INSTITUT DE FRANCE
Académie des sciences

Comptes Rendus
Mathématique

Ahmad Younso

Consistency of the k-nearest neighbors rule for functional data

Volume 361 (2023), p. 237-242

https://doi.org/10.5802/crmath.402

Les Comptes Rendus. Mathématique *sont membres du Centre Mersenne pour l'édition scientifique ouverte*
www.centre-mersenne.org
e-ISSN : 1778-3569

Comptes Rendus
Mathématique
2023, Vol. 361, p. 237-242
https://doi.org/10.5802/crmath.402

Statistics / *Statistiques*

Consistency of the k-nearest neighbors rule for functional data

Consistance de la règle des k-plus proches voisins pour des données fonctionnelles

Ahmad Younso [a, b]

[a] MISTEA, Université Montpellier, INRAE, Institut Agro,Montpellier, France

[b] Department of mathematical statistics, Damascus university, Damascus, Syria

E-mail: ahmad.younso@inrae.fr

Abstract. The problem of nonparametric classification by k-nearest neighbors rule in a general metric space will be considered. Consistency and strong consistency of the classifier will be established under mild conditions.

Résumé. Le problème de la classification non paramétrique par la règle des k- plus proches voisins dans un espace métrique général sera considéré. La consistance et la forte consistance du classifieur seront établies sous des conditions légères.

2020 Mathematics Subject Classification. 00X99.

Manuscript received 20 April 2021, accepted 26 June 2022.

1. Introduction

Let X be a random variable defined on some probability space $(\Omega, \mathscr{A}, \mathbb{P})$ with values in a separable metric space $(\mathscr{F}, \rho)$ where $\mathscr{F}$ is a function space and ρ denotes the metric on $\mathscr{F}$, and let Y be a random variable with values 0 or 1. In this paper, we study the classical binary supervised classification model for data from $\mathscr{F}$. Given a new incoming observation X, our goal is to predict its corresponding label Y. The distribution of the pair (X, Y) is well defined by (μ, η) where $\mu(B) = \mathbb{P}(X \in B)$, for all Borel sets B on $\mathscr{F}$, and $\eta(x) = \mathbb{E}(Y|X = x)$ the regression function of Y given $X = x$, for all $x \in \mathscr{F}$. In order to predict the unknown label Y of an observation $X = x$, we use a classifier that provides a decision rule for this problem. Formally, a classifier is a measurable mapping $g : \mathscr{F} \longrightarrow \{0, 1\}$. Given a classifier g, its corresponding miss-classification error is given by $L = L(g) = \mathbb{P}\{g(X) \neq Y\}$. In practice, the best classifier is that associated with the smallest possible error. It is well known that the Bayes classifier given by

$$g^*(x) = \mathbb{1}_{\{\eta(x) \geq 1/2\}},$$

where $\mathbb{1}_A$ denotes the indicator function of the set A, leads to the lowest possible miss-classification error, i.e.,

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

Unfortunately, g^* is not available since it depends on the distribution of (X,Y) which is generally unknown. But it is often possible to construct a classifier from a set of n independent and identically distributed copies $D_n = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ of (X,Y). The set D_n is called the training data. Among the various ways to define a classifier from a training data, one of the most wide spread and simplest is the k-nearest neighbors (k-NN) classifier given by

$$g_n(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^n w_{ni} Y_i \leq 1/2 \\ 1 & \text{otherwise,} \end{cases} \tag{1}$$

where $w_{ni} = w_{ni}(x; D_n)$ is $1/k$ if X_i is one of the k-nearest neighbor of x in D_n with respect to the metric ρ, and w_{ni} is zero otherwise with $k = k(n)$ is a sequence of positive integers satisfying

$$k \longrightarrow \infty \quad \text{and} \quad k/n \longrightarrow 0 \text{ as } n \to \infty. \tag{2}$$

Ties are broken by preferring points earlier in the sequence. According to the rule (1), an unclassified element is assigned to the class represented by a majority of its k-nearest neighbors in the training set. If we let $\eta_n(x) = \sum_{i=1}^n w_{ni} Y_i$ be the k-nearest neighbor estimator of $\eta(x)$, the classifier (1) can be re-written as follows

$$g_n(x) = \begin{cases} 0 & \text{if } \eta_n(x) \leq 1/2 \\ 1 & \text{otherwise.} \end{cases}$$

