SVM fonctionnels par interpolation spline

Nathalie Villa & Fabrice Rossi

Département de Mathématiques & Informatique (Équipe GRIMM), Université Toulouse Le Mirail, 5 allées A. Machado, 31058 Toulouse cedex 9, France
Projet AxIS, INRIA-Rocquencourt, Domaine de Voluceau,Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France

Résumé

Dans un nombre croissant d'applications, les données statistiques ne sont plus des vecteurs réels classiques mais le résultat de la discrétisation de fonctions sous-jacentes. L'étude de ces données fonctionnelles a conduit à la généralisation de nombreuses méthodes statistiques multi-dimensionnelles au cadre d'espaces de Hilbert de dimension infinie. Dans cet article, nous poursuivons l'étude de l'utilisation des SVM (machines à vecteurs de support) pour des données fonctionnelles que nous avions initiée dans Rossi & Villa (2006). Il s'agit ici de proposer une méthodologie pour utiliser les SVM sur des dérivées des fonctions d'origine, ce type de pré-traitement ayant montré son efficacité dans des problèmes réels comme en spectrométrie, par exemple. Nous proposons, à la fois une méthodologie permettant de calculer directement des noyaux sur les dérivées en utilisant les propriétés de l'interpolation spline dans les espaces de Sobolev, ainsi qu'un résultat de consistance universelle pour ce type de noyau.
Mots clés : analyse des données fonctionnelles, discrimination, support vector machine, apprentissage, splines

Abstract

In a growing number of applications, data are not classical real vectors but sampled functions. Functional data analysis (FDA) is a collection of traditional data analysis tools modified to handle correctly functional inputs taking their values in infinite dimensional spaces. In this article, we propose to apply Support Vector Machines (SVMs) to functional data analysis. We extend our earlier work Rossi & Villa (2006) by studying the case where SVMs are applied to the derivatives of the original data. This type of pre-processing had already shown its efficiency on real problems such as spectra classification. We develop here a methodology that allows to calculate directly kernels on derivatives by the use of interpolating splines properties in Sobolev spaces and also present an universal consistency result for this kind of kernels.
Key words : functional data analysis, discrimination, support vector machine, statistical learning theory, splines

1 Introduction

1.1 Analyse des données fonctionnelles

Le développement des appareils de mesure de grandeurs physiques a vu l'apparition, de manière de plus en plus fréquente, de données qui prennent la forme de fonctions discrétisées. De nombreux exemples de ce type de données se retrouvent dans des domaines d'applications divers comme la reconnaissance vocale, l'analyse des séries temporelles, la spectrométrie, etc.

Ces données sont de nature très particulière : la dimension de l'espace dans lequel elles prennent leurs valeurs est infinie (et, de manière pratique, le nombre de points de discrétisation est souvent très supérieur au nombre de points d'observations) et il existe une grande corrélation entre les différents points de discrétisation d'une même observation fonctionnelle. Ainsi, les outils de la statistique classique

conduisent, si ils sont appliqués directement aux données discrétisées, à des problèmes mal posés et ne permettent pas de construire des classifieurs ou des régresseurs pertinents sur ces données. Ces dernières années, de nombreux algorithmes ont été adaptés au traitement de ces données fonctionnelles et sont regroupés sous le nom générique d'« analyse des données fonctionnelles » (FDA). C'est le cas, par exemple, des analyses factorielles (e.g., Deville (1974), Dauxois & Pousse (1976), Besse & Ramsay (1986)), du modèle linéaire fonctionnel (Cardot *et al.* (1999)) ou de la régression PLS (Preda & Saporta (2002)). Une introduction complète aux méthodes linéaires fonctionnelles est disponible dans Ramsay & Silverman (1997). Enfin, plus récemment, des modèles fonctionnels non linéaires ont été développés, comme par exemple la régression inverse (modèle de réduction de la dimension semi-linéaire, généralisé par Ferré & Yao (2003)), les réseaux de neurones (Rossi & Conan-Guez (2005), Rossi *et al.* (2005) et Ferré & Villa (2006)), les modèles de régression non-paramétrique à noyaux (Ferraty & Vieu (2002)) ou encore les k-plus proches voisins (Biau *et al.* (2005)).

