Kernel Classification Rule for Functional Random Fields

Ahmad Younso

Damascus University, Mathematical Statistics Department, Damascus, Syria
E-mail: ahyounso@yahoo.fr

Résumé. Nous considérons la règle de la fenêtre mobile pour classifier des données fonctionnelles spatialement dépendantes. Nous étudions les propriétés asymptotiques de cette règle de classification non paramétrique basée sur des données d'apprentissage tirées d'un champ aléatoire α ou $\beta-$mélangeant à valeurs en espace de dimension infinie. Nous étendons les résultats d'Abraham et al. (2006) et Younso (2017) concernant la consistance et la consistance forte au cas spatialement dépendant sous des hypothèses légères.

Mots-clés. Règle de Bayes; Données d'apprentissage; Règle de la fenêtre mobile; champ aléatoire mélangeant; Facteur de lissage; consistance.

Abstract. We consider the moving window rule of classification for functional spatially dependent data. We investigate asymptotic properties of this nonparametric classification rule based on training data drawn from α or $\beta-$mixing random field taking values in infinite space. We extend the results of Abraham et al. (2006) and Younso (2017) concerning both the consistency and the strong consistency to a spatially dependent case under mild assumptions.

Keywords. Bayes rule; training data; moving window rule; mixing random field, smoothing factor; consistency.

1 Introduction

In many studies, the observations can be collected as spatially dependent curves. This type of data arises in a variety of fields including econometrics, epidemiology, environmental sciences, image analysis, oceanography and many others. In some studies, it can be interesting to see spatio-temporal data as spatially dependent data. The need to classify spatial functional data occurs in many scientific problems. For example, in medical imaging modalities, an important problem is how to classify image pixels into spatial regions in which the pixels exhibit similar temporal behavior. In this research, we propose a nonparametric classification rule for functional random fields satisfying mixing conditions defined later.

2 Moving window rule for functional random fields

Let $(\mathcal{E}, d)$ be a metric space where $\mathcal{E}$ is a function space and d is the metric on $\mathcal{E}$. Denote the integer lattice points in the N-dimensional Euclidean space by $\mathbb{Z}^N$, $N \geq 1$. Consider a stationary random field $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}_{\mathbf{i}\in\mathbb{Z}^N}$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $\mathcal{E} \times \{0,1\}$. A point $\mathbf{i} = (i_1, ..., i_N) \in \mathbb{Z}^N$ will be referred to as a site. For $\mathbf{n} = (n_1, ..., n_N) \in (\mathbb{N}^*)^N$, we define the rectangular region $\mathcal{I}_{\mathbf{n}}$ by

$$\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} \in \mathbb{Z}^N : \ 1 \leq i_k \leq n_k, \ \forall k = 1, ..., N\}.$$

We will write $\mathbf{n} \to \infty$ if

$$\min_{k=1,\dots,N} n_k \to \infty.$$

Define $\hat{\mathbf{n}} = n_1 \times ... \times n_N = \mathrm{Card}(\mathcal{I}_{\mathbf{n}})$. For the sake of simplicity, we suppose that $(X_{\mathbf{i}}, Y_{\mathbf{i}})$ has the same distribution as (X, Y) for all $\mathbf{i} \in \mathbb{Z}^N$. Classical procedure of classification deals with predicting the unknown nature Y called a class (0 or 1) of an observation X with values in $\mathcal{E}$. The statistician creates a classifier $g : \mathcal{E} \longrightarrow \{0,1\}$ which maps a new observation $x \in \mathcal{E}$ into its predicted label $g(x)$. Since it is not assumed that the covariate X fully determines the label, the same value x may give rise to different labels, it is certainly possible to wrongly specify its associated label. An error occurs if $g(X) \neq Y$, and the probability of error for a classifier g is $L(g) = \mathbb{P}\{g(X) \neq Y\}$. There exists a classifier, called Bayes rule, given by

$$g^*(x) = \begin{cases} 0 \text{ if } \mathbb{P}\{Y = 0|X = x\} \geq \mathbb{P}\{Y = 1|X = x\} \\ 1 \text{ otherwise.} \end{cases}$$