Let $L_n = L(g_n) = \mathbb{P}\{g_n(X) \neq Y\}$ be the miss-classification error of $g_n(x)$. The main challenge in this supervised classification setting is to construct a classifier g_n whose miss-classification error will be as close as possible to L^*. In this context, the classifier $g_n(x)$ is called consistent if

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

and called strongly consistent if

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

A classifier can be consistent for certain class of distribution of (X,Y), but not be consistent for others. The classifier $g_n(x)$ is called (strongly) universally consistent, if it is (strongly) consistent for all distribution of (X,Y). In finite-dimensional spaces, the k-nearest neighbors rule is universally strongly consistent under classical conditions. (see [5] and [9]). [2] prove the consistency of the k-NN rule in separable Hilbert spaces. [3] give some examples showing that the results of [6] on the consistency are no more valid in a general functional metric space $(\mathscr{F},\rho)$ and they establish consistency of the k-NN rule on a separable metric space. [1] show that the moving window rule is not consistent in general metric spaces and give conditions on the space and the regression function to ensure the (strong) consistency of the estimator. More recently, [7] extend the Stone's seminal result to the case of metric spaces when the probability measure of the explanatory variables is tight. Then, under slight variations on the hypotheses, they extend the result to some general metric measure spaces. In this paper, the consistency of the classifier (1) will be proved under weaker assumptions than that of [3] and [7]. Furthermore, the strong consistency of the classifier (1) will also be established under some weak conditions. Denote $B_{x,\epsilon}$ the closed ball of radius $\epsilon > 0$ and center at $x \in \mathscr{F}$. To establish the main results, we will need the following Besicovitch condition, for every $\epsilon > 0$,

$$\lim_{\delta\to 0} \mu\left\{x \in \mathscr{F} : \left|\frac{1}{\mu(B_{x,\delta})}\int_{B_{x,\delta}} \eta \mathrm{d}\mu - \eta(x)\right| > \epsilon\right\} = 0. \tag{3}$$

Note that (3) a classical assumption for this kind of results and it holds for example if $\eta(x)$ is μ-continuous (for more detail on this topic, see [3]). Besicovitch condition plays an important role also in the consistency of kernel rules (see [1]). In finite dimension, (3) holds automatically for any integrable function since it is just the differentiation theorem with respect to a finite measure.

2. Main results

Before we state the main results of this paper, we introduce the following lemma which is needed in the proofs. Denote

$$\text{supp}(\mu) = \{x \in \mathscr{F} : \mu(B_{x,\epsilon}) > 0, \forall\, \epsilon > 0\}.$$

Lemma 1. *If* (2) *is fulfilled then, for each* $x \in supp(\mu)$*, there exists a sequence* $a_n = a_n(x) \to 0$ *as* $n \to \infty$ *such that* $\mu(B_{x,a_n}) \geq \sqrt{k/n}$.

For the proof of Lemma 1, the reader is referred to [7].

2.1. *Consistency of the classifier*

The following theorem state the consistency of the classifier (1).

Theorem 2. *If* (2)–(3) *are fulfilled then the k-NN classifier is consistent.*

Note that classical condition (2) is the same as that used by [6] in the finite dimensional case and by [3] in the infinite dimensional case. Condition (3) is used by [3].

Proof. By Theorem 2.2 in [6], whose extension to the infinite dimensional setting is straightforward, we can write

$$\mathbb{E}(L_n) - L^* \leq 2\mathbb{E}\int_{\mathscr{F}} |\eta(x) - \eta_n(x)|\mu(\mathrm{d}x). \tag{4}$$

Therefore, the theorem will be proved if we show that

$$\mathbb{E}\int_{\mathscr{F}} |\eta(x) - \eta_n(x)|\mu(\mathrm{d}x) \longrightarrow 0 \text{ as } n \to \infty. \tag{5}$$