1.2 Analyse de données fonctionnelles par SVM

Dans Rossi & Villa (2006), nous étudions la discrimination de données fonctionnelles par machines à vecteurs de support (SVM). Nous proposons des noyaux qui sont spécialement conçus pour le traitement des données fonctionnelles et qui tirent partie de la nature particulière des données. De manière plus formelle, si $\mathcal{X}$ est un espace de Hilbert muni du produit scalaire $\langle ., . \rangle$, les noyaux proposés sont du type :

$$\forall\, u, v \in \mathcal{X}, \qquad Q(u, v) = K(\mathcal{P}(u), \mathcal{P}(v))$$

où $\mathcal{P}$ est un pré-traitement fonctionnel. De nombreux types de pré-traitements sont proposés et leurs capacités respectives sont illustrées par des exemples sur des données réelles. De plus, un résultat de consistance est démontré dans le cas où $\mathcal{P}$ est un opérateur de projection sur un espace engendré par une base hilbertienne tronquée de $L^2(\mu)$ (comme dans Biau *et al.* (2005)).

Dans le présent article, nous nous proposons d'étendre ce résultat de consistance à des noyaux différents. En effet, dans Rossi & Villa (2006), nous soulignons l'efficacité, sur certains types de données, d'un pré-traitement consistant à calculer des dérivées des fonctions d'origine. Ce type de pré-traitement se heurte à plusieurs difficultés : comme nous ne connaissons pas exactement la totalité de la fonction mais seulement sa valeur en certains points de discrétisation, nous devons reconstruire une représentation de la fonction observée avant de pouvoir calculer ses dérivées. La représentation des données par des splines est une méthode fréquemment utilisée lorsque l'on veut utiliser des dérivées des fonctions initiales ; malheureusement, les splines ne sont pas une base hilbertienne de $L^2(\mu)$ et le résultat de consistance décrit dans Rossi & Villa (2006) ne s'applique donc pas à cette méthodologie.

Nous proposons ici d'utiliser l'interpolation spline dans les espaces de Sobolev pour introduire des noyaux qui sont capables de mettre en œuvre, de manière naturelle, des SVM sur les dérivées en utilisant directement les données discrétisées d'origine. Cette méthodologie conduit à interpoler indirectement les données de la manière la plus régulière possible (au sens d'une pénalité définie par un opérateur linéaire). Dans la section 2, nous introduisons le lien entre espaces de Sobolev, splines d'interpolation et noyau reproduisant, puis, dans la section 3, nous présentons la méthode de SVM fonctionnelle sur dérivées et nous montrons que celle-ci est universellement consistante.

2 Interpolation spline et espaces de Hilbert à noyau reproduisant

2.1 Les données

Nous nous restreindrons ici à un problème de discrimination binaire pour lequel la variable explicative est une fonction de $\mathbb{R}$ dans $\mathbb{R}$. De manière plus précise, nous étudions un couple de variables

aléatoires (X, Y) où X est fonctionnelle et $Y \in \{-1, 1\}$, connu seulement grâce à n observations de (X, Y), $(x_1, y_1), \ldots, (x_n, y_n)$. On cherche à construire, à partir de ces observations, un classifieur capable de prédire Y sachant X. En fait, les x_i ne sont pas connues de façon exacte car nous disposons seulement, pour tout $i = 1, \ldots, n$, du vecteur $(x_i(t_1), \ldots, x_i(t_d))$ (les points de discrétisation sont les mêmes pour tous les x_i).

Par ailleurs, nous supposons que la variable aléatoire X est régulière, c'est-à-dire qu'elle prend ses valeurs dans l'espace de Sobolev $\mathcal{H}^m([0,1]) = \{h \in L^2([0,1]) \ : \ \forall\, j = 1, \ldots, m, D^j h$ existe (au sens faible) et $D^m h \in L^2([0,1])\}$. Ainsi, en tirant partie de la structure d'espace de Hilbert à noyau reproduisant (RKHS) de $\mathcal{H}^m$, X sera représentée par une interpolation spline.

