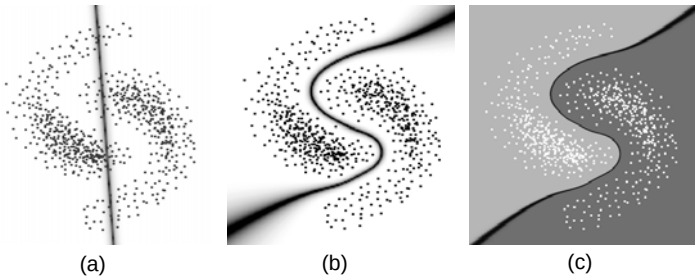
le cnam

AFD à noyaux

- L'analyse factorielle discriminante peut être effectuée dans l'espace d'arrivée, sur le nuage des points $\phi(\mathbf{x}_i)$
- Une version à 2 classes a été proposée dans [MRW99] et des versions à plusieurs classes dans [BA00], [RS00]
- Si les SVM se focalisent sur les vecteurs les plus proches de la frontière (la maximisation de marge mène à une solution creuse), l'analyse discriminante à noyaux tient compte de la répartition des « masses » de points et la solution est en général moins creuse

le cnam

AFD à noyaux : exemple 1

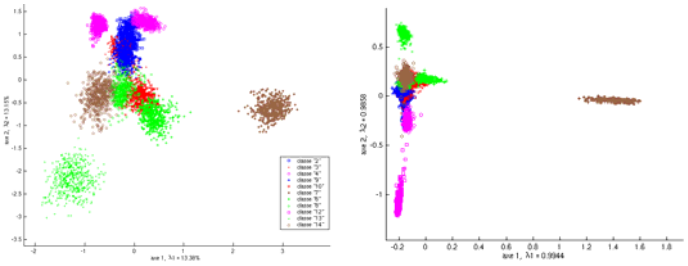


(a) (b) (c)

Discrimination obtenue par (a) AFD linéaire, (b) AFD à noyau angulaire et (c) machine à vecteurs support (SVM) à noyau angulaire (figure de [CAB04])

le cnam

AFD à noyaux : exemple 2



(a) (b)

Projections des individus des données « Textures » sur les deux premiers axes factoriels discriminants pour (a) AFD linéaire et (b) AFD à noyau angulaire (figure de [CAB04])

25 septembre 2013 RCP209 41

le cnam

Ingénierie des noyaux

25 septembre 2013 RCP209 42

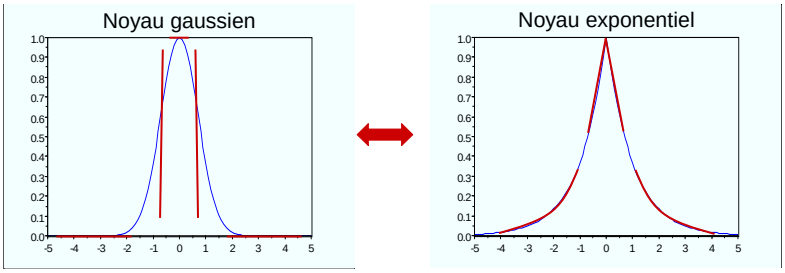
le cnam

Qu'est-ce qu'un noyau ?

- Noyau $\overset{?}{\sim}$ mesure de similarité
- Définition d'une mesure de similarité $s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:

$\mathbf{x}, \mathbf{y} \in \mathcal{X} \quad s(\mathbf{x}, \mathbf{y}) \geq 0 \quad s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$

$\forall \mathbf{y},\ s(\mathbf{x}, \mathbf{y}) \leq s(\mathbf{x}, \mathbf{x}) \quad s(\mathbf{x}, \mathbf{y}) = s(\mathbf{x}, \mathbf{x}) \Leftrightarrow \mathbf{x} = \mathbf{y}$



le cnam

Noyaux valides

- Soit $\mathcal{X}$ compact dans $\mathbb{R}^d$ et $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symétrique tels que