Define

$$\widehat{\eta}_n(x) = \frac{1}{n\mu(B_{x,a_n})} \sum_{i=1}^{n} Y_i \mathbb{1}_{\{X_i \in B_{x,a_n}\}} \tag{6}$$

with $a_n \to 0$ is the sequence given in Lemma 1. We have

$$\mathbb{E}\int_{\mathscr{F}} |\eta(x) - \eta_n(x)|\mu(\mathrm{d}x) \leq \mathbb{E}\int_{\mathscr{F}} |\eta(x) - \mathbb{E}\widehat{\eta}_n(x)|\mu(\mathrm{d}x) + \mathbb{E}\int_{\mathscr{F}} |\mathbb{E}\widehat{\eta}_n(x) - \eta_n(x)|\mu(\mathrm{d}x). \tag{7}$$

Therefore, it suffices to prove that each term in the right-hand side of (7) tends to zero as $n \to 0$. Besicovitch condition yields as $n \to 0$,

$$\int_{\mathscr{F}} |\eta(x) - \mathbb{E}\widehat{\eta}_n(x)|\mu(\mathrm{d}x) \to 0. \tag{8}$$

It remains to prove that as $n \to 0$,

$$\mathbb{E}\int_{\mathscr{F}} |\mathbb{E}\widehat{\eta}_n(x) - \eta_n(x)|\mu(\mathrm{d}x) \to 0. \tag{9}$$

Clearly, we have

$$\mathbb{E}\int_{\mathscr{F}} |\mathbb{E}\widehat{\eta}_n(x) - \eta_n(x)|\mu(\mathrm{d}x) \leq \mathbb{E}\int_{\mathscr{F}} |\mathbb{E}\widehat{\eta}_n(x) - \widehat{\eta}_n(x)|\mu(\mathrm{d}x) + \mathbb{E}\int_{\mathscr{F}} |\widehat{\eta}_n(x) - \eta_n(x)|\mu(\mathrm{d}x). \tag{10}$$

Therefore, we will prove that each term in the right-hand side of the above inequality tends to zero. Cauchy–Schwartz inequality yields

$$\mathbb{E}|\widehat{\eta}_n(x)-\mathbb{E}\widehat{\eta}_n(x)| \le \mathbb{E}((\widehat{\eta}_n(x)-\mathbb{E}\widehat{\eta}_n(x))^2)^{1/2} \le (\mathrm{var}(\widehat{\eta}_n(x)))^{1/2} \le \left(\frac{\mathbb{E}(Y\mathbb{1}_{\{X\in B_{x,a_n}\}})^2}{n(\mu(B_{x,a_n}))^2}\right)^{1/2}.$$

Since $|Y|\le 1$, we obtain

$$\mathbb{E}|\widehat{\eta}_n(x)-\mathbb{E}\widehat{\eta}_n(x)| \le \left(\frac{1}{n\mu(B_{x,a_n})}\right)^{1/2}. \tag{11}$$

Hence, Lemma 1 yields

$$\frac{1}{n\mu(B_{x,a_n})} \le \frac{\sqrt{n/k}}{n} = \frac{1}{\sqrt{kn}}. \tag{12}$$

By (11)–(12) together with Fubini's theorem and the dominated convergence theorem, we get as $n\to\infty$,

$$\mathbb{E}\int_{\mathscr{F}} |\mathbb{E}\widehat{\eta}_n(x)-\widehat{\eta}_n(x)|\mu(\mathrm{d}x) \to 0. \tag{13}$$

Now, let $X_{(k)}(x)$ be the k-nearest neighbor of x in D_n and denote $r_n = r_n(x) = \rho(X_{(k)}(x),x)$. Clearly,

$$\eta_n(x) = \frac{1}{k}\sum_{i=1}^{n} Y_i \mathbb{1}_{\{X_i\in B_{x,r_n}\}}.$$

Then, since $|Y_i|\le 1$,

$$\begin{aligned}
|\widehat{\eta}_n(x)-\eta_n(x)| &= \left|\frac{1}{n\mu(B_{x,a_n})}\sum_{i=1}^{n} Y_i\mathbb{1}_{\{X_i\in B_{x,a_n}\}} - \frac{1}{k}\sum_{i=1}^{n} Y_i\mathbb{1}_{\{X_i\in B_{x,r_n}\}}\right| \\
&\le \sum_{i=1}^{n}\left|\frac{1}{n\mu(B_{x,a_n})}\mathbb{1}_{\{X_i\in B_{x,a_n}\}} - \frac{1}{k}\mathbb{1}_{\{X_i\in B_{x,r_n}\}}\right| \\
&\le \left|\frac{1}{n\mu(B_{x,a_n})}\sum_{i=1}^{n}\mathbb{1}_{\{X_i\in B_{x,a_n}\}} - 1\right| := |\widetilde{\eta}_n(x)-\mathbb{E}\widetilde{\eta}_n(x)|.
\end{aligned} \tag{14}$$

