INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Nearest neighbor classification in infinite dimension

Frédéric Cérou and Arnaud Guyader

N° 5536

March 2005

Thème NUM

ISSN 0249-6399 ISRN INRIA/RR--5536--FR+ENG

INRIA
RENNES

Nearest neighbor classification in infinite dimension

Frédéric Cérou* and Arnaud Guyader†

Thème NUM — Systèmes numériques
Projet Aspi

Rapport de recherche n° 5536 — March 2005 — 23 pages

Abstract: Let X be a random element in a metric space $(\mathcal{F}, d)$, and let Y be a random variable with value 0 or 1. Y is called the class, or the label, of X. Assume n i.i.d. copies $(X_i, Y_i)_{1 \leq i \leq n}$. The problem of classification is to predict the label of a new random element X. The k-nearest neighbor classifier consists in the simple following rule: look at the k nearest neighbors of X and choose 0 or 1 for its label according to the majority vote. If $(\mathcal{F}, d) = (\mathbb{R}^d, ||.||)$, Stone has proved in 1977 the universal consistency of this classifier: its probability of error converges to the Bayes error, whatever the distribution of (X, Y). We show in this paper that this result is no more valid in general metric spaces. However, if $(\mathcal{F}, d)$ is separable and if a regularity condition is assumed, then the k-nearest neighbor classifier is weakly consistent.

Key-words: Classification, Consistency, Non parametric statistics

* Frederic.Cerou@inria.fr
† Arnaud.Guyader@uhb.fr

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 Rennes Cedex (France)
Téléphone : +33 2 99 84 71 00 — Télécopie : +33 2 99 84 71 71

Classification par les plus proches voisins en dimension infinie

Résumé : Soit X un élément aléatoire dans un espace métrique $(\mathcal{F}, d)$, et soit Y une variable aléatoire prenant pour valeur 0 ou 1. Supposons n copies i.i.d. $(X_i, Y_i)_{1 \leq i \leq n}$. Le problème de classification consiste à prédire la classe d'un nouvel élément aléatoire X. Le classifieur des k plus proches voisins applique la simple règle suivante : on considère les k plus proches voisins de X et on décide s'il est de la classe 0 ou 1 par un vote à la majorité. Si $(\mathcal{F}, d) = (\mathbb{R}^d, ||.||)$, Stone a prouvé en 1977 la consistence universelle de ce classifieur : sa probabilité d'erreur converge vers l'erreur de Bayes, quelle que soit la loi de (X, Y). Dans cet article nous montrons que ce résultat n'est plus valide dans un espace métrique quelconque. Cependant, si $(\mathcal{F}, d)$ est séparable et si on fait une hypothèse de régularité, alors le classifieur des k plus proches voisins est faiblement consistant.

Mots-clés : Classification, Consistence, Statistiques non paramétriques

1 General definitions and results about classification

Let X be a random element with values in a metric space $(\mathcal{F}, d)$, and let Y be a random variable with values 0 or 1. The distribution of the pair (X, Y) is defined by:

- the probability measure μ of X:

$$\mu(B) = \mathbb{P}(X \in B) \text{ for all Borel sets } B \text{ on } \mathcal{F},$$

- and the regression function η of Y on X:

$$\eta(x) = \mathbb{P}(Y = 1 | X = x) \text{ for all } x \in \mathcal{F}.$$

Assume n independent and identically distributed copies $(X_i, Y_i)_{1 \leq i \leq n}$: they are called the training data, and briefly denoted by D_n. Now we would like to guess the label Y of a new random element X, with $X \sim \mu$ independent of the training data. In this aim, one has to construct a function $g_n : \mathcal{F} \to \{0, 1\}$, called a classifier. This classifier is usually obtained by thresholding an approximation η_n of η. It is easy to prove that the best possible solution is the Bayes classifier (see Figure 1):

$$g^*(x) \stackrel{\Delta}{=} \mathbb{1}_{\{\eta(X) \geq 1/2\}}.$$

We shall precise this point (see [7] for a proof).