2.2 Les L-splines

L'interpolation par L-spline d'une fonction discrétisée x consiste à la représenter par une fonction qui l'interpole exactement aux points de discrétisation et qui minimise une pénalité définie à partir d'un opérateur différentiel L.

Considérons plus précisément l'opérateur d'ordre m suivant

$$L = D^m + \sum_{j=0}^{m-1} a_j D^j,$$

pour lequel Ker$L$ est un sous-espace de dimension m de $\mathcal{H}^m$, noté $\mathcal{H}_0$. On montre (cf Besse & Ramsay (1986)) que l'espace $\mathcal{H}^m$ peut être décomposé en une somme directe de deux espaces :

$$\mathcal{H}^m = \mathcal{H}_0 \oplus \mathcal{H}_1,$$

où $\mathcal{H}_1$ est un espace de dimension infinie induit par m conditions aux bornes sur $\mathcal{H}^m$; la $j^{\text{ème}}$ condition est notée, pour tout $h \in \mathcal{H}_1$, $B^j h = 0$.

Supposons que $x \in \mathcal{H}_1$. On munit ce sous-espace de Hilbert du produit scalaire suivant : $\forall\, u, v \in \mathcal{H}_1$, $\langle u, v\rangle = \int Lu(t)Lv(t)dt$. L'interpolation L-spline de x est alors une fonction de $\mathcal{H}_1$ qui coïncide parfaitement avec x aux points de discrétisation et qui minimise la norme induite par le produit scalaire (et qui est donc la plus régulière de $\mathcal{H}_1$ au sens de L).

Or, $\mathcal{H}_1$ est un espace de Hilbert à noyau reproduisant (voir, par exemple, Heckman & Ramsay (2000)), ce qui permet de représenter simplement la fonction d'interpolation. On rappelle que dans un RKHS (cf Berlinet & Thomas-Agnan (2004) pour une présentation complète de la notion et de ses applications), il existe une fonction, K, appelée noyau, de $[0,1] \times [0,1]$ dans $\mathbb{R}$, telle que, pour tout $u \in \mathcal{H}_1$ et tout $t \in \mathbb{R}$, $\langle u, K(t, .)\rangle = u(t)$.

On a alors :

Théorème 1 (Besse & Ramsay (1986))**.** *Soit $x \in \mathcal{H}_1$ une fonction connue aux points de discrétisation $t_1, \ldots t_d$. Supposons, en outre, que la matrice $\mathbf{K} = (K(t_i, t_j))_{i,j}$ soit définie positive. Alors, il existe une unique fonction d'interpolation $h \in \mathcal{H}_1$ aux points $t_1, \ldots t_d$ telle que $\| h \| \leq \| u \|$ pour toute fonction d'interpolation $u \in \mathcal{H}_1$:*

$$h = \sum_{i=1}^{d} c_i K(t_i, .)$$

où $c = \mathbf{K}^{-1}\mathbf{x}$ avec $\mathbf{x} = (x(t_1), \ldots, x(t_d))$.

De plus, si h_1 et h_2 sont les deux fonctions d'interpolation de x_1, $x_2 \in \mathcal{H}_1$ comme définies ci-dessus, alors

$$\langle h_1, h_2\rangle = \mathbf{x}_1' \mathbf{K}^{-1} \mathbf{x}_2 = \langle \mathbf{x}_1, \mathbf{x}_2\rangle_{(\mathbb{R}^d, \mathbf{K}^{-1})} \tag{1}$$

où $(\mathbb{R}^d, \mathbf{K}^{-1})$ est l'espace $\mathbb{R}^d$ muni du produit scalaire induit par la matrice $\mathbf{K}^{-1}$.