It is easily to see that the Bayes rule has the smallest probability of error, that is

$$L^* = L(g^*) = \inf_{g:\mathbb{R}^d \to \{0,1\}} \mathbb{P}\{g(X) \neq Y\},$$

see Theorem 2.1 in Devroye et al. (1996) for the finite dimensional case. Unfortunately, the Bayes rule depends on the distribution of (X, Y) which is generally unknown to the statistician. But it is often possible to observe data according to the distribution of the random pair (X, Y) and to construct a classifier from these data. To be more precise, let $D_{\mathbf{n}} = \{(X_{\mathbf{i}}, Y_{\mathbf{i}}), \ \mathbf{i} \in \mathcal{I}_{\mathbf{n}}\}$ be a set of observations of the random field defined above. The set $D_{\mathbf{n}}$ is called training data. In the classification problem, one want to use the data $D_{\mathbf{n}}$ in order to construct a classifier $g_{\mathbf{n}}(x)$ being a measurable function with respect to x and to the data $D_{\mathbf{n}}$. Among the various ways to define such classifier from a training data, one of the most simple and popular is the moving window rule given by

$$g_{\mathbf{n}}(x) = \begin{cases} 0 \text{ if } \displaystyle\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \mathbb{I}_{\{Y_{\mathbf{i}}=0, X_{\mathbf{i}}\in B_{x,b}\}} \geq \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \mathbb{I}_{\{Y_{\mathbf{i}}=1, X_{\mathbf{i}}\in B_{x,b}\}} \\ \\ 1 \text{ otherwise,} \end{cases} \tag{1}$$

where $\mathbb{I}_A$ denotes the indicator function of the set A, $b = b(\mathbf{n})$ the smoothing factor, is a strictly positive number tending to 0 when $\mathbf{n} \to \infty$ and $B_{x,b}$ denotes the closed ball of radius b centered at x. Clearly, the moving window rule is a kernel rule with a particular choice of the kernel. The best we can expect from $g_{\mathbf{n}}(x)$ is to achieve the Bayes error probability, and we rather require that the error probability $L_{\mathbf{n}} = L(g_{\mathbf{n}}) = \mathbb{P}\{g_{\mathbf{n}}(X) \neq Y\}$ get arbitrarily close to L^* with large probability. The classifier $g_{\mathbf{n}}(x)$ is said to be consistent if $L_{\mathbf{n}} = \mathbb{P}\{g_{\mathbf{n}}(X) \neq Y|D_{\mathbf{n}}\}$ acheives the Bayes risk L^* in the limit, as $\mathbf{n} \to \infty$. More precisely, the classifier $g_{\mathbf{n}}(x)$ is called consistent if

$$\mathbb{E}L_{\mathbf{n}} \longrightarrow L^* \text{ as } \mathbf{n} \to \infty$$

and said to be strongly consistent if

$$L_{\mathbf{n}} \longrightarrow L^* \text{ as } \mathbf{n} \to \infty \text{ with probability one .}$$

A classifier can be consistent for certain class of distribution of (X, Y), but not be consistent for others. The classifier $g_{\mathbf{n}}(x)$ is called universally (strongly) consistent, if it is (strongly) consistent for all distribution of (X, Y). In the independent functional case, Abraham et al. (2006) give examples showing that the moving window rule is not universally consistent. They give sufficient conditions to extend the consistency result of Theorem 10.1 in Devroye et al. (1996) to a function space. In this research, we will investigate the consistency and the strong consistency of the moving window rule (1).

3 Hypotheses and results

We suppose that the random field $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}_{\mathbf{i}\in\mathbb{Z}^N}$ is said to be α-mixing or strongly mixing if

$$\alpha(t) = \sup_{A\in\mathcal{B}(\mathcal{S}),C\in\mathcal{B}(\mathcal{S}'):\text{dist}(\mathcal{S},\mathcal{S}')\geq t} \{|\mathbb{P}(A\cap C) - \mathbb{P}(A)\mathbb{P}(C)|\} \downarrow 0 \text{ as } t \to \infty, \tag{2}$$

where $\mathcal{B}(\mathcal{S})$ and $\mathcal{B}(\mathcal{S}')$ denote the Borel σ-fields generated by $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}_{\mathbf{i}\in\mathcal{S}}$ and $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}_{\mathbf{i}\in\mathcal{S}'}$ respectively for any $\mathcal{S}, \mathcal{S}' \subset \mathbb{Z}^N$,

$$\text{dist}(\mathcal{S}, \mathcal{S}') = \inf_{\mathbf{i}\in\mathcal{S},\mathbf{j}\in\mathcal{S}'} \|\mathbf{i} - \mathbf{j}\|$$

and $\|.\|$ denotes the Euclidean norm. The random field $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}_{\mathbf{i}\in\mathbb{Z}^N}$ is said to be β-mixing or absolutely regular if

$$\beta(t) = \sup_{A\in\mathcal{B}(\mathcal{S}):\text{dist}(\mathcal{S},\mathcal{S}')\geq t} \mathbb{E}\{|\mathbb{P}(A|\mathcal{B}(\mathcal{S}')) - \mathbb{P}(A)|\} \downarrow 0 \text{ as } t \to \infty.$$

