Apprentissage, réseaux de neurones et modèles graphiques (RCP209)

Méthodes à noyaux et SVM non linéaires

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

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

23 mars 2017

Plan du cours

Objectif

"La raison d'être des statistiques, c'est de vous donner raison." — Abe Burrows

Méthodes à noyaux :

- Ingénierie des noyaux
  - Définitions
  - Noyaux valides, noyaux positif définis
  - Condition de Mercer
  - Transformer et combiner des noyaux
  - Noyaux structurés (noyaux pour ensembles)
- Le truc à noyaux (*the kernel trick*)
- SVM non linéaire

Plan du cours

Ingénierie des noyaux

Qu'est-ce qu'un noyau ? (intuition)

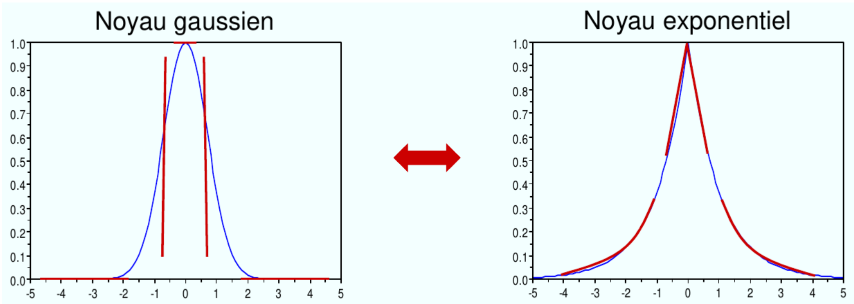
- Noyau $\approx$ mesure de similarité
- Définition d'une mesure de similarité :

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

$$\forall y \in \mathcal{X}, y \neq x \quad s(x, y) > s(x, x)$$

$$s(x, y) = s(x, x) \Leftrightarrow x = y$$

- A comparer avec la définition d'une distance



Exemples noyaux

Ingénierie des noyaux

Théorème de Mercer :
$\mathcal{X}$ compact dans R^d et $K : \mathcal{X} \times \mathcal{X} \rightarrow R$ symétrique
De plus, $\forall f \in L_2(\mathcal{X})$:

$$\int_{\mathcal{X}} K(x, y) f(x) f(y) dxdy \geq 0 \quad \text{(condition de Mercer)}$$

Alors : il existe un espace de Hilbert $\mathcal{H}$ et $\phi : \mathcal{X} \rightarrow \mathcal{H}$ tel que $\forall x, y \in \mathcal{X}$:

$$K(x, y) = \langle \phi(x), \phi(y) \rangle \text{ (produit scalaire)}$$

- $K(x, y)$ s'appelle **noyau positif défini**

Ingénierie des noyaux

Condition équivalente (**noyau positif défini**) :
$\forall n \in N$ et $\{x_i\}_{i=1,\dots,n} \subset \mathcal{X}$ la matrice de Gramm

$$K = [K_{i,j}]_{i=1,\dots,n} = [K(x_i, x_j)]_{i=1,\dots,n}$$

est définie positive, c.t.d :

$$\forall c \in R^n, c \neq 0, \text{ on a } c^T K c > 0$$

- Un noyau valide garantit donc l'existence de $\mathcal{H}$ et peut s'exprimer donc comme un produit scalaire dans $\mathcal{H}$
- Un noyau valide garantit aussi la convexité du problème d'optimisation quadratique sous contraintes des SVM

Ingénierie des noyaux

Un noyau est **conditionnellement défini positif** si
$\forall n \in N$ et $\{x_i\}_{i=1,\dots,n} \subset \mathcal{X}$ la matrice de Gramm

$$K = [K_{i,j}]_{i=1,\dots,n} = [K(x_i, x_j)]_{i=1,\dots,n}$$

est **conditionnellement** définie positive, c.t.d :

$$\forall c \in R^n, c \neq 0 \text{ tel que } \sum_{i=1}^{n} c_i = 0, \text{ on a } c^T K c > 0$$

Noyau conditionnellement défini positif (CDP)

Étant donné un **noyau symétrique conditionnellement défini positif**, il existe

- Un espace vectoriel $\mathcal{V}$;
- Une transformation $\phi : \mathcal{X} \to \mathcal{V}$
- Une forme bilinéaire $Q : \mathcal{V} \times \mathcal{V} \to R$