Remarque 1. La fonction d'interpolation spline est donc simplement $h = \mathcal{P}_{\text{Vect}\{K(t_k,.),\ k=1,\ldots,d\}}(x)$.

2.3 Exemples

Pour illustrer la section précédente, nous donnons deux exemples de la décomposition $\mathcal{H}^m = \mathcal{H}_0 \oplus \mathcal{H}_1$.

Exemple 1. Pour $m = 1$ et $L = I + D$, on a $\mathcal{H}_0 = \text{Vect}\{t \to e^{-t}\}$. Un choix possible pour $\mathcal{H}_1$ est $\{h \in \mathcal{H}^1 \ : \ h(0) = 0\}$.

Exemple 2. Pour $m = 2$, $L = I + D^2$, on a $\mathcal{H}_0 = \text{Vect}\{\cos, \sin\}$. Un choix possible pour $\mathcal{H}_1$ est $\{h \in \mathcal{H}^2 \ : \ h(0) = Dh(0) = 0\}$.

On trouvera d'autres exemples dans Besse & Ramsay (1986) (avec des illustrations de l'importance des conditions aux bornes) ou Heckman & Ramsay (2000) (pour des opérateurs différentiels à coefficients non constants). On renvoie aux mêmes articles pour une description de la méthode permettant de déterminer, à partir de L et des conditions aux bornes, le noyau K, par le biais de la *fonction de Green*, $G : [0,1] \times [0,1] \to \mathbb{R}$, vérifiant

$$\forall\, h \in \mathcal{H}_1, \qquad \forall\, s \in [0,1], \qquad h(s) = \int G(s,t) L h(t) dt.$$

Les propriétés de la fonction de Green (cf Roach (1982) ou Stakgold (1979)) permettent de la déterminer entièrement à partir de L et des conditions aux bornes. Dans le cas de l'exemple 1, cela donne :

- $\forall\, t \in [0,1]$, $G(.,t) \in \mathcal{H}_0$ donc $G(s,t) = \begin{cases} a_1(t)e^{-s} & \text{si } t \leq s \\ b_1(t)e^{-s} & \text{si } t > s \end{cases}$;
- $\forall\, t \in [0,1]$, $G(.,t) \in \mathcal{H}_1$ donc $\forall\, t \geq 0$, $G(0,t) = 0 \Leftrightarrow b_1(t) = 0$;
- $\forall\, s \in [0,1]$ et $\forall\, j = 1 \ldots m-2$, $D_1^j G(s,s^+) = D_1^j G(s,s^-)$ et $D_1^{m-1} G(s,s^+) - D_1^{m-1} G(s,s^-) = 1$, ce qui implique, dans notre exemple, $a_1 : t \to e^t$;
- enfin, $K(s,t) = \langle G(s,.), G(t,.) \rangle = \int_0^{\inf(s,t)} e^{u-s} e^{u-t} du = e^{-\sup(s,t)} \sinh(\inf(s,t))$.

3 SVM fonctionnels sur dérivées

3.1 Méthodologie

Les résultats du Théorème 1 nous permettent de mettre en œuvre simplement des SVM sur les dérivées des observations en utilisant un noyau construit sur les discrétisations. En effet, supposons que $X \in \mathcal{H}_1$ et notons, pour tout $i = 1, \ldots, n$, h_i l'interpolation spline de x_i aux points de discrétisation $t_1, \ldots, t_d$. Alors, si on suppose que la matrice $\mathbf{K}$ est inversible, on obtient directement le résultat suivant :