Proposition 1 (Optimality of the Bayes classifier). *The quantity $L^* = \mathbb{P}(g^*(X) \neq Y)$ is called the Bayes (probability of) error, or the Bayes risk. For every classifier $g_n : \mathcal{F} \to \{0, 1\}$*

$$\mathbb{P}(g_n(X) \neq Y) \geq L^*.$$

More precisely, if $g_n(x) = \mathbb{1}_{\{\eta_n(x) \geq 1/2\}}$, then:

$$0 \leq \mathbb{P}(g_n(X) \neq Y) - L^* = 2 \int_{\mathcal{F}} |\eta(x) - \frac{1}{2}| \mathbb{1}_{\{g_n(x) \neq g^*(x)\}} \mu(dx) \leq 2\mathbb{E}|(\eta - \eta_n)(X)|.$$

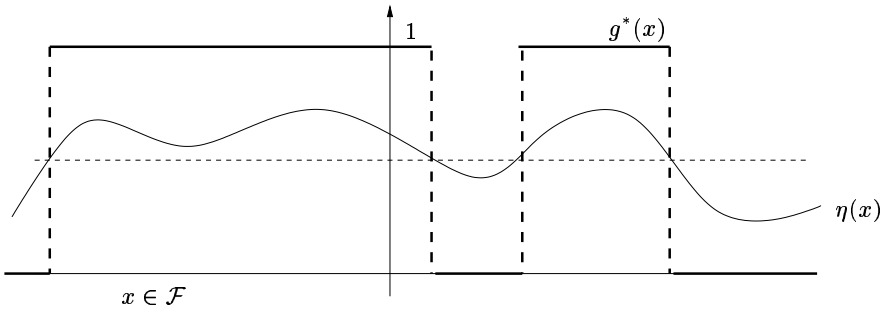

Figure 1: The regression function η and the associated Bayes classifier g^*.

Remark. If the label Y is a deterministic function of X, then $L^* = 0$.

Of course, in general, one does not know the regression function η, nor the Bayes classifier g^*. From now on, we focus our attention on the k-nearest neighbor classifier, sometimes simply called "nearest neighbor classifier" in the following. Let us define the approximate regression function by:

$$\eta_n(X) \stackrel{\Delta}{=} \sum_{i=1}^{n} \frac{1}{k} 1_{\{X_i \in k(X)\}} Y_i = \sum_{i=1}^{k} \frac{1}{k} Y_{(i)},$$

where "$X_i \in k(X)$" means "X_i is one of the k nearest neighbors of X" and the notation $(X_{(i)}, Y_{(i)})$ (we should write $(X_{(i)}(X), Y_{(i)}(X))$ to be completely rigorous) means that the pairs $(X_i, Y_i)_{1 \leq i \leq n}$ have been re-indexed so that:

$$d(X, X_{(1)}) \leq d(X, X_{(2)}) \leq \cdots \leq d(X, X_{(n)}).$$

In case of equality, the ties are broken by comparing auxiliary i.i.d. variables $\beta_1, \ldots, \beta_n$, independent of all the other random objects, and uniformly distributed in $(0, 1)$. This rule has the interesting feature of making all the $n!$ orderings have the same probability to occur.
The associated decision function is

$$g_n(X) \stackrel{\Delta}{=} 1_{\{\eta_n(x) \geq 1/2\}}.$$

The error probability conditional on D_n is defined by:

$$L_n \stackrel{\Delta}{=} \mathbb{P}(g_n(X) \neq Y | D_n).$$

L_n is a random variable, and its expectation $\mathbb{E}[L_n] = \mathbb{P}(g_n(X) \neq Y)$ is a real number depending on the parameters (k, n). We are interested in the asymptotic comportment, which means: $n \to \infty$, $k \to \infty$ and $\frac{k}{n} \to 0$. By convention, in the following, we will simply write " $n \to \infty$" for these asymptotic.

Definition 1 (Universal Consistency). *The k-nearest neighbor classifier is:*

- *universally weakly consistent if:* $\lim_{n\to\infty} \mathbb{E}[L_n] = L^*$.
- *universally strongly consistent if:* $\lim_{n\to\infty} L_n = L^*$ *almost surely.*

The term "universally" means that the result is independent of the distribution μ and of the regression function η. In the following, we are only interested in weak consistency. The principal result is due to Stone.

Theorem 1 (Stone (1977)). *With* $(\mathcal{F}, d) = (\mathbb{R}^d, ||.||)$, *the k-nearest neighbor classifier is universally weakly consistent.*

For the proof, we refer the reader to [7] or [13]. It is based on geometrical result, known as Stone's Lemma. This powerful and elegant argument can unfortunately not be generalized in infinite dimension.