Tel que :

$$K(x, y) = Q(\phi(x), \phi(y))$$

- Si K n'est pas défini positif alors Q n'est pas un produit scalaire
- Un noyau CDP peut être utilisé pour les SVM en discrimination car les contraintes du problème d'optimisation quadratique incluent la condition $\sum_{i=1}^{n} \alpha_i y_i = 0$ $(c_i = \alpha_i y_i)$

Plan du cours

Construction des noyaux définis positif

- **Construction directe** : Définition de $\mathcal{H}, \phi : \mathcal{X} \to \mathcal{H}$ et ensuite construction du noyau

$$K : \mathcal{X} \times \mathcal{X} \to R \text{ par } K(x, y) = \langle \phi(x), \phi(y) \rangle \text{ (produit scalaire)}$$

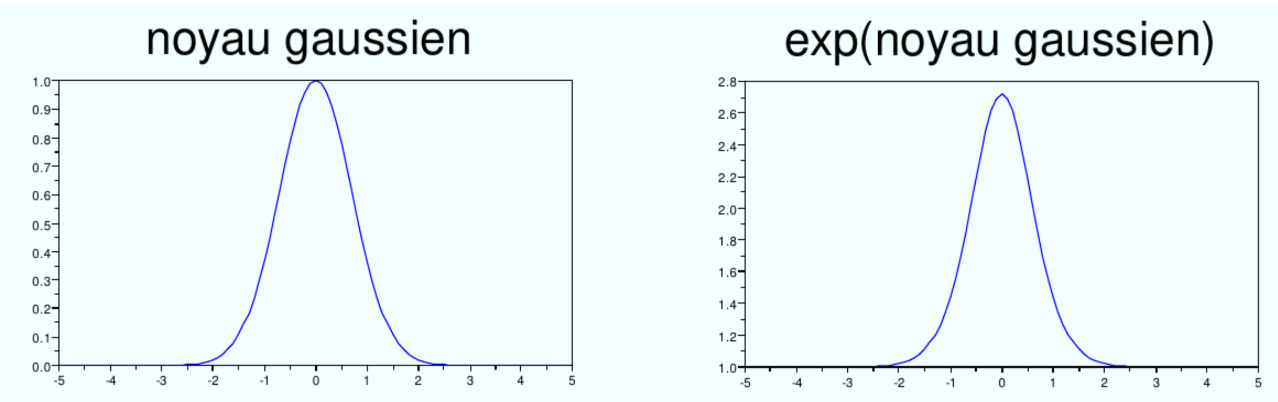
- Si $f : \mathcal{X} \to R$, alors $K(x, y) = f(x) \cdot f(y)$ ($K : \mathcal{X} \times \mathcal{X} \to R, \mathcal{X}$ compact dans R) est défini positif (**conformal kernel**).

Attention : ces noyaux conformes ne peuvent pas être interprétés comme des similarités.
Exemples :

- $f : R \to R, f(x) = x : K(x, y) = x \cdot y$ (le noyau linéaire)
- $f : R \to R, f(x) = e^x , K(x, y) = e^{x+y}$

Transformer des noyaux

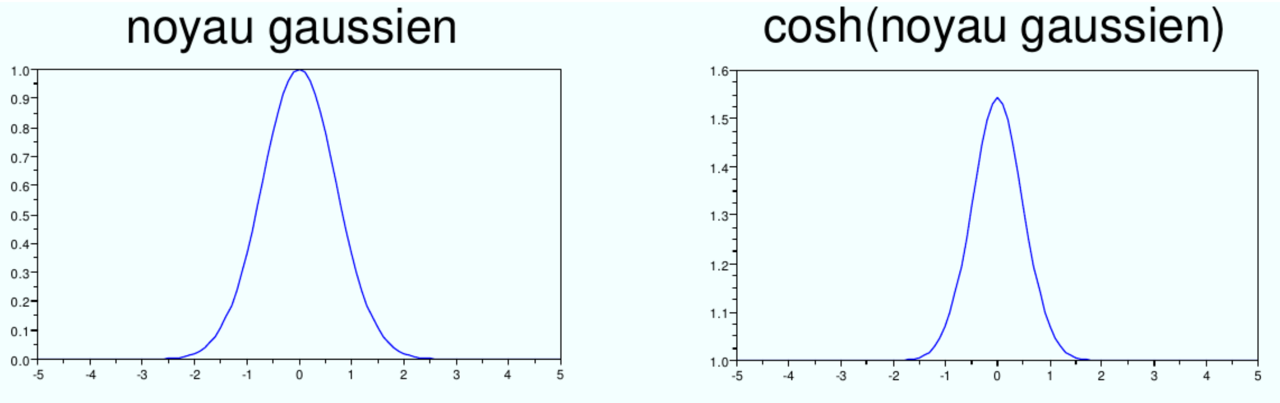
Si $K : \mathcal{X} \times \mathcal{X} \to R$ est défini positif alors $exp(K)$ est défini positif aussi