Théorème 2. *Soit G_γ^d le noyau gaussien de paramètre γ sur $\mathbb{R}^d$ et G_γ^∞ le noyau gaussien de paramètre γ sur $L^2([0,1])$ ($G_\gamma(u,v) = e^{-\gamma \|u-v\|_{\mathbb{R}^d \text{ ou } L^2}}$). Alors, le SVM sur les dérivées des fonctions $h_1, \ldots, h_n$ (noté $\phi_h^{n,d}$) défini par*

$$\begin{array}{ll} \max_\alpha \sum_{i=1}^n \alpha_i - \sum_{i,j=1}^n \alpha_i \alpha_j G_\gamma^\infty(Lh_i, Lh_j) \\ \textit{avec} \quad \sum_{i,j=1}^n \alpha_i y_i = 0, \\ \qquad 0 \leq \alpha_i \leq C, \ 1 \leq i \leq n, \end{array}$$

est équivalent au SVM sur les discrétisations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (noté $\phi_{\mathbf{x}}^{n,d}$) :

$$\begin{array}{ll} \max_\alpha \sum_{i=1}^n \alpha_i - \sum_{i,j=1}^n \alpha_i \alpha_j G_\gamma^d \circ \mathbf{K}^{-1/2}(\mathbf{x}_i, \mathbf{x}_j) \\ \textit{avec} \quad \sum_{i,j=1}^n \alpha_i y_i = 0, \\ \qquad 0 \leq \alpha_i \leq C, \ 1 \leq i \leq n, \end{array}$$

Remarque 2. Ce SVM est équivalent à la construction d'un SVM fonctionnel ayant pour noyau $\mathcal{Q}(u,v) = e^{-\gamma \|L(\mathcal{P}_d u) - L(\mathcal{P}_d u)\|_{L^2}^2}$, soit si on se réfère à la terminologie employée dans Rossi & Villa (2006), à utiliser le

pré-traitement fonctionnel $\mathcal{P} : h \in \mathcal{H}_1 \subset L^2_{[0;1]} \to L(\mathcal{P}_d h)$ sans avoir à le calculer explicitement. Selon les choix de L, on peut donc construire des SVM sur dérivées en utilisant, directement sur les discrétisations, un noyau gaussien classique « perturbé » par la matrice $\mathbf{K}^{-1/2}$.

3.2 Consistance

Par le biais de l'utilisation de tels noyaux, on peut démontrer la consistance des SVM définis sur les dérivées des fonctions initiales. Pour introduire ce résultat, qui se présente sous la forme d'une double limite (limite quand le nombre d'observations tend vers $+\infty$ et le nombre de points de discrétisation tend vers $+\infty$), on démontre tout d'abord qu'étant donnés des points de discrétisation $t_1, \ldots, t_d$, on peut trouver une suite de points de discrétisation $(\tau_D)_{D\geq 1}$ telle que $\tau_1 = (t_1, \ldots, t_d)$ et qui assure la consistance de la méthode :

Proposition 1. *Soit $(t_k)_{k=1,\ldots,d}$ les points de discrétisation des fonctions observées. Quitte à retirer des points, on peut toujours supposer que $(K(t_k, .))_{k=1,\ldots,d}$ sont linéairement indépendants. Alors, il existe un ensemble dénombrable $\mathcal{D}_0 = (t_k)_{k\geq 1} \subset [0, 1]$ tel que*
- *$Vect\{K(t, .),\ t \in \mathcal{D}_0\}$ est dense dans $\mathcal{H}_1$:*
- *pour tout $D \geq 1$, la matrice $(K(t_i, t_j))_{i,j=1,\ldots,D}$ est inversible.*

On note alors : $\tau_1 = \{t_1, \ldots, t_d\}$ et $\forall\, D \geq 1$, $\tau_{D+1} = \tau_D \cup \{t_{d+D}\}$.

Ceci nous amène au résultat de consistance suivant :

Théorème 3. *Le SVM défini comme dans le Théorème 2, $\phi_h^{n,D}$ pour les points d'interpolation $\mathcal{T}_D$ et la suite de régularisation $(C_n^D)_n = \mathcal{O}(n^{1-\beta_D})$ avec $0 < \beta_D < 1/D$, est universellement consistant :*

$$\lim_{n\to+\infty} \lim_{D\to+\infty} L\phi_h^{n,D} = L^*$$

où $L\phi = P(\phi(X) \neq Y)$ et $L^* = \inf_{\phi:\mathcal{H}_1\to\{-1;1\}} P(\phi(X) \neq Y) = P(\phi^*(X) \neq Y)$ avec $\phi^*(x) = \begin{cases} 1 & si\ P(Y = 1|X = x) > 1/2 \\ -1 & sinon \end{cases}$.