The notation $(\mathcal{F}, d) = (\mathbb{R}^d, ||.||)$ means that the metric d derives from a vector norm on $\mathbb{R}^d$. As we will see in the next section, this point is essential for the validity of the result. The universal strong consistency in $(\mathbb{R}^d, ||.||)$ has been proved by Devroye *et al.* [6].

2 Consistency in general metric spaces

2.1 Separability of the space

To generalize Stone's result, the first natural assumption is the separability of the metric space $(\mathcal{F}, d)$. The following example shows that this condition is

required even in finite dimension.

Example: a pathological distance on $[0,1]$
Let us define the distance d on $[0,1]$ as follows:

$$d(x,x') = \begin{cases} 0 & \text{if } x = x' \\ 1 & \text{if } xx' = 0 \text{ and } x \neq x' \\ 2 & \text{if } xx' \neq 0 \text{ and } x \neq x' \end{cases}$$

Since the triangle inequality is verified, d is a distance on $[0,1]$. But $([0,1],d)$ is clearly not separable.

The distribution μ on $[0,1]$ is very simple: with probability one half, one picks the origin 0 ; with probability one half, one picks a point uniformly in $[0,1]$. Mathematically speaking, if $\lambda_{[0,1]}$ denotes Lebesgue's measure on $[0,1]$ and δ_0 Dirac's measure at the origin:

$$\mu = \frac{1}{2}\delta_0 + \frac{1}{2}\lambda_{[0,1]}$$

The way to attribute the label y of a point x in $[0,1]$ is deterministic: if $x = 0$ then $y = 0$; if $0 < x \leq 1$ then $y = 1$. Y is deterministically attributed, so the Bayes risk L^* is equal to zero. Nevertheless, it is intuitively clear that the asymptotic probability of error with the nearest neighbors rule does not converge to 0:

$$\lim_{n\to\infty} \mathbb{E}[L_n] = \frac{1}{2} > L^* = 0.$$

So the nearest neighbors classifier is not weakly consistent in this context, although we are in finite dimension.

In general metric spaces, the separability assumption is sufficient to have convergence of the nearest neighbor to the point of interest. That's what Cover and Hart noticed in 1967 [3]. From now on we will assume that $(\mathcal{F},d)$ is a separable metric space.

Proposition 2 (Cover and Hart (1967)). *If x is in the support of μ and $\lim_{n\to\infty} k/n = 0$, then $\lim_{n\to\infty} d(X_k(x),x) = 0$ with probability one. If X is*

independent of the data and has probability measure μ, then

$$\lim_{n\to\infty} d(X_k(X), X) = 0$$

with probability one whenever $k/n \to 0$.

We refer to [7] for the proof[1].

2.2 The Besicovich condition

As we will see later, the separability of the metric space is not a sufficient assumption for the consistency of the nearest neighbor classifier. It is necessary to put an assumption on the regularity of the regression function η with respect to the measure μ. More precisely, we will request a differentiation hypothesis that will be called "Besicovich condition". In what follows, we will use the symbol $B_{x,\delta}$ for the closed ball of radius δ centered at x.

Hypothesis ($\mathcal{H}$): Besicovich condition). For every $\varepsilon > 0$

$$\lim_{\delta\to 0} \mu\{x \in \mathcal{F} : \frac{1}{\mu(B_{x,\delta})} \int_{B_{x,\delta}} |\eta - \eta(x)|\, d\mu > \varepsilon\} = 0.$$

Another way to say it is the following convergence in probability:

$$\frac{1}{\mu(B_{X,\delta})} \int_{B_{X,\delta}} |\eta - \eta(X)|\, d\mu \xrightarrow[\delta\to 0]{\mathbb{P}} 0$$

We will discuss this condition in the final section. Let us give now the main result of this paper.

Theorem 2 (Consistency of the nearest neighbor classifier). *If $(\mathcal{F}, d)$ is separable and if Besicovich condition $\mathcal{H}$ is fulfilled, then the nearest neighbor classifier is weakly consistent*

$$\mathbb{E}[L_n] \xrightarrow[n\to\infty]{} L^*.$$

[1]The proof there is written in $\mathbb{R}^d$ with its usual norm, but the argument still works in any separable metric space.