$\forall f \in L_2(\mathcal{X}), \quad \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) f(\mathbf{y}) d\mathbf{x}\, d\mathbf{y} \geq 0$ (condition de Mercer)

alors il existe un espace de Hilbert $\mathcal{H}$ et $\phi : \mathcal{X} \to \mathcal{H}$ tels que $K(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$

- Condition équivalente (noyau défini positif) : $\forall n \in \mathbb{N}^*$ et $\{\mathbf{x}_i\}_n \subset \mathcal{X}$, la matrice de Gram $\mathbf{K}_n : \{\mathbf{K}_n\}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j),\ 1 \leq i, j \leq n$, est définie positive, c'est à dire $\forall \mathbf{c}_n \in \mathbb{R}^n$ différent du vecteur nul de $\mathbb{R}^n$ on a $\mathbf{c}_n^T \mathbf{K}_n \mathbf{c}_n > 0$
- Un noyau valide garantit donc l'existence de $\mathcal{H}$ et peut s'exprimer comme un produit scalaire dans $\mathcal{H}$; aussi, il garantit la convexité du problème d'optimisation quadratique sous contraintes

le cnam

Noyaux valides (2)

- Noyau conditionnellement défini positif (cdp) : $\forall n \in \mathbb{N}^*$, $\{\mathbf{x}_i\}_n \subset \mathcal{X}$, la matrice de Gram $\mathbf{K}_n : \{\mathbf{K}_n\}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, $1 \le i, j \le n$, est conditionnellement définie positive, c'est à dire $\forall \mathbf{c}_n \in \mathbb{R}^n$ différent du vecteur nul de $\mathbb{R}^n$ et tel que $\sum_{i=1}^n c_i = 0$, on a $\mathbf{c}_n^T \mathbf{K}_n \mathbf{c}_n > 0$
- Étant donné un noyau symétrique K (un noyau cdp est symétrique), il existe un espace vectoriel $\mathcal{V}$, une transformation $\phi : \mathcal{X} \to \mathcal{V}$ et une forme bilinéaire $Q : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ tels que $K(\mathbf{x}, \mathbf{y}) = Q[\phi(\mathbf{x}), \phi(\mathbf{y})]$ (si K n'est pas défini positif, Q ne sera pas un produit scalaire) [Sch01]
- Un noyau cdp peut être utilisé pour les SVM en discrimination car les contraintes du problème d'optimisation quadratique incluent la condition $\sum_{i=1}^n \alpha_i y_i = 0$ $(c_i = \alpha_i y_i)$

le cnam

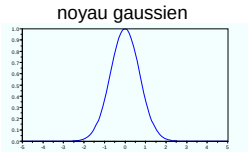
Construire des noyaux définis positifs

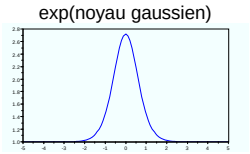
- Construction directe : définition de $\mathcal{H}$, $\phi : \mathcal{X} \to \mathcal{H}$ et ensuite du noyau $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ par $K(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$
- Soit $f : \mathcal{X} \to \mathbb{R}$, alors $K(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) \cdot f(\mathbf{y})$ ($K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\mathcal{X}$ compact dans $\mathbb{R}$) est défini positif (*conformal kernel*)
  - Exemples (ces noyaux ne peuvent pas être interprétés comme des mesures de similarité) :
    1. $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x$, $K(x,y) = x \cdot y$ (le noyau linéaire)
    2. $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, $K(x,y) = e^{x+y}$

le cnam

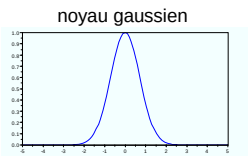
Transformer des noyaux

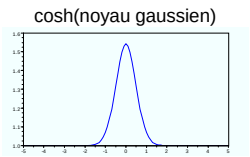
- Si $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est défini positif alors $\exp(K)$ est défini positif
  - Exemple :





- Si $K : \mathcal{X} \times \mathcal{X} \to [-1, 1]$ est défini positif alors $\cosh(K)$ est défini positif
  - Exemple :