La preuve de ce résultat est basée sur une démonstration en deux temps : tout d'abord, on montre que la régression de Y sur X est approchée de manière arbitrairement précise par la régression de Y sur $\mathcal{P}_D(X)$ lorsque D tend vers $+\infty$. Ensuite, on montre, pour D fixé, en utilisant le résultat de consistance des SVM multi-dimensionnels démontré par Steinwart (2002), que l'erreur de Bayes commise par le SVM $\phi_{\mathbf{x}}^{n,D}$ tend vers l'erreur de Bayes du couple $(\mathcal{P}_D(X), Y)$ lorsque le nombre d'observations tend vers $+\infty$. La combinaison de ces deux résultats permet de conclure.

4 Conclusion et ouvertures

Nous avons introduit une méthode d'utilisation des SVM pour des données de type fonctionnel avec un pré-traitement qui prend la forme d'un opérateur différentiel. L'avantage de notre approche est qu'elle permet d'effectuer ce pré-traitement de manière transparente en utilisant une simple perturbation du noyau d'origine sur les discrétisations des fonctions. Un résultat de consistance en découle obtenu à partir des résultats de consistance existant en dimension finie. Ce résultat pourrait être étendu, moyennant quelques adaptations, par une approche de splines de lissage qui autoriserait une interpolation imparfaite des données et donc la prise en compte d'un éventuel bruit sur les mesures ; cette nouvelle approche entraînerait néanmoins la détermination d'un paramètre de régularisation (le paramètre de lissage des splines) supplémentaire.

Références

Berlinet, A. & Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publisher.

Besse, P. & Ramsay, J. (1986). Principal component analysis of sampled curves. *Psychometrica*, **51**, 285–311.

Biau, G., Bunea, F. & Wegkamp, M. (2005). Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory*, **51**, 2163–2172.

Cardot, H., Ferraty, F. & Sarda, P. (1999). Functional Linear Model. *Statistics and Probability Letter*, **45**, 11–22.

Dauxois, J. & Pousse, A. (1976). Les analyses factorielles en calcul des probabilités et en statistique : essai d'étude synthétique. Thèse, Université Toulouse III.

Deville, J. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE*, **15**(Janvier–Avril), 3–97.

Ferraty, F. & Vieu, P. (2002). The functional nonparametric model and application to spectrometric data. *Computational Statistics*, **17**, 515–561.

Ferré, L. & Villa, N. (2006). Multi-layer neural network with functional inputs. *Scandinavian Journal of Statistics*. A paraître.

Ferré, L. & Yao, A. (2003). Functional sliced inverse regression analysis. *Statistics*, **37**, 475–488.

Heckman, N. & Ramsay, J. (2000). Penalized regression with model-based penalties. *The Canadian Journal of Statistics*, **28**, 241–258.

Preda, C. & Saporta, G. (2002). Régression PLS sur un processus stochastique. *Revue de statistique appliquée*, **L**(2).

Ramsay, J. & Silverman, B. (1997). *Functional Data Analysis*. Springer Verlag, New York.

Roach, G. (1982). *Green's Functions*. Cambridge University Press, Cambridge.

Rossi, F. & Conan-Guez, B. (2005). Functional multi-layer perceptron : a nonlinear tool for functional data anlysis. *Neural Networks*, **18**(1), 45–60.

Rossi, F., Delannay, N., Conan-Guez, B. & Verleysen, M. (2005). Representation of functional data in neural networks. *Neurocomputing*, **64**, 183–210.

Rossi, F. & Villa, N. (2006). Support vector machine for functional data classification. *Neurocomputing*. A paraître.

Stakgold, I. (1979). *Green's Functions and Boundary Value Problems*. Wiley, New York.

Steinwart, I. (2002). Support vector machines are universally consistent. *J. Complexity*, **18**, 768–791.