$exp(K)$

Transformer des noyaux

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



$$cosh(K) = \frac{exp(K)+exp(-K)}{2}$$

Combiner des noyaux définis positifs

Si $K_1\,, K_2 : \mathcal{X} \times \mathcal{X} \to R$ sont définis positifs et $\alpha_1\,, \alpha_2 > 0$ alors sont également défini positifs les noyaux suivants :

- Combinaison linéaire : $K(x, y) = \alpha_1 K_1(x, y) + \alpha_2 K_2(x, y)$
- Produit simple : $K(x, y) = \alpha_1 K_1(x, y) \cdot \alpha_2 K_2(x, y)$

ou $K : \mathcal{X} \times \mathcal{X} \to R$.

Si $K_1 : \mathcal{X}_1 \times \mathcal{X}_1 \to R$ et $K_2 : \mathcal{X}_2 \times \mathcal{X}_2 \to R$ sont définis positifs alors sont également défini positifs :

- Somme directe : $K_1 \oplus K_2 = K_1 + K_2$
- Produit tensoriel : $K_1 \otimes K_2 = K_1 \cdot K_2$

$\Longrightarrow$ Construction des noyaux hybrides

Noyaux d'appariement intermédiaire

Nature des données (exemple issu de [Boughorbel 2005]) : ensembles de descripteurs locaux d'images.

L'espace $\mathcal{X}$: ensemble des parties finies mais de cardinalité variable de R^d : $\mathcal{X} = \mathcal{P}_f(R^d)$

Objectif : évaluer la similarité entre ensembles de vecteurs $\mathcal{E}, \mathcal{E}' \in \mathcal{P}_f(R^d)$ à travers les proximités entre vecteurs similaires de $\mathcal{E}, \mathcal{E}'$ (*noyaux d'appariement intermédiaire*)

Problème : le noyau d'appariement direct :

$$K(\mathcal{E}, \mathcal{E}') = \frac{1}{2}\left[\sum_{x_i \in \mathcal{E}} \max_{x'_j \in \mathcal{E}'} K(x_i, x'_j) + \sum_{x'_j \in \mathcal{E}'} \max_{x_i \in \mathcal{E}} K(x'_j, x_i)\right]$$

ou $K(x_i, x'_j)$ est un noyau classique entre les vecteurs x_i et x'_j, n'est pas défini positif !

Noyaux d'appariement intermédiaire

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 $\mathcal{E}$ et $\mathcal{E}'$.

Soit m vecteurs pivot $p_1, \ldots, p_m \in R^d$. Pour chaque vecteur p_l on défini une fonction $\psi_l : \mathcal{X} \to R^d, \psi_l(\mathcal{E}) = x_l^* = \arg\min_{x \in \mathcal{E}} ||x - p_l||$

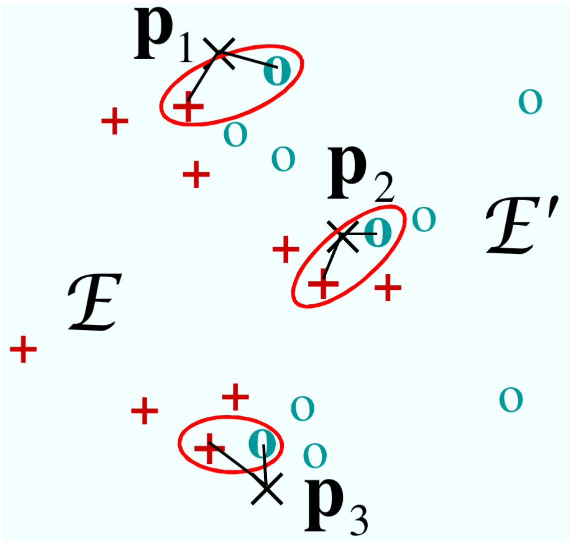
Le noyau d'appariement intermédiaire construit à partir des pivots $p_1, \ldots, p_m$ est défini par :

$$K_M(\mathcal{E}, \mathcal{E}') = \sum_{l=1}^{m} K(x_l, x'_l)$$

et on peut montrer qu'il est positif défini.