25 septembre 2013 RCP209 47

le cnam

Combiner des noyaux définis positifs

- Si $K_1, K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ sont définis positifs et $\alpha_1, \alpha_2 \in \mathbb{R}^+$, alors sont également définis positifs les noyaux suivants
  - Combinaison linéaire : $K = \alpha_1 K_1 + \alpha_2 K_2$
  - Produit simple : $K = K_1 \cdot K_2$

    avec $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
- Si $K_1 : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathbb{R}$ et $K_2 : \mathcal{X}_2 \times \mathcal{X}_2 \to \mathbb{R}$ sont définis positifs, alors sont également définis positifs les noyaux
  - Somme directe : $K_1 \oplus K_2 = K_1 + K_2$
  - Produit tensoriel : $K_1 \otimes K_2 = K_1 \cdot K_2$

    avec $K_1 \oplus K_2, K_1 \otimes K_2 : (\mathcal{X}_1 \times \mathcal{X}_2) \times (\mathcal{X}_1 \times \mathcal{X}_2) \to \mathbb{R}$

  → construction de noyaux hybrides

25 septembre 2013 RCP209 48

le cnam

Qu'est-ce qu'un « bon » noyau ?

- Propriétés d'un « bon » noyau (suivant [Gär03]) :
  - Préalable : définition de « concepts » $c : \mathcal{X} \to \{0, 1\}$
  1. « Complétude » : dans quelle mesure la connaissance incorporée dans le noyau est suffisante pour résoudre le problème $\left(K(\mathbf{x},\cdot) = K(\mathbf{y},\cdot) \;\Rightarrow\; c(\mathbf{x}) = c(\mathbf{y})\right)$
  2. « Correctitude » : dans quelle mesure la sémantique du problème est respectée par le noyau (l'outil qui exploite le noyau, SVM dans notre cas, permet de trouver une solution pour chaque concept du problème)
  3. « Justesse » : dans quelle mesure le noyau reflète bien la mesure de similarité qui correspond aux « concepts » du problème (des bornes polynomiales de généralisation peuvent être obtenues pour l'outil qui exploite le noyau)

le cnam

Exemple : noyau pour ensembles

- Nature des données (exemple issu de [Bou05]) : ensembles de descripteurs locaux d'images
- Définition de l'espace $\mathcal{X}$: ensemble des parties finies mais de cardinalité variable de $\mathbb{R}^d$: $\mathcal{X} = \mathcal{P}_f\left(\mathbb{R}^d\right)$
- Objectif : évaluer la similarité entre ensembles de vecteurs $\mathcal{E}, \mathcal{E}' \in \mathcal{P}_f\left(\mathbb{R}^d\right)$ à travers les proximités entre vecteurs similaires de $\mathcal{E}, \mathcal{E}' \to$ on décrit ici le noyau d'appariement intermédiaire (*intermediate matching kernel*, [Bou05])
- Problème : le noyau d'appariement direct,

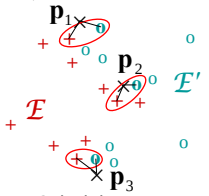
  $$K_{\mathcal{A}}(\mathcal{E}, \mathcal{E}') = (1/2)\left[\sum_{\mathbf{x}_i \in \mathcal{E}} \max_{\mathbf{x}'_j \in \mathcal{E}'} K(\mathbf{x}_i, \mathbf{x}'_j) + \sum_{\mathbf{x}'_j \in \mathcal{E}'} \max_{\mathbf{x}_i \in \mathcal{E}} K(\mathbf{x}_i, \mathbf{x}'_j)\right]$$

  où $K(\mathbf{x}_i, \mathbf{x}'_j)$ est un noyau classique entre les vecteurs $\mathbf{x}_i, \mathbf{x}'_j$, n'est pas défini positif !

le cnam

Exemple : noyau pour ensembles (2)