Proof. Thanks to Proposition 1, it is sufficient to show the convergence in L^1:

$$\lim_{n\to\infty} \mathbb{E}|(\eta - \eta_n)(X)| = 0.$$

Let us introduce another approximation $\tilde{\eta}_n$ of the regression function:

$$\tilde{\eta}_n(x) = \frac{1}{k}\sum_{i=1}^{k} \eta(X_{(i)}(x)).$$

Then the triangle inequality gives:

$$\mathbb{E}\,|(\eta - \eta_n)(X)| \leq \mathbb{E}|(\eta - \tilde{\eta}_n)(X)| + \mathbb{E}|(\tilde{\eta}_n - \eta_n)(X)|.$$

- $\mathbb{E}|(\tilde{\eta}_n - \eta_n)(X)|$?
  This step is very classical. Cauchy-Schwarz inequality implies

$$\begin{aligned}\mathbb{E}|(\eta_n - \tilde{\eta}_n)(X)| &\leq (\mathbb{E}[(\eta_n - \tilde{\eta}_n)^2(X)])^{1/2} \\ &= \left\{\mathbb{E}\left[\left(\sum_{i=1}^{k}\frac{1}{k}(Y_{(i)} - \eta(X_{(i)}))\right)^2\right]\right\}^{1/2},\end{aligned}$$

  which gives

$$\mathbb{E}|(\eta_n - \tilde{\eta}_n)(X)| \leq \left\{\frac{1}{k^2}\sum_{1\leq i,j\leq k}\mathbb{E}\left[(Y_{(i)} - \eta(X_{(i)}))\cdot(Y_{(j)} - \eta(X_{(j)}))\right]\right\}^{1/2}.$$

  We use the conditioning trick

$$\begin{aligned}&\mathbb{E}\left[(Y_{(i)} - \eta(X_{(i)}))\cdot(Y_{(j)} - \eta(X_{(j)}))\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[(Y_{(i)} - \eta(X_{(i)}))\cdot(Y_{(j)} - \eta(X_{(j)}))|X_1,\ldots,X_n,X\right]\right],\end{aligned}$$

  and remark that, conditionally on $X_1,\ldots,X_n,X$, the quantities $(Y_{(i)} - \eta(X_{(i)}))$ are i.i.d. Bernoulli random variables with zero mean. So if $i \neq j$

$$\mathbb{E}\left[(Y_{(i)} - \eta(X_{(i)}))\cdot(Y_{(j)} - \eta(X_{(j)}))\right] = 0.$$

Thus there remains only

$$\mathbb{E}|(\eta_n - \tilde{\eta}_n)(X)| \leq \left\{ \frac{1}{k^2} \sum_{i=1}^{k} \mathbb{E}[(Y_{(i)} - \eta(X_{(i)}))^2] \right\}^{1/2},$$

and since $|Y_{(i)} - \eta(X_{(i)})| \leq 1$, we have finally

$$\mathbb{E}|(\eta_n - \tilde{\eta}_n)(X)| \leq \frac{1}{\sqrt{k}},$$

which proves that $\lim_{n\to\infty} \mathbb{E}|(\eta_n - \tilde{\eta}_n)(X)| = 0$.

- $\mathbb{E}|(\eta - \tilde{\eta}_n)(X)|$?
  Let $\mathcal{F}_0$ denote the support of μ. Then

$$\mathbb{E}|(\eta - \tilde{\eta}_n)(X)| = \int_{\mathcal{F}_0} \mathbb{E}|(\tilde{\eta}_n - \eta)(x)| \, d\mu(x).$$

We use the conditioning trick again

$$\begin{aligned} \mathbb{E}|(\tilde{\eta}_n - \eta)(x)| &= \mathbb{E}\left| \frac{1}{k} \sum_{i=1}^{k} (\eta(X_{(i)}) - \eta(x)) \right| \\ &= \mathbb{E}\left[\mathbb{E}\left[\left| \frac{1}{k} \sum_{i=1}^{k} (\eta(X_{(i)}) - \eta(x)) \right| \, | d(x, X_{(k+1)}(x)) \right] \right]. \end{aligned}$$