Noyaux d'appariement intermédiaire

Choix possible des vecteurs pivots : prototypes des groupes de vecteurs obtenus par classification automatique des données d'apprentissage.



Calcul du noyau d'appariement intermédiaire

Plan du cours

L'astuce à noyaux

SVM est un séparateur linéaire (avantages : pb. d'optimisation convexe, algorithmes efficaces)

Question : 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) et ensuite appliquer l'algorithme SVM sur les données transposées.

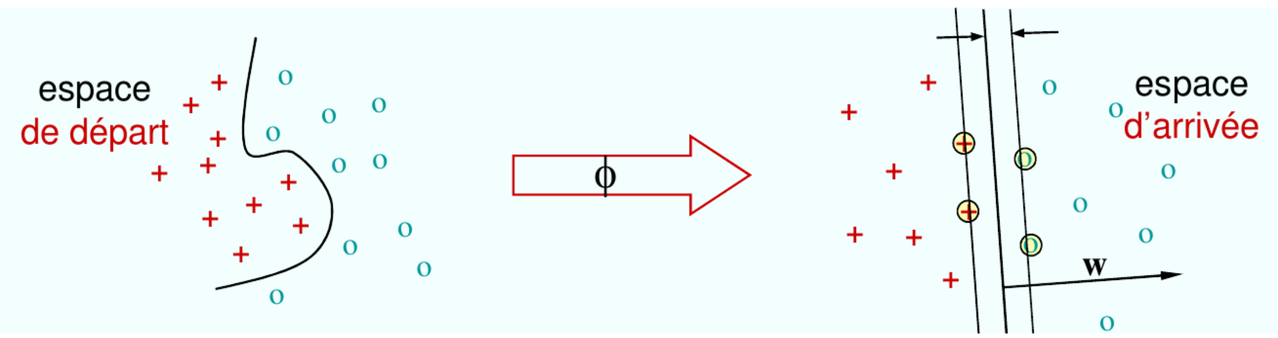
Transformation $\phi : R^d \to \mathcal{H}\,, x \to \phi(x)\,, \mathcal{H}$ espace de Hilbert.

L'astuce à noyaux

Chercher une transformation $\phi : R^d \to \mathcal{H}\,, x \to \phi(x)\,, \mathcal{H}$ espace de Hilbert.

Si K est un noyau défini positif ($K : R^d \times R^d \to R$), alors l'existence de ϕ et $\mathcal{H}$ est garantie (condition de Mercer) et :