Observe that $\widetilde{\eta}_n(x) = \widehat{\eta}_n(x)$ if we let $Y_i = 1$ for all $i = 1,\dots,n$. Consequently, the proof of the limit

$$\mathbb{E}\int_{\mathscr{F}} |\widetilde{\eta}_n(x)-\mathbb{E}\widetilde{\eta}_n(x)| \to 0 \tag{15}$$

is the same as that of (13). Finally, by (7)–(10), (13) and (14)–(15) the theorem is proved. □

2.2. *Strong consistency of the classifier*

The following theorem state the strong consistency of the classifier (1).

Theorem 3. *Suppose that* (2)–(3) *are fulfilled. If in addition* $k/(\log n)\to\infty$, *then the* k*-NN classifier is strongly consistent.*

Note that the classical condition $k/(\log n)\to\infty$, used by [4], is crucial to get the strong consistency of the classifier (2).

Proof. By Theorem 2.2 in [6], the theorem will be proved if we show that

$$\int_{\mathscr{F}} |\eta(x)-\eta_n(x)|\mu(\mathrm{d}x) \longrightarrow 0 \quad \text{as } n\to\infty \text{ with probability one.} \tag{16}$$

Using the term $\widehat{\eta}_n(x)$ defined above in (6), we can write

$$\int_{\mathscr{F}} |\eta(x)-\eta_n(x)|\mu(\mathrm{d}x) \le \int_{\mathscr{F}} |\eta(x)-\mathbb{E}\widehat{\eta}_n(x)|\mu(\mathrm{d}x) + \int_{\mathscr{F}} |\mathbb{E}\widehat{\eta}_n(x)-\eta_n(x)|\mu(\mathrm{d}x). \tag{17}$$

As a consequence, by (8), it suffices to show that

$$\int_{\mathscr{F}} |\mathbb{E}\widehat{\eta}_n(x)-\eta_n(x)|\mu(\mathrm{d}x) \to 0 \text{ with probability one as } n\to\infty. \tag{18}$$

Clearly, we have

$$\int_{\mathscr{F}} |\mathbb{E}\widehat{\eta}_n(x) - \eta_n(x)|\mu(\mathrm{d}x) \leq \int_{\mathscr{F}} |\mathbb{E}\widehat{\eta}_n(x) - \widehat{\eta}_n(x)|\mu(\mathrm{d}x) + \int_{\mathscr{F}} |\widehat{\eta}_n(x) - \eta_n(x)|\mu(\mathrm{d}x). \tag{19}$$

Hence, we will prove each term in the right-hand side of the above inequality tends to zero as $n \to \infty$ with probability one. Let us first deal with the first term in the right-hand side of (19). If we replace (X_i, Y_i) by (X_i', Y_i'), suppose that the value of $\widehat{\eta}_n(x)$ is changed to $\widehat{\eta}_{ni}(x)$. Then,

$$\begin{aligned}\Big| \int_{\mathscr{F}} |\mathbb{E}\widehat{\eta}_n(x) - \widehat{\eta}_n(x))|\mu(\mathrm{d}x) - \int_{\mathscr{F}} |\mathbb{E}\widehat{\eta}_n(x) - \widehat{\eta}_{ni}(x)|\mu(\mathrm{d}x)\Big| \\ \leq \int_{\mathscr{F}} |\widehat{\eta}_n(x) - \widehat{\eta}_{ni}(x)|\mu(\mathrm{d}x) \leq \int_{\mathscr{F}} \frac{1}{n\mu(B_{x,a_n})}\mu(\mathrm{d}x) \leq \frac{1}{\sqrt{nk}}. \end{aligned} \tag{20}$$