- Principe du noyau d'appariement intermédiaire : faire l'appariement par rapport à des vecteurs pivots, fixés pour un ensemble d'apprentissage donné (ne dépendant donc pas de $\mathcal{E}, \mathcal{E}'$)
- Soient *m* vecteurs pivots $\mathbf{p}_1, \ldots, \mathbf{p}_m \in \mathbb{R}^d$, pour chaque vecteur $\mathbf{p}_l$ on définit une fonction $\psi_l : \mathcal{X} \to \mathbb{R}^d$, $\psi_l(\mathcal{E}) = \mathbf{x}_l^* = \arg\min_{\mathbf{x} \in \mathcal{E}} \| \mathbf{x} - \mathbf{p}_l \|$
- Le noyau d'appariement intermédiaire construit à partir des pivots $\mathbf{p}_1, \ldots, \mathbf{p}_m$ est défini par $K_{\mathcal{M}}(\mathcal{E}, \mathcal{E}') = \sum_{l=1}^{m} K(\mathbf{x}_l^*, \mathbf{x}_l^{*'})$ et on peut facilement montrer qu'il est positif défini
- Choix possible des vecteurs pivots : prototypes des groupes de vecteurs obtenus par classification automatique des données d'apprentissage

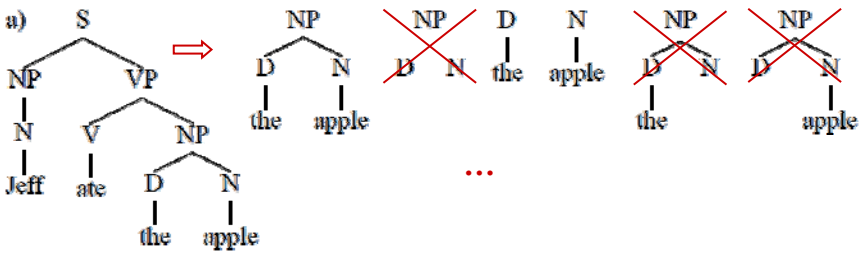


Calcul du noyau d'appariement intermédiaire

le cnam

Exemple : noyau pour arbres

- Nature des données (exemple issu de [CD02]) : arbres de *parsing* employés dans le traitement automatique des langues
- Définition de l'espace $\mathcal{X}$: ensemble des arbres étiquetés et ordonnés, avec des étiquettes dans un ensemble *E*
- Sous-arbre : sous-graphe connexe d'un arbre qui, pour un nœud de l'arbre, contient soit tous ses fils, soit aucun



le cnam

Exemple : noyau pour arbres (2)

- On considère une énumération de tous les sous-arbres possibles (nombre fini)
- On note par $h_l(T)$ le nombre d'occurrences du sous-arbre *l* dans l'arbre T
- L'arbre T sera représenté par $\phi(T) = [h_1(T)\, h_2(T) \ldots h_l(T) \ldots]$
- Pour deux arbres $T_1, T_2 \in \mathcal{X}$ on définit la valeur du noyau comme étant le produit scalaire entre ces deux représentations, $K(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$ ou, plus précisément, $K(T_1, T_2) = \sum_l h_l(T_1)\, h_l(T_2)$
- Soit $\mathcal{N}_1, \mathcal{N}_2$ les ensembles des nœuds des arbres T_1, T_2 et $S(v_1, v_2), v_1 \in \mathcal{N}_1, v_2 \in \mathcal{N}_2$ le nombre de sous-arbres isomorphes ayant pour racine v_1 et respectivement v_2
- Alors $K(T_1, T_2) = \sum_{v_1 \in \mathcal{N}_1, v_2 \in \mathcal{N}_2} S(v_1, v_2)$, calcul récursif $O(|\mathcal{N}_1| \cdot |\mathcal{N}_2|)$

le cnam

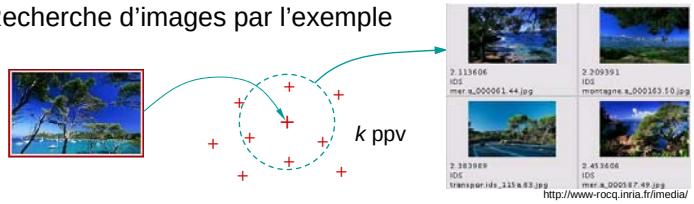
Exemples d'application

le cnam

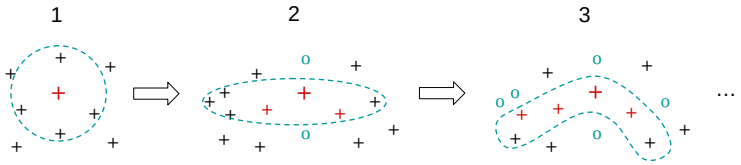
SVM pour le contrôle de pertinence

- Contrôle de pertinence (*relevance feedback*) : tenir compte du feedback de l'utilisateur dans la recherche itérative par le contenu