Since $2\alpha(t) \leq \beta(t)$, any $\beta-$mixing random field is $\alpha-$mixing. Before we formulate the main results, we introduce some important notations concerning covering number and metric entropy, see Kolmogorov and Tihomirov (1961). The idea of a covering number is to count how many small spherical balls would be needed to completely cover a given set. For a given subset $\mathcal{G}$ of the metric space $(\mathcal{E}, d)$ and a given $\epsilon > 0$, the ϵ-covering number denoted by $\mathcal{N}(\epsilon, \mathcal{G}, d)$, is defined by

$$\mathcal{N}(\epsilon.\mathcal{G}, d) = \inf \left\{ k \geq 1 : \ \exists x_1, ..., x_k \in \mathcal{E} \text{ with } \mathcal{G} \subset \cup_{i=1}^{k} S_{x_i,\epsilon} \right\},$$

where $S_{x,\epsilon}$ the open ball of radius ϵ and center at $x \in \mathcal{E}$. Recall that the logarithm of $\mathcal{N}(\epsilon, \mathcal{G}, d)$ is called metric entropy or ϵ-entropy. The set $\mathcal{G}$ is said to be totally bounded if $\mathcal{N}(\epsilon, \mathcal{G}, d) < \infty$ for all $\epsilon > 0$. In particular, every relatively compact set is totally bounded and all totally bounded sets are bounded. Let us introduce the following hypotheses:

Assumption 1 There exists a sequence $(\mathcal{E}_k)_{k\geq 1}$ of totally bounded subsets of $\mathcal{E}$ such that $\mathcal{E}_k \subset \mathcal{E}_{k+1}$ for all $k \geq 1$ and $\mu\left(\bigcup_{k\geq 1} \mathcal{E}_k\right) = 1$.

Assumption 2 For any positive integer $k \geq 1$, any $\mathbf{i} \neq \mathbf{j}$ and $\epsilon_1 \in]0, 1]$, $\mathbb{P}((X_{\mathbf{i}}, X_{\mathbf{j}}) \in B_{x,b} \times B_{x,b}) \leq C[\mu(B_{x,b})]^{1+\epsilon_1}$, for all $x \in \mathcal{E}_k$, and some $C > 0$.

Assumption 3 The following Besicovich condition holds, for every $\epsilon > 0$,

$$\lim_{h\to 0+} \mu \left\{ x \in \mathcal{E} : \left| \frac{1}{\mu(B_{x,b})} \int_{B_{x,b}} \eta d\mu - \eta(x) \right| > \epsilon \right\} = 0.$$

Note that, for example, Assumption 1 and Assumption 3 are used by Abraham et al. (2006) and Younso (2017). Assumption 2 is used by Ternynck (2014) in the spatial case and by Younso (2017) in the temporal case.
Suppose that the training data $D_{\mathbf{n}}$ is drawn from strictly stationary and arithmetically α-mixing random field in the sense that there exist $C > 0$ and $\theta > 0$ such that

$$\alpha(t) \leq Ct^{-\theta} \text{ for all } t \in \mathbb{R}_+^*. \tag{3}$$

The notion $\mathcal{G}^c$ stands for the complement of any subset $\mathcal{G}$ of $\mathcal{E}$. For simplicity of notation, we write $\mathcal{N}_k(\epsilon)$ instead of $\mathcal{N}(\epsilon, \mathcal{E}_k, \rho)$. The first main result on the consistency is stated in the following theorem.