The last inequality is due to Lemma 1. Now, for any $\epsilon > 0$, and n large enough, we have for n large enough

$$\mathbb{P}\left(\int_{\mathscr{F}} |\mathbb{E}\widehat{\eta}_n(x) - \widehat{\eta}_n(x)|\mu(\mathrm{d}x) > \epsilon\right) \leq \mathbb{P}\left(\left|\int_{\mathscr{F}} |\mathbb{E}\widehat{\eta}_n(x) - \widehat{\eta}_n(x)|\mu(\mathrm{d}x - \int_{\mathscr{F}} |\mathbb{E}\widehat{\eta}_n(x) - \widehat{\eta}_n(x)|\mu(\mathrm{d}x)\right| > \epsilon/2\right).$$

Hence, using McDiarmid's inequality (see [8]) with taking into account (20), we get for n large enough

$$\mathbb{P}\left(\int_{\mathscr{F}} |\mathbb{E}\widehat{\eta}_n(x) - \widehat{\eta}_n(x)|\mu(\mathrm{d}x) > \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2 k}{2}\right)$$

Since $k/(\log n) \to \infty$ by assumption, Borel–Cantelli lemma yields

$$\int_{\mathscr{F}} |\mathbb{E}\widehat{\eta}_n(x) - \widehat{\eta}_n(x)|\mu(\mathrm{d}x) \to 0 \text{ with probability one as } n \to \infty. \tag{21}$$

Let us now deal with the second term in the right-hand side of (19). By (14),

$$|\widehat{\eta}_n(x) - \eta_n(x)| \leq |\widetilde{\eta}_n(x) - \mathbb{E}\widetilde{\eta}_n(x)|, \tag{22}$$

with $\widetilde{\eta}_n(x) = \widehat{\eta}_n(x)$ if we let $Y_i = 1$ for all $i = 1, \ldots, n$. Hence, if we follow the same arguments that use to prove (21), we get

$$\int_{\mathscr{F}} |\widetilde{\eta}_n(x) - \mathbb{E}\widetilde{\eta}_n(x)|\mu(\mathrm{d}x) \to 0 \text{ with probability one as } n \to \infty. \tag{23}$$

Finally, by (8), (17), (19) and (21)–(23) we get (16) and the proof is completed. □

Conclusion. The consistency result of Theorem 2 is obtained under weaker assumptions than that of [7]. If we let $(\mathscr{F}, \rho) = (\mathbb{R}^d, \|\cdot\|)$ with $\|\cdot\|$ denotes the Euclidean norm, then the Stone's theorem is considered as a particular case of Theorem 2. Furthermore, Theorem 3 is an extension of Theorem 11.1 of [6] to a general metric space. In the latter result the authors suppose that μ is absolutely continuous while μ in Theorem 2 is any probability measure.

References

[1] C. Abraham, G. Biau, B. Cadre, "On the kernel rule for function classification", *Ann. Inst. Stat. Math.* **58**.

[2] G. Biau, F. Bunea, M. H. Wegkamp, "Functional classification in Hilbert spaces", *IEEE Trans. Inf. Theory* **51** (2005), no. 6, p. 2163-2172.

[3] F. Cérou, A. Guyader, "Nearest neighbor classification in infinite dimension", *ESAIM, Probab. Stat.* **10** (2006), p. 340-355.

[4] K. Chaudhuri, S. Dasgupta, "Rates of Convergence for Nearest Neighbor Classification", 2014, https://arxiv.org/abs/1407.0067.

[5] L. Devroye, L. Györfi, A. Krzyżak, G. Lugosi, "On the Strong Universal Consistency of Nearest Neighbor Regression Function Estimates", *Ann. Stat.* **22** (1994), no. 3, p. 1371-1385.

[6] L. Devroye, L. Györfi, G. Lugosi, *A probabilistic Theory of Pattern Recognition*, Applications of Mathematics, vol. 31, Springer, 1996.

[7] L. Forzani, R. Fraiman, P. Llop, “Consistent Nonparametric Regression for Functional Data Under the Stone–Besicovitch Conditions”, *IEEE Trans. Inf. Theory* **58** (2012), no. 11, p. 6697-6708.
[8] C. McDiarmid, “On the method of bounded differences”, in *Surveys of combinatorics*, London Mathematical Society Lecture Note Series, vol. 141, Cambridge University Press, 1989, p. 148-188.
[9] C. J. Stone, “Consistent nonparametric regression”, *Ann. Stat.* **5** (1977), no. 4, p. 595-620.