Une approche de type k-plus proches voisins pour la régression fonctionnelle

Thomas Laloë

UNE APPROCHE DE TYPE k-PLUS PROCHES VOISINS POUR LA REGRESSION FONCTIONNELLE

Thomas LALOË

Institut de Mathématiques et de Modélisation de Montpellier
UMR CNRS 5149, Equipe de Probabilités et Statistique
Université Montpellier II, Cc 051
Place Eugène Bataillon, 34095 Montpellier Cedex 5, France
tlaloe@math.univ-montp2.fr

Résumé

Soit (X, Y) un couple aléatoire à valeurs dans $\mathcal{H} \times \mathbb{R}$, où $\mathcal{H}$ est un espace de Hilbert de dimension infinie. Nous établissons la convergence faible d'un estimateur de type plus proches voisins de la fonction de régression de Y sur X, construit à partir d'observations indépendantes du couple (X, Y). Comme méthode générale, nous proposons de réduire la dimension de $\mathcal{H}$ en ne considérant que les d premiers coefficients de la projection de X dans une base orhonormale de $\mathcal{H}$, puis d'appliquer une régression de type plus proches voisins dans $\mathbb{R}^d$. La dimension et le nombre de voisins sont automatiquement choisis à partir des observations, en utilisant un outil classique de "data-splitting".

Mots clés : Plus proches voisins, Espaces de Hilbert de dimension infinie, Régression.

Abstract

Let (X, Y) be a random pair taking values in $\mathcal{H} \times \mathbb{R}$, where $\mathcal{H}$ is an infinite dimensional separable Hilbert space. We establish weak consistency of a nearest neighbor-type estimator of the regression function of Y on X based on independent observations of the pair (X, Y). As a general strategy, we propose to reduce the infinite dimension of $\mathcal{H}$ by considering only the first d coefficients of an expansion of X in an orthonormal system of $\mathcal{H}$, and then to perform k-nearest neighbor regression in $\mathbb{R}^d$. Both the dimension and the number of neighbors are automatically selected from the observations using a simple data-dependent splitting device.

Key-words : Nearest Neighbor, Infinite-dimensional Hilbert space, Régression.

1 Introduction

La régression consiste à prédire le comportement d'une variable à partir d'un certain nombre d'observations. Une observation est habituellement donnée par un nombre fini de mesures numériques, représentées par un vecteur de dimension d. Cependant, dans nombre de problèmes pratiques, les données prennent la forme de fonctions aléatoires (enregistrement de voix, images), et cela place la régression dans le problème plus général de l'analyse de données fonctionnelles. Même si en pratique de telles données sont observées en un nombre fini de points, le défi est d'inférer la structure des données en utilisant la nature fonctionnelle des données. Les dernières années ont vu se multiplier les résultats, tant pratiques que théoriques, en analyse des données fonctionnelles, et nombre d'outils classiques d'analyse des données ont été adaptés pour gérer des données fonctionnelles. Le livre de Ramsay et Silverman [6] offre une introduction compréhensible et complète au sujet. Nous recommandons également les travaux de Cerou et Guyader [3], Rossi et Villa [7] et Tuleau [8].

Dans cette présentation, nous considérons le cadre de la régression fonctionnelle, dans lequel le but est de prédire une réponse scalaire Y à partir d'observations de dimension infinie X. Plus précisément, nous noterons (X, Y) un couple aléatoire à valeurs dans $\mathcal{Z} = \mathcal{H} \times \mathbb{R}$, où $\mathcal{H}$ est un espace de Hilbert séparable de dimension infinie. Nous noterons également ρ la distribution (inconnue) de (X, Y), et ρ_X la distribution marginale de X. A partir de n copies indépendantes $(X_1, Y_1), \ldots, (X_n, Y_n)$ de (X, Y), nous construisons un estimateur f_n de la fonction de régression $f_\rho(x) = \mathbb{E}[Y|X = x]$ de la façon suivante : tout d'abord nous réduisons la dimension de $\mathcal{H}$ en ne considérant que les d premiers coefficients de la projection de chaque observation dans un système orthonormal de $\mathcal{H}$. Ensuite nous appliquons une régression de type k-plus proches voisins (Györfi, Kohler, Krzyzak et Walk [5]) dans $\mathbb{R}^d$. Nous choisissons simultanément la dimension d et le nombre de voisins k en utilisant une méthode de type "data-splitting". Notre principal résultat établit la convergence faible de l'estimateur obtenu, généralisant ainsi la stratégie introduite par Biau, Bunea et Wegkamp [2] dans le contexte de la classification (c'est-à-dire lorsque Y est à valeur dans un ensemble fini).

2 Notations

On note respectivement $\langle .|. \rangle$ et $\|.\|$ le produit scalaire et la norme associée sur $\mathcal{H}$, et on se donne $(\phi_j)_{j \geq 1}$ une base orthonormale de $\mathcal{H}$ (Akhiezer et Glazman [1]). Pour chaque observation X_i, on définit $X_{ij} = \langle X_i | \phi_j \rangle$. On sait que

$$X_i = \sum_{j=1}^{\infty} X_{ij} \phi_j.$$

Soit $\mathcal{H}^{(d)}$, l'espace vectoriel de dimension finie engendré par les fonctions $\{\phi_1, \phi_2, \ldots, \phi_d\}$, et soit, pour tout X_i,

$$X_i^{(d)} = \sum_{j=1}^{d} X_{ij}\phi_j.$$

Finalement, on note respectivement f_ρ et $f_{\rho,d}$ les fonctions de régression dans $\mathcal{H}$ et $\mathcal{H}^{(d)}$, et σ_ρ^2 et $\sigma_{\rho,d}^2$ leurs erreurs L^2 respectives. Plus précisément, on a $f_\rho(x) = \mathbb{E}[Y|X = x]$, $\sigma_\rho^2 = \int_{\mathcal{Z}}(y - f_\rho(x))^2 d\rho(x, y)$, et pareillement dans $\mathcal{H}^{(d)} \times \mathbb{R}$ pour $f_{\rho,d}$ et $\sigma_{\rho,d}^2$. Tout au long de la présentation, on supposera que $\mathbb{E}(Y^2) < \infty$ *p.s.*, et que toutes les integrales sont calculées par rapport à ρ ou ρ_X.