Theorem 1 **(Consistency)** Let $(\mathcal{E}_k)_{k\geq 1}$ be a sequence of totally bounded subsets of $\mathcal{E}$. Assume that $D_{\mathbf{n}}$ is α-mixing such that (3). If Assumptions 1-3 are satisfied, and $b \to 0$ and for every $k \geq 1$, $\frac{\mathcal{N}_k(b/2)}{\hat{\mathbf{n}}} \longrightarrow 0$ as $\mathbf{n} \to \infty$, then, for $\theta > 2N$,

$$\mathbb{E}L_{\mathbf{n}} \longrightarrow L^* \text{ as } \mathbf{n} \to \infty.$$

Now, in order to establish the strong consistency, we suppose that the training data $D_{\mathbf{n}}$ is drawn from strictly stationary and arithmetically β-mixing random field in the sense that there exist $C > 0$ and $\theta > 0$ such that

$$\beta(t) \leq Ct^{-\theta} \text{ for all } t \in \mathbb{R}_+^*. \tag{4}$$

We suppose that $n_1 = n_2 = ... = n_N = n$. This isotropic assumption is used by El Machkouri (2007). For the sake of simplicity, we will write

$$\mathcal{I}_n := \mathcal{I}_{\mathbf{n}} = \{\mathbf{i} \in \mathbb{Z}^N : \ 1 \leq i_k \leq n, \ \forall k = 1, ..., N\}, \ \hat{\mathbf{n}} = \text{Card}(\mathcal{I}_n) = n^N$$

and

$$g_n(x) := g_{\mathbf{n}}(x) = \begin{cases} 0 \text{ if } \eta_n(x) \leq \frac{\sum_{\mathbf{i}\in\mathcal{I}_n}(1-Y_{\mathbf{i}})\mathbb{1}_{\{X_{\mathbf{i}}\in B_{x,b}\}}}{n^N\mu(B_{x,b})} \\ \\ 1 \text{ otherwise,} \end{cases}$$

where

$$\eta_n(x) := \eta_{\mathbf{n}}(x) = \frac{\sum_{\mathbf{i}\in\mathcal{I}_n} Y_{\mathbf{i}}\mathbb{1}_{\{X_{\mathbf{i}}\in B_{x,b}\}}}{n^N\mu(B_{x,b})},$$

$L_n := \mathbb{P}(g_n(X) \neq Y)$ and $b = b(n)$. Clearly, the limit $\mathbf{n} \to \infty$ is equivalent to the limit $n \to \infty$.

Theorem 2 **(Strong consistency)** Let $(\mathcal{E}_k)_{k\geq 1}$ be a sequence of totally bounded subsets of $\mathcal{E}$. Assume that $D_{\mathbf{n}}$ is β-mixing such that (4) for some $\theta > 2N$, the metric space $(\mathcal{E}, d)$ is Polish and Assumptions 1-3 hold. Let $(k_n)_{n\geq 1}$ be an increasing sequence of positive integers such that

$$\sum_{n\geq 1} \mu(\mathcal{E}_{k_n}^c) < \infty \text{ and } \sum_{n\geq 1} \mathcal{N}_{k_n}\left(\frac{b}{2}\right)\beta(p_n) < \infty,$$

for some integer $p = p_n \in [1, n/2]$ with $p_n \to \infty$ as $n \to \infty$. If $b \to 0$ and

$$\frac{n^N}{Np^N \log(n)\mathcal{N}_{k_n}^2(b/2)} \longrightarrow \infty \quad \text{as} \quad n \to \infty,$$

then,

$$L_n \longrightarrow L^* \quad \text{as} \quad n \to \infty \quad \text{with probability one.}$$

Bibliographie

[1] Abraham, C., Biau, G. and Cadre, B. (2006). On the kernel rule for function classification. *Ann. Inst. Statist. Math.*, 58 , pp. 619–633.

[2] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A probabilitic Theory of Pattern Recognition*, Springer-Verlag, New York.
[3] El Machkouri, M. (2007). Nonparametric regression estimation for random fields in a fixed-design, Statistical Inference for Stochastic Processes,10, No. 1, 29–47.
[4] Kolmogorov, A.N. and Tihomirov, V.M. (1961). ϵ-entropy and ϵ-capacity of sets in functional spaces, American Mathematical Society Translations, 17, No. 2, pp. 277–364.
[5] Ternynck, C. (2014). Spatial regression estimation for functional data with spatial dependency. *Journal de la Société Française de Statistique.*, 155, No. 2, 138–160.
[6] Younso, A. (2017). On nonparametric classification for weakly dependent functional processes. *ESAIM: PS.*, "https://doi.org/10.1051/ps/2017002".