To simplify the writings, let us denote $d_{(j)} = d_{(j)}(x) = d(x, X_{(j)})$, for $1 \leq j \leq n$. Then, using Lemma 2 given in the appendix, we get:

$$\begin{aligned} & \mathbb{E}\left[\left| \frac{1}{k} \sum_{i=1}^{k} (\eta(X_{(i)}) - \eta(x)) \right| \, | d_{(k+1)} \right] \\ &\leq \mathbb{E}\left[\frac{1}{k} \sum_{i=1}^{k} \left| (\eta(X_{(i)}) - \eta(x)) \right| \, | d_{(k+1)} \right] \\ &= \left(\tilde{\mu}(B_{x, d_{(k+1)}}) \right)^{-1} \int_{B_{x, d_{(k+1)}}} |(\eta(x') - \eta(x))| \, d\tilde{\mu}(x'), \end{aligned}$$

where $\tilde{\mu} = (\mathbb{1}_{U_{x,d_{(k+1)}}} + \frac{1}{2}\mathbb{1}_{S_{x,d_{(k+1)}}})\,\mu$. Now it is clear that for any measurable positive function φ, we have

$$-\frac{1}{2}\int_{B_{x,d_{(k+1)}}} \varphi\, d\mu \leq \int_{B_{x,d_{(k+1)}}} \varphi\, d\tilde{\mu} - \int_{B_{x,d_{(k+1)}}} \varphi\, d\mu \leq 0,$$

so that

$$\begin{aligned}\mathbb{E}\left[\left|\frac{1}{k}\sum_{i=1}^{k}(\eta(X_{(i)}) - \eta(x))\right| |d_{(k+1)}\right] \\ \leq \quad 2\left(\mu(B_{x,d_{(k+1)}})\right)^{-1}\int_{B_{x,d_{(k+1)}}} |(\eta - \eta(x))|\, d\mu.\end{aligned}$$

Thus we have:

$$\mathbb{E}|(\eta - \tilde{\eta}_n)(X)| \leq 2\mathbb{E}\left[\left(\mu(B_{X,d_{(k+1)}})\right)^{-1}\int_{B_{X,d_{(k+1)}}} |(\eta - \eta(X))|\, d\mu\right].$$

Since the random variables $\left(\mu(B_{X,d_{(k+1)}})\right)^{-1}\int_{B_{X,d_{(k+1)}}} |(\eta - \eta(X))|\, d\mu$ are all less than 1, the proof will be complete if we show that they converge in probability to 0. For this, fix $\varepsilon > 0$, then for every $\delta_0 > 0$:

$$\begin{aligned}&\mathbb{P}\left(\left(\mu(B_{X,d_{(k+1)}})\right)^{-1}\int_{B_{X,d_{(k+1)}}} |(\eta - \eta(X))|\, d\mu > \varepsilon\right) \\ &\leq \quad \mathbb{P}(d_{(k+1)(X)} \geq \delta_0) \\ &\qquad + \sup_{0\leq\delta\leq\delta_0} \mathbb{P}\left((\mu(B_{X,\delta})^{-1}\int_{B_{X,\delta}} |(\eta - \eta(X))|\, d\mu > \varepsilon\right).\end{aligned}$$

Now, the first terms goes to 0 thanks to Cover and Hart's result and the second one also thanks to Besicovich assumption $\mathcal{H}$.

■

Remark. In finite dimension, Devroye [5] already mentioned that Besicovich condition was the cornerstone for nearest neighbor estimates as well as for kernel estimates.

3 Discussion

3.1 Continuity of the regression function

It is clear that if η is continuous on $(\mathcal{F}, d)$, then Besicovich condition is automatically fulfilled. Nevertheless, intuitively, continuity is not necessary, since the principle of the nearest neighbor classifier is the following: to guess the label Y of a new point X, just average the labels Y_i for points X_i around X. The continuous version which ensures the validity of this averaging method has an integral form: this is exactly Besicovich condition.

To account for this, we will exhibit an example where the continuity condition on η is not fulfilled, but where the k-nearest neighbor classifier is consistent anyway. Before that, we formulate a stronger but more tractable assumption than Besicovich condition.