3 k-plus proches voisins dans $\mathcal{H}^{(d)}$

Commençons par définir notre estimateur de type k-plus proches voisins. Pour cela, on considère la suite $(X_1^{(d)}, Y_1), \ldots, (X_n^{(d)}, Y_n)$ où les observations ont été projetées sur $\mathcal{H}^{(d)}$. Pour chaque x dans $\mathcal{H}^{(d)}$, on ordonne les observations

$$\left(X_{(1)}^{(d)}(x), Y_{(1)}(x)\right), \ldots, \left(X_{(n)}^{(d)}(x), Y_{(n)}(x)\right),$$

selon les distances euclidiennes croissantes $\|X_i^{(d)} - x\|$ entre $X_i^{(d)}$ et x. Autrement dit, $X_{(i)}^{(d)}(x)$ est le i-ème plus proche voisin de x parmi les $X_j^{(d)}, j = 1, \ldots, n$. Si $\|X_i^{(d)} - x\| = \|X_j^{(d)} - x\|$, $X_i^{(d)}$ est le plus proche de x si $i < j$. L'estimateur de type k-plus proches voisins de f_ρ est alors défini (Györfi, Kohler, Krzyzak et Walk [5]) par

$$f_{n,k,d}(x) = \frac{1}{k}\sum_{i=1}^{k} Y_{(i)}(x). \tag{1}$$

Pour choisir simultanément la dimension d et le nombre de voisins k, nous suggérons la méthode de type "data-splitting" suivante . Premièrement on coupe les données en un ensemble d'apprentissage $\{(X_i, Y_i), i \in \mathcal{I}_\ell\}$ de taille ℓ, et un ensemble de validation $\{(X_j, Y_j), j \in \mathcal{J}_m\}$ de taille m, avec $m + \ell = n$ (ℓ et m peuvent être des fonctions de n). Pour tout $d \geq 1$, $1 \leq k \leq \ell$, on construit un estimateur de type plus proches voisins à partir de l'ensemble d'apprentissage. Ensuite on utilise l'ensemble de validation pour choisir $\hat{d}$ et $\hat{k}$ de la façon suivante :

$$(\hat{d}, \hat{k}) \in \underset{d\geq 1, 1\leq k\leq \ell}{\arg\min} \left[\frac{1}{m}\sum_{j\in\mathcal{J}_m}\left(Y_j - f_{\ell,k,d}(X_j^{(d)})\right)^2 + \frac{\lambda_d}{\sqrt{m}}\right]. \tag{2}$$

Le terme $\lambda_d/\sqrt{m}$ est un terme de penalité donné qui tend vers l'infini avec d pour éviter de trop grandes valeurs pour d.

Cette méthode, qui est très facile à programmer, conduit à l'estimateur

$$\hat{f}_n(x) := f_{\ell,\hat{k},\hat{d}}(x^{(\hat{d})}), \tag{3}$$

qui a une erreur

$$\mathcal{E}(\hat{f}_n) = \int_{\mathcal{Z}} \left(y - \hat{f}_n(x)\right)^2 d\rho(x,y) = \int_{\mathcal{H}} \left(\hat{f}_n(x) - f_\rho(x)\right)^2 d\rho_X(x) \ + \sigma_\rho^2 \, .$$

Finalement, sous des hypothèses similaires à celles de Biau, Bunea et Wegkamp [2], on montre que $\hat{f}_n$ converge faiblement vers f_ρ, i.e.,

$$\mathbb{E} \int_{\mathcal{H}} \left(\hat{f}_n(x) - f_\rho(x)\right)^2 \to 0 \ \textit{quand} \ n \to \infty.$$

Nous finirons l'exposé en présentant quelques résultats pratiques obtenus avec cet estimateur sur des données réelles et simulées.

Références

[1] N.I. AKHIEZER et I.M. GLAZMAN (1961). *Theory of linear operators in Hilbert space*, Frederick Ungar Publishing Co., New York.

[2] G.BIAU, F. BUNEA, et M.H. WEGKAMP (2005). Functional classification in Hilbert Spaces, *IEEE Transactions on Information Theory*, Vol. 51, pp. 2163-2172.

[3] F. CEROU et A. GUYADER (2005). Nearest neighbor classification in infinite dimension, *Research Report INRIA,* RR 5536.

[4] F. CUCKER et S. SMALE (2001). On the mathematical foundations of learning, *bulletin (New Series) of the american mathematical society*, Vol. 39, pp. 1-49.

[5] L. GYÖRFI, M. KOHLER, A. KRZYZAK, et H. WALK (2002). *A distribution free theory of nonparametric regression*, Springer Verlag, New York.

[6] J.O. RAMSAY et B.W. SILVERMAN (2002). *Functional data analysis*, Springer, New-York.

[7] F. ROSSI et N. VILLA (2006). Support Vector Machine For Functional Data Classification, *Neurocomputing* Vol 69, pp. 730-742.

[8] C. TULEAU (2005). Sélection de variables pour la discrimination en grande dimension, classification de données fonctionnelles, *PhD thesis, University Paris XI.*