1. Recherche d'images par l'exemple



2. Recherche itérative avec contrôle de pertinence



25 septembre 2013 RCP209 55

le cnam

Composantes du mécanisme

1. *Learner* : à partir de l'information disponible (notamment des exemples positifs et/ou négatifs), estimer l'ensemble d'images visé
2. Sélecteur : à partir de l'estimation produite par le *learner*, choisir les images que l'utilisateur doit marquer lors de l'itération suivante
3. Utilisateur : fournir à chaque itération le retour pour les images choisies par le sélecteur
   - Les évaluations sont souvent faites à l'aide d'une vérité terrain, en émulant l'utilisateur

25 septembre 2013 RCP209 56

le cnam

Difficultés pour l'apprentissage

- Très peu d'exemples étiquetés : leur nombre est souvent inférieur au nombre de dimensions de l'espace de description !
- Déséquilibre important entre le nombre d'exemples positifs et le nombre d'exemples négatifs
- Forme potentiellement complexe de l'ensemble d'images visé, qui peut même présenter plusieurs modes distants dans l'espace de description
- L'interactivité exige un temps de réponse très court, à la fois pour le *learner* et pour le sélecteur

le cnam

Sélecteur : objectifs et critères

- Objectifs
  1. Retourner un maximum d'images pertinentes à l'utilisateur
  2. Maximiser le transfert d'information utilisateur → système
- Critères de sélection
  - « *Les plus positives* » (MP) : retourner les images les plus pertinentes suivant l'estimation actuelle faite par le *learner* – critère classique le plus utilisé
  - « *Les plus informatives* » (MI) : retourner les images qui permettent à l'utilisateur de fournir un maximum d'information sur sa cible → minimiser le nombre d'exemples

le cnam

Critère « les plus informatives »

- Composantes complémentaires du critère MI
  1. Ambiguïté élevée (de chaque image sélectionnée) par rapport à l'estimation courante faite par le *learner*
     - Comme critère individuel : « les plus ambiguës »
  2. Faible redondance de l'ensemble des *s* images retournées

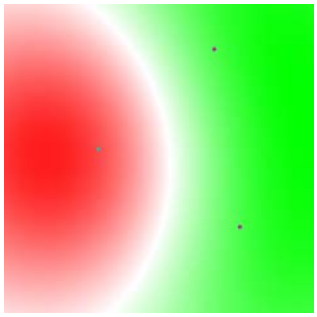
- Un critère « les plus informatives » pour SVM [FCB04]
  1. Présélectionner les *t* > *s* images pour lesquelles les valeurs de la fonction de décision SVM sont les plus proches de 0 (images les plus ambiguës)
  2. Choisir itérativement les *s* images pour lesquelles

$$\mathbf{x}_j = \arg\min_{\mathbf{x} \in \text{présélection}} \max_i K(\mathbf{x}, \mathbf{x}_j)$$