Hypothesis ($(\mathcal{H}')$: μ-continuity). For every $\varepsilon > 0$, for μ almost every $x \in \mathcal{F}$

$$\lim_{\delta \to 0} \mu \left\{ z \in \mathcal{F} : |\eta(z) - \eta(x)| > \varepsilon | d(x,z) < \delta \right\} = 0.$$

This is a sort of continuity of η with respect to the measure μ, whence the name μ-continuity (see figure 2). Another way to say it is the following almost sure convergence:

$$\frac{1}{\mu(B_{X,\delta})} \int_{B_{X,\delta})} \mathbb{1}_{\{\eta - \eta(X) > \varepsilon\}} \, d\mu \xrightarrow[\delta \to 0]{} 0 \qquad a.s.$$

Proposition 3 (μ-continuity $\Rightarrow$ Besicovich condition). *If the regression function η is μ-continuous, then Besicovich condition is fulfilled.*

Proof. For μ almost every $x \in \mathcal{F}$, take any $\varepsilon > 0$. Let us consider the following decomposition:

$$\begin{aligned} \frac{1}{\mu(B_{x,\delta})} \int_{B_{x,\delta}} |\eta(z) - \eta(x)| \, d\mu &= \frac{1}{\mu(B_{x,\delta})} \int_{B_{x,\delta}} |\eta(z) - \eta(x)| \mathbb{1}_{\{\eta(z) - \eta(x) \leq \varepsilon\}} \, d\mu \\ &+ \frac{1}{\mu(B_{x,\delta})} \int_{B_{x,\delta}} |\eta(z) - \eta(x)| \mathbb{1}_{\{\eta(z) - \eta(x) > \varepsilon\}} \, d\mu. \end{aligned}$$

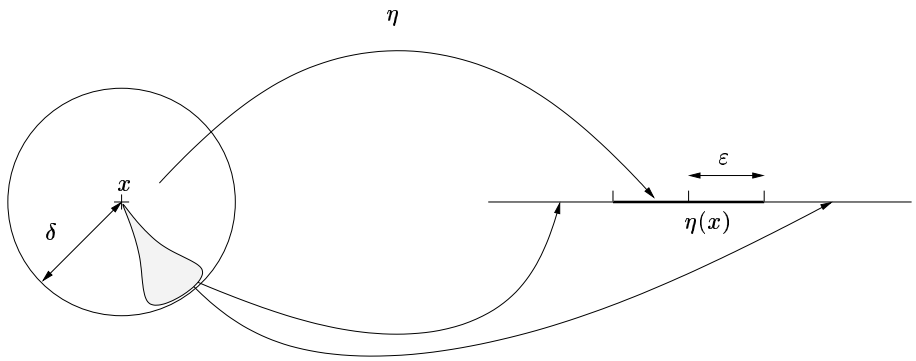

Figure 2: The $\mu-$continuity: another way to see Besicovich condition.

Now we use the fact that for all x and z in $\mathcal{F}^2$, $|\eta(z)-\eta(x)| \leq 1$. This gives:

$$\frac{1}{\mu(B_{x,\delta})}\int_{B_{x,\delta}} |\eta(z)-\eta(x)|\, d\mu \leq \varepsilon + \mu\left\{z \in \mathcal{F} : |\eta(z)-\eta(x)| > \varepsilon | d(x,z) < \delta\right\}$$

So we have

$$\overline{\lim}_{\delta\to 0} \frac{1}{\mu(B_{x,\delta})}\int_{B_{x,\delta}} |\eta(z)-\eta(x)|\, d\mu \leq \varepsilon.$$

Since ε is arbitrary, that gives

$$\lim_{\delta\to 0} \frac{1}{\mu(B_{x,\delta})}\int_{B_{x,\delta}} |\eta(z)-\eta(x)|\, d\mu = 0.$$

Of course, this almost sure convergence implies Besicovich condition $\mathcal{H}$.

■

More precisely, one can easily see that the μ-continuity is rigorously equivalent to the almost sure convergence in Besicovich condition:

$$\frac{1}{\mu(B_{X,\delta})}\int_{B_{X,\delta}} |\eta-\eta(X)|\, d\mu \xrightarrow[\delta\to 0]{} 0 \qquad a.s.$$

As we see in the following example, the $\mu-$continuity may be easier to check than Besicovich condition.