$$K(x, y) = \langle \phi(x), \phi(y) \rangle$$

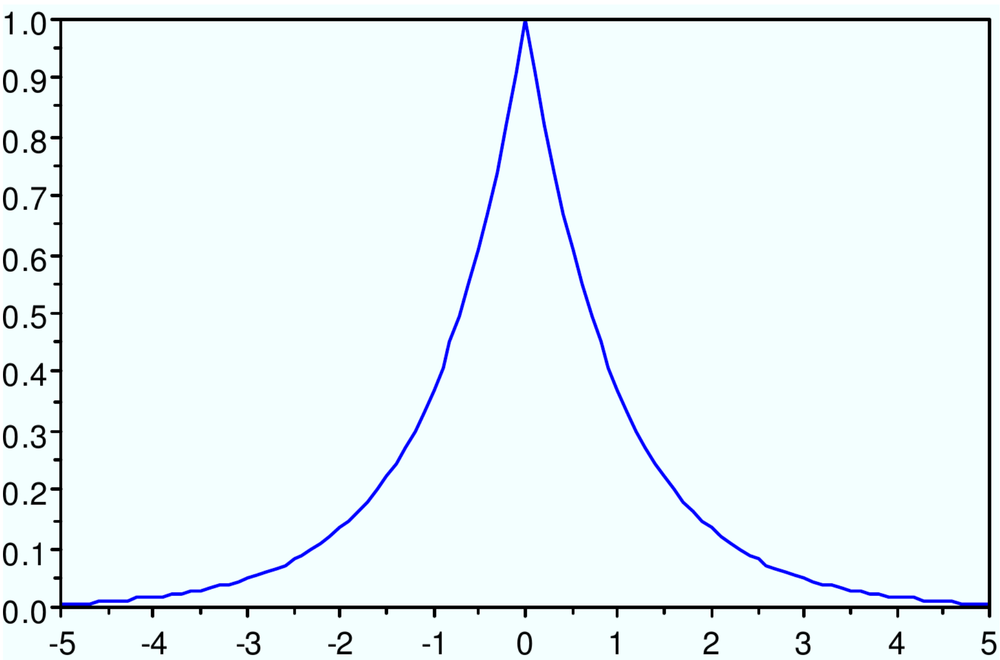


L'astuce à noyaux

- On n'a pas besoin de trouver la fonction ϕ
- Tout algorithme qui utilise seulement des produits scalaires entres les échantillons de données peuvent toute suite être appliqué dans l'espace $\mathcal{H}$, car le produit scalaire dans cet espace se calcule directement via le noyau $(K(x, y) = \langle \phi(x), \phi(y) \rangle)$, sans avoir besoin d'expliciter la projection.

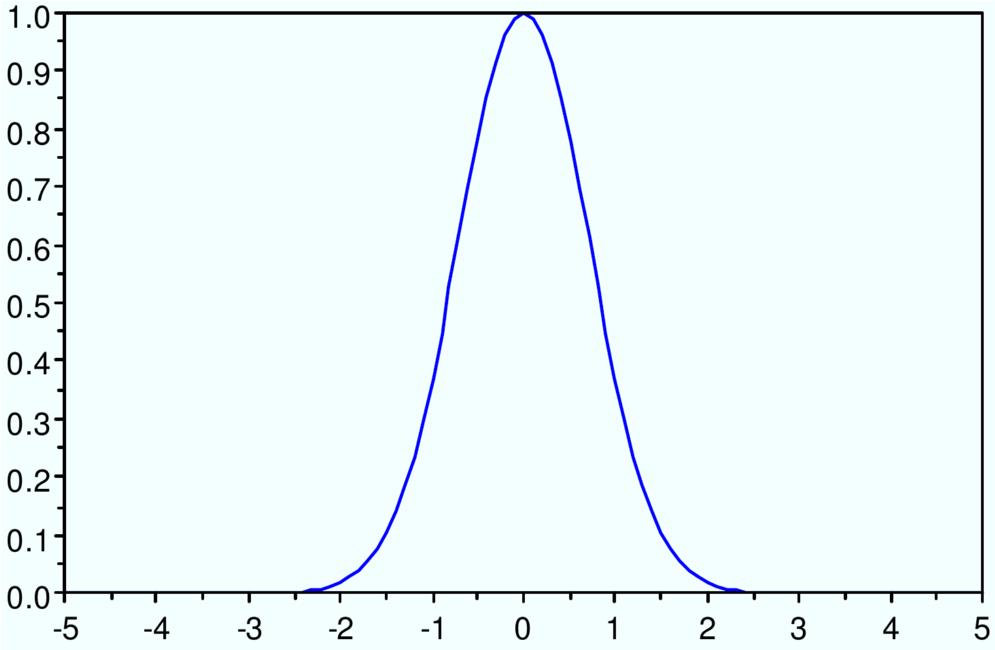
Exemples de noyaux

Noyaux linéaire : $K(x_i, x_j) = x_i^T x_j = \langle x_i, x_j \rangle$



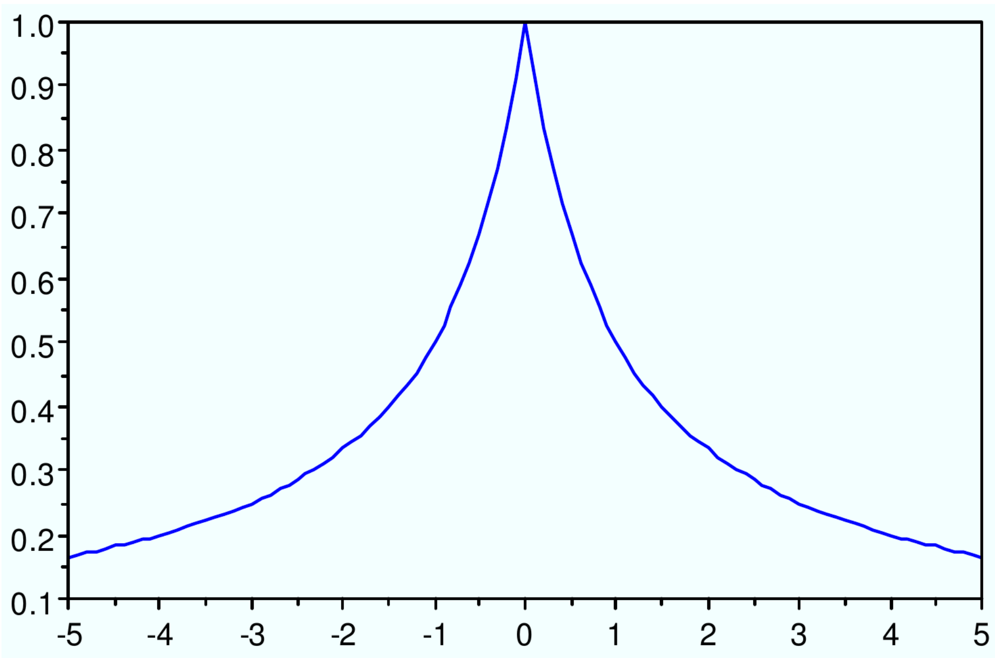
Noyau exponentiel : $K(x_i, x_j) = exp(-\gamma||x_i - x_j||)$