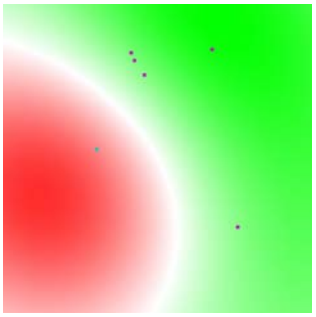
le cnam

Plus ambiguës : illustration

- Les *s* images les plus proches de la frontière peuvent être redondantes



Avant sélection

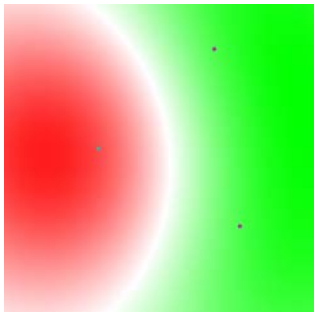


Après sélection, *feedback*, estimation

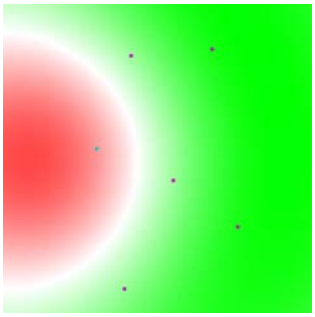
le cnam

Plus informatives : illustration

- Le critère conjoint minimise également la redondance
  ⇒ focalisation plus rapide sur les images recherchées



Avant sélection



Après sélection, *feedback*, estimation

le cnam

Contrôle de pertinence : exemple (3)

Objectif : retrouver des portraits

Base de 7500 images, dont 110 portraits

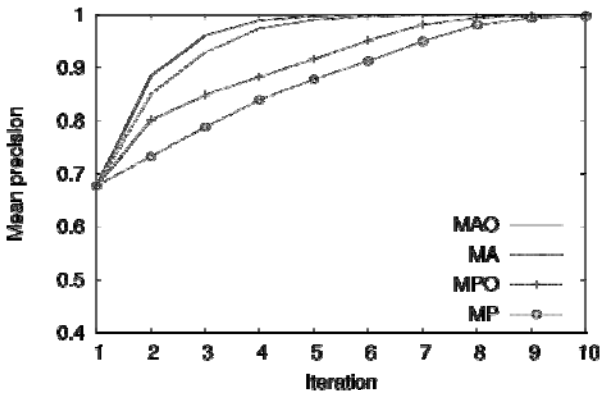
Disponible : description globale (couleur, texture, forme)

Première page de résultats après 4 itérations →

Voir [FCB04] et http://www-rocq.inria.fr/imedia

le cnam

Rapidité de convergence : *ranking*

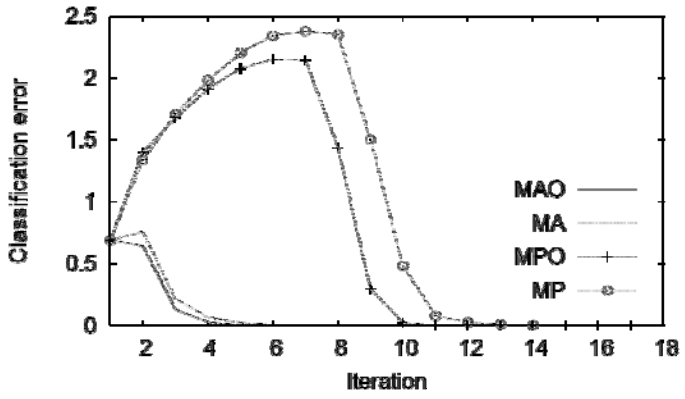


Évolution de la précision moyenne lors d'itérations successives

[Fer05], [FCB04]

le cnam

Rapidité de convergence : séparation

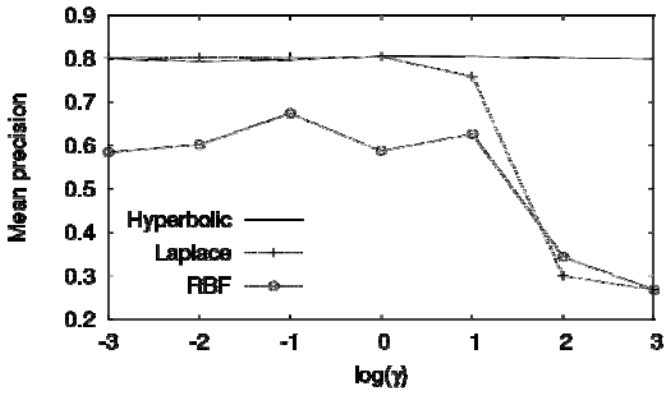


Évolution de l'erreur de discrimination lors d'itérations successives

[Fer05], [FCB04]

le cnam

Dépendance du paramètre d'échelle



Dépendance des résultats (précision moyenne durant les 15 premières itérations) du paramètre d'échelle du noyau