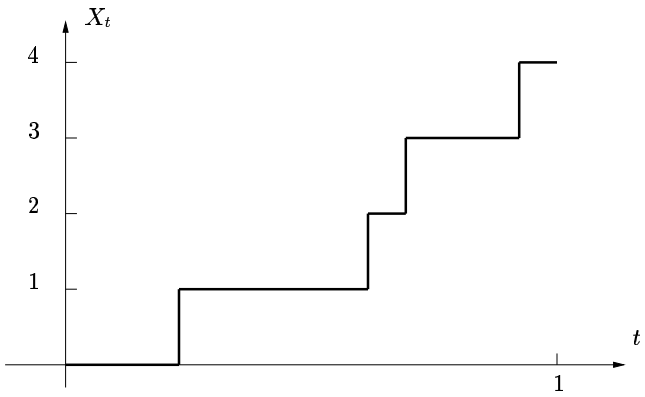

Figure 3: A trajectory $(x_t)_{0\leq t\leq 1}$ of a Poisson process.

Now the example: $\mathcal{F}$ is the space of all Poisson processes of fixed intensity 1 between initial time 0 and final time 1. Its elements are denoted $x = (x_t)_{0\leq t\leq 1}$ or $z = (z_t)_{0\leq t\leq 1}$. The distance on $\mathcal{F}$ is derived from the L_1 norm:

$$d(x,z) = \|x - z\|_1 = \int_0^1 |x_t - z_t|\, dt$$

It is clear that $(\mathcal{F}, ||.||_1)$ is separable: consider for example the processes that jump at rational times between time 0 and time 1. This is a countable set and for every $\delta > 0$ and every $x \in \mathcal{F}$, there exists such a process in the ball $B_{x,\delta}$.

The label of a process x is deterministic and depends only on the final point of the process: if x_1 is even, then $y = 0$. If x_1 is odd, then $y = 1$. Since the label is deterministic, the Bayes risk L^* is null. Moreover, it is not difficult to see that η is nowhere continuous. Indeed, let us fix $x \in \mathcal{F}$, $\delta \in]0,1[$ and consider $z \in \mathcal{F}$ defined as follows (see figure 4):

$$z(t) = \begin{cases} x(t) & \text{if } 0 \leq t \leq 1-\delta \\ x(t)+1 & \text{if } 1-\delta < t \leq 1 \end{cases}$$

So z is at distance δ of x but has not the same label as x: since δ is arbitrary, this proves that η is not continuous at point x. Since x is arbitrary, this proves

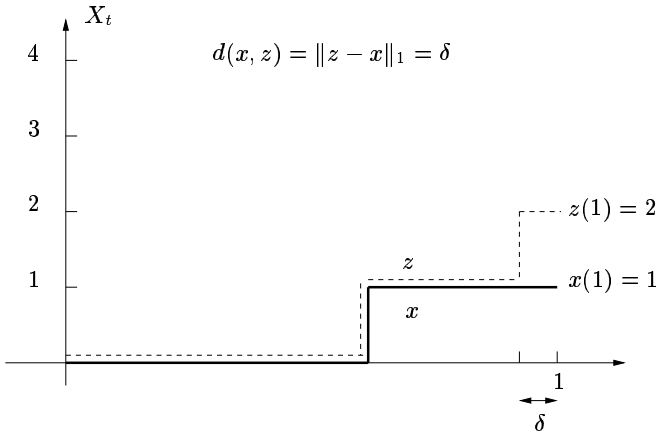

Figure 4: The curves x and x' have not the same label.

that η is nowhere continuous. Nevertheless, we prove now that the nearest neighbor rule is consistent, since Besicovich condition is fulfilled. For this, we use the more tractable formulation $\mathcal{H}'$. Let us fix $\varepsilon > 0$ and $x \in \mathcal{F}$. The aim is to show that

$$\lim_{\delta \to 0} \frac{\mu\left(\{|\eta(z) - \eta(x)| > \varepsilon\} \cap B_{x,\delta}\right)}{\mu(B_{x,\delta})} = 0. \tag{1}$$

We can estimate this quantity.