Exemples de noyaux



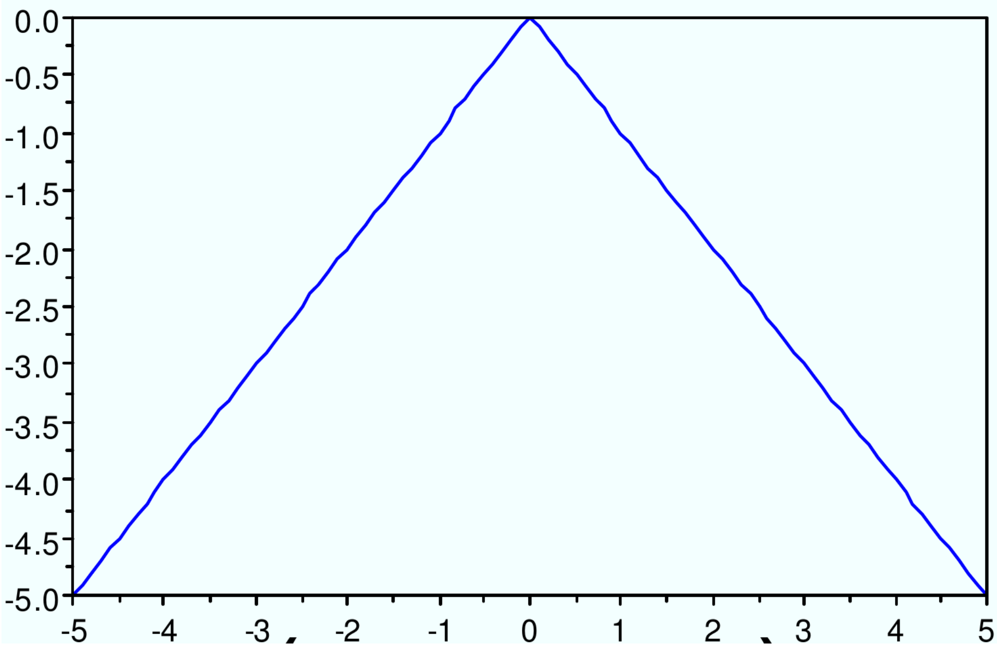
Noyau gaussien : $K(x_i\,, x_j) = exp(-\gamma||x_i - x_j||^2)$

Exemples de noyaux



Noyau hyperbolique : $K(x_i\,, x_j) = \frac{1}{\epsilon + \gamma||x_i - x_j||}$

Exemples de noyaux



Noyau angulaire : $K(x_i, x_j) = -||x_i - x_j||$

Noyau puissance : $K(x_i, x_j) = -||x_i - x_j||^\beta, 0 < \beta < 2$

Exemples : noyau polynomial

Noyau polynomial de degré 2 : $K(\mathbf{x_i}, \mathbf{x_j}) = (1 + \mathbf{x_i}^T \mathbf{x_j})^2$

$\mathcal{X} = R^2, \mathbf{x_i} = (x_i, y_i), \mathbf{x_j} = (x_j, y_j)$

En développant on obtient :

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

On obtient un produit scalaire en dimension 6 (et l'expression analytique de la fonction ϕ).