[Fer05], [FCB04]

le cnam

Application : conclusion

- Avantages des SVM pour le contrôle de pertinence
  - La fonction de décision associée permet à la fois la définition d'une frontière et le classement des images
  - Avec un large choix des noyaux, les SVM permettent une grande liberté dans la forme des classes (avec un contrôle par la régularisation)
  - D'autres sources d'information (en dehors des exemples) permettent de définir des noyaux appropriés
  - Apprentissage très rapide avec le nombre relativement limité d'exemples fournis par le contrôle de pertinence
  - Moindre sensibilité au déséquilibre entre exemples positifs et négatifs
- Inconvénients
  - Par rapport aux noyaux de Parzen, absence de caractère incrémental (dans la formulation de base) et donc étape de sélection plus coûteuse

le cnam

Références

[BA00] G. Baudat, Anouar F., Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12: 2385-2404, 2000.

[Bou05] S. Boughorbel, Noyaux pour la classification d'images par les Machines à Vecteurs de Support. *Thèse de doctorat*, Université d'Orsay, juillet 2005.

[Bur98] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2: 121-167, 1998.

[CAB04] M. Crucianu, J.-P. Asselin de Beauville, R. Boné, *Méthodes factorielles pour l'analyse des données : méthodes linéaires et extensions non-linéaires*. Hermès, 2004, 288 p.

[CD02] M. Collins, N. Duffy. Convolution kernels for natural language. In T.G. Dietterich, S. Becker and Z. Ghahramani (Eds.) *Advances in Neural Information Processing Systems* Vol. 15. MIT Press, 2002.

[Fer05] M. Ferecatu. Image retrieval with active relevance feedback using both visual and keyword-based descriptors, *Thèse de doctorat*, Université de Versailles Saint-Quentin-en-Yvelines, juillet 2005.

le cnam

Références

[FCB04] Ferecatu, M., Crucianu, M., Boujemaa, N. Retrieval of Difficult Image Classes Using SVM-Based Relevance Feedback, *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2004, New York, pp. 23-30.

[Gär03] Gärtner, T. A survey of kernels for structured data. *SIGKDD Explorer Newsletter* 5: 49-58, 2003.

[MRW99] S. Mika, Rätsch G., Weston J., Schölkopf B., Muller K.-R., Fischer discriminant analysis with kernels, *Proceedings of the IXth international workshop on Neural Networks for Signal Processing*, pp. 41-48, IEEE Press, 1999.

[MMR01] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf. An introduction to kernel-based learning methods. *IEEE Transactions on Neural Networks* 12: 181-202, 2001.

[RS00] V. Roth, Steinhage V., Nonlinear discriminant analysis using kernel functions, *Advances in Neural Information Processing Systems*, vol. 12, pp. 568-574, MIT Press, 2000.

le cnam

Références

[Sch01] Schölkopf, B. The kernel trick for distances. *Advances in Neural Information Processing Systems* 13, 301-307. (Eds.) Leen, T. K., T. G. Dietterich, V. Tresp, MIT Press, Cambridge, MA, USA, 2001.

[SPS99] B. Schölkopf, J. Platt, A. Smola, J. Shawe-Taylor, R. J. Williamson. Estimating the support of a high-dimensional distribution. *Technical Report* 87, Microsoft Research, Redmond, WA, 1999.

[SSM96] B. Schölkopf, Smola A., Müller K.-R., Nonlinear component analysis as a kernel eigenvalue problem, *Technical Report* 44, Max-Planck-Institut für biologische Kybernetik, Tübingen, Germany, 1996.

[SS02] B. Schölkopf, A. Smola. *Learning with Kernels*. MIT Press, 2002.

[SS04] A. Smola, B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing* 14: 199-222, 2004.

[TD99] D. Tax, R. Duin. Data domain description by support vectors. *Proceedings ESANN'99*, pp. 251-256. D-Facto Press, Brussels. 1999.

[Vap98] V. Vapnik. *Statistical Learning Theory*. John Wiley, New York. 1998.