Plan du cours

SVM non-linéaire

SVM : problème d'optimisation dans le cas des données non-séparable :

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

- La séparation entre les classes étant non linéaire, un séparateur linéaire n'est pas adapté (produira beaucoup d'erreurs de classification)
- Les données x_i sont projetées dans l'espace $\mathcal{H}$ de très grande dimension, $x_i \to \phi(x_i)$. Dans $\mathcal{H}$ les projections $\phi(x_i)$ ont plus de chance d'être séparables linéairement.
- Le problème d'optimisation dans l'espace $\mathcal{H}$ est :

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i \\ t.q. \\ y_i(\langle w, \phi(x_i)\rangle + b) \geq 1 - \xi_i, i = 1, \dots, n \\ \xi_i \geq 0, i = 1, \dots, n \end{cases}$$

SVM non-linéaire

SVM : problème dual est donc :

$$\begin{cases} \max\limits_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle \\ t.q. \\ C \geq \alpha_i \geq 0, i = 1, \ldots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

Mais par l'astuce à noyau $\langle \phi(x_i)\,, \phi(x_j) \rangle = K(x_i\,, x_j)$ et le problème peut être résolu sans expliciter la projection ϕ (ce qui est souvent très difficile).

$$\begin{cases} \max\limits_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ t.q. \\ C \geq \alpha_i \geq 0, i = 1, \ldots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

SVM non-linéaire

La fonction de décision est donnée par :

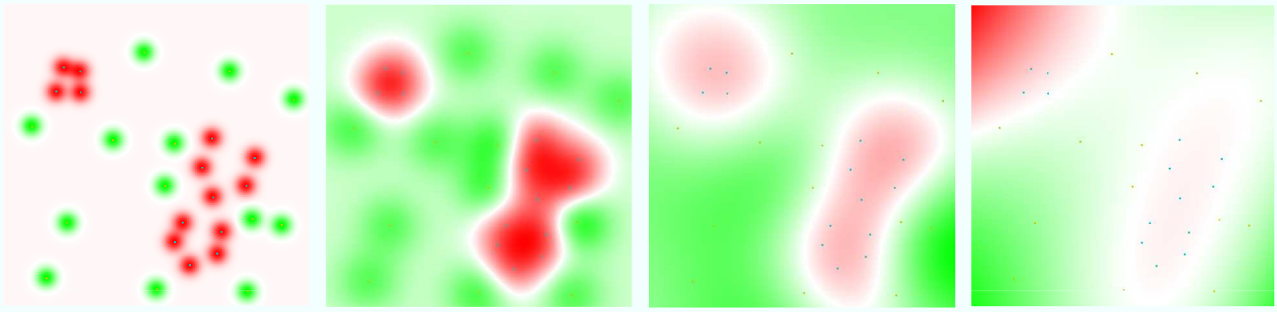
$$f^*(x) = \sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x) + b^*$$

- α^* sont issus du problème dual
- Pour une meilleure stabilité numérique, b^* est obtenu à partir de la moyenne sur l'ensemble I des vecteurs pour lesquels $0 < \alpha_i < C$:

$$b^* = \frac{1}{|I|} \sum_{x_j \in I} \left(y_j - \sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x_j) \right)$$

- Paramètres : C et les paramètres du noyau (souvent le paramètre d'échelle γ)

Effet de l'échelle le noyau Gaussien (RBF)



$1/\gamma$=0.001, 27 SV | $1/\gamma$=0.01, 26 SV | $1/\gamma$=0.1, 12 SV | $1/\gamma$=1.0, 10 SV

Sur-apprentissage : mauvaise généralisation, apprentissage du bruit

Sous-apprentissage : mauvaise généralisation, frontière imprécise

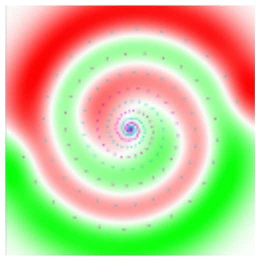
SVM à noyaux : le problème de l'échelle

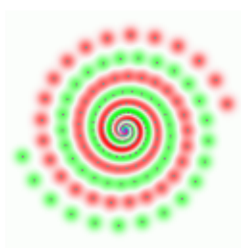
Noyaux classiques:

- paramètres d'échelle
- difficile a adapter en ligne

Noyau Gaussien

$$K(\mathrm{x},\mathrm{y}) = \exp(-\gamma\|\mathrm{x}-\mathrm{y}\|^2)$$





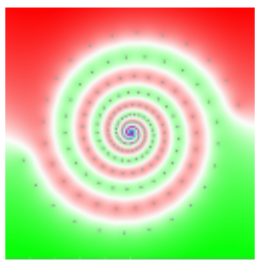
Noyaux non sensibles à l'échelle des données

Noyau triangulaire

$$K(\mathrm{x},\mathrm{y}) = -\|\mathrm{x}-\mathrm{y}\|$$

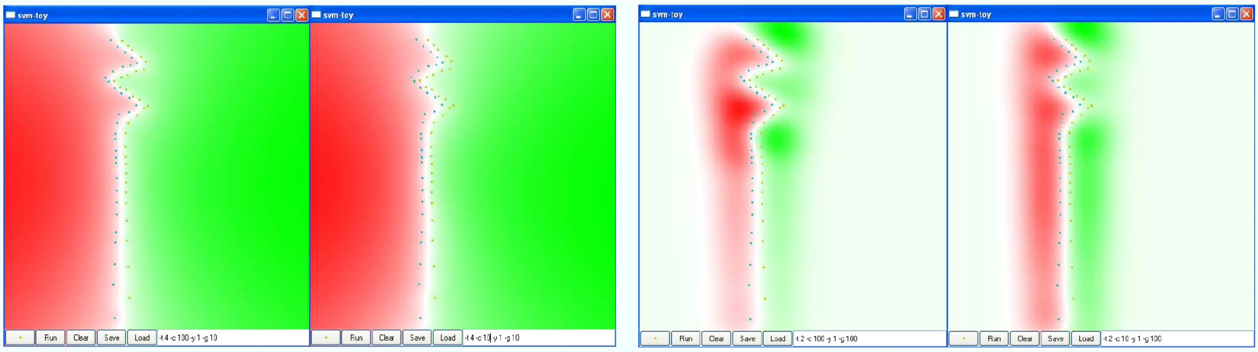
Noyau hyperbolique

$$K(\mathrm{x},\mathrm{y}) = \frac{1}{\varepsilon + \gamma\|\mathrm{x}-\mathrm{y}\|}$$

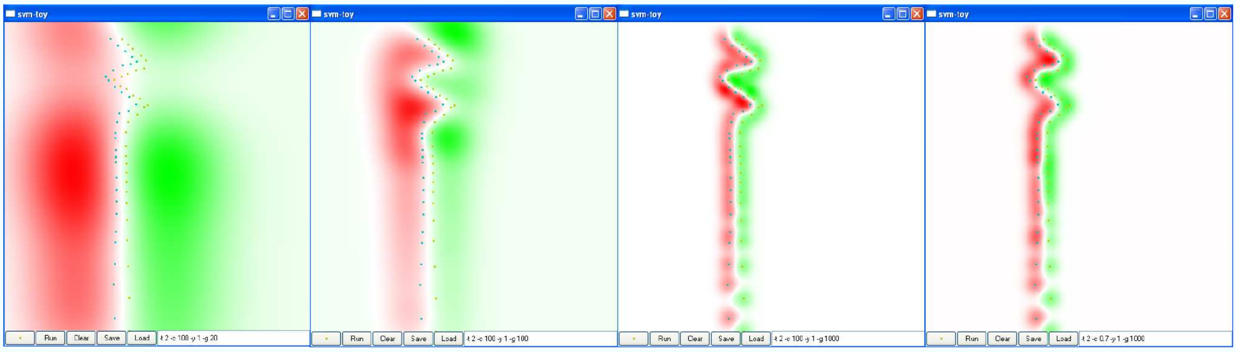


Tuning a kernel involves adapting its parameters to fit best the characteristics of the database. This is possible only for databases with a ground truth.

Effet de l'échelle et de C pour le noyau Gaussien (RBF) et angulaire



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



Noyau RBF avec comme valeurs pour $\gamma \setminus C$: 20 \ 100, puis 100 \ 100, puis 1000 \ 100, puis 1000 \ 0,7

SVM à noyaux

Implémentations software : Torch, LibSVM, LibLinear, Scikit-Learn

- Toch, http ://torch.ch/
- LibSVM, https ://www.csie.ntu.edu.tw/˜cjlin/libsvm/
- LibLinear, https ://www.csie.ntu.edu.tw/˜cjlin/liblinear/
- Scikit-Learn, http ://scikit-learn.org/

Pratiquement tous les grands environnement de modélisation mathématique possèdent implémentations performantes pour les SVM et méthodes à noyaux (R, Matlab, Mathematica, Scipy, Torch, Scikit-learn, etc.)

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
- Boughorbel et al., *Noyaux pour la classification d'images par les Machines à Vecteurs de Support*. Thèse de doctorat, Université d'Orsay, 2005.