Lemma 1 (The ratio of small balls).

$$\lim_{\delta \to 0} \frac{\mu(\{|\eta(z) - \eta(x)| > \varepsilon\} \cap B_{x,\delta})}{\mu(B_{x,\delta})} = 0.$$

Proof. In a first time, let us suppose that the number of jumps M of the process x is strictly positive and denote $t_1, \ldots, t_M$ the times of jumps. For the denominator, a process z is in $B_{x,\delta}$ if it jumps at times $t'_1, \ldots, t'_M$ which are close enough to $t_1, \ldots, t_M$:

$$\forall i \in \{1, \ldots, M\} \qquad |t'_i - t_i| < \frac{\delta}{M}.$$

Some calculation on the Poisson process gives

$$\mu\left(z : \forall i \in \{1,\dots,M\},\ |t'_i - t_i| < \frac{\delta}{M}\right) \sim \left(\frac{2\delta}{M}\right)^M.$$

But doing like this, we have not enumerated all processes $z \in B_{x,\delta}$, so we can only conclude that

$$\mu(B_{x,\delta}) \geq f(\delta)\left(\frac{2\delta}{M}\right)^M \qquad with\ \lim_{\delta\to 0} f(\delta) = 1.$$

For the numerator, since η takes values 0 and 1, we have

$$\mu\left(\{|\eta(z) - \eta(x)| > \varepsilon\} \cap B_{x,\delta}\right) = \mu\left(\{\eta(z) \neq \eta(x)\} \cap B_{x,\delta}\right).$$

Thus, if we consider the processes z with M jumps at times $t'_1,\dots,t'_M$ such that

$$\forall i \in \{1,\dots,M\} \qquad |t'_i - t_i| < \delta,$$

and other jumps at times $t'_{M+1},\dots$ such that

$$\forall i > M \qquad |t'_i - 1| < \delta,$$

we get a set S of Poisson processes which is bigger than the one of interest. Briefly speaking

$$(\{|\eta(z) - \eta(x)| > \varepsilon\} \cap B_{x,\delta}) \subset S$$

Some calculation on the Poisson process gives this time

$$\mu(S) \sim (2\delta)^{M+1},$$

so that

$$\mu\left(\{|\eta(z) - \eta(x)| > \varepsilon\} \cap B_{x,\delta}\right) \leq g(\delta)(2\delta)^{M+1} \qquad with\ \lim_{\delta\to 0} g(\delta) = 1.$$

The ration of small balls is now

$$\frac{\mu(\{|\eta(z) - \eta(x)| > \varepsilon\} \cap B_{x,\delta})}{\mu(B_{x,\delta})} \leq \frac{g(\delta)(2\delta)^{M+1}}{f(\delta)\left(\frac{2\delta}{M}\right)^M} \xrightarrow[\delta\to 0]{} 0.$$

If x has no jump, the reasoning is the same.

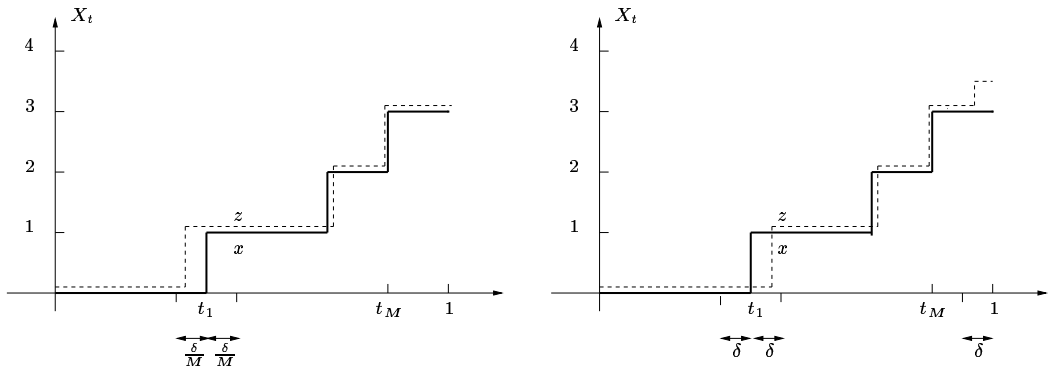

Figure 5: Estimation of the numerator and denominator of equation (1).

■

The result of the Lemma ensures that for every $x \in \mathcal{F}$ and for every $\varepsilon > 0$

$$\lim_{\delta \to 0} \mu \left\{ z \in \mathcal{F} : |\eta(z) - \eta(x)| > \varepsilon | d(x,z) < \delta \right\} = 0.$$

As we have seen before, this implies Besicovich condition. So the nearest neighbor rule is consistent, although η is nowhere continuous.

3.2 Besicovich assumption in infinite dimension

In this section, we discuss the Besicovich condition. If we restrict ourselves to finite dimension and suppose that distance d is issued from a vectorial norm, the essential result is the following one (see for instance [8], Chapter 1.7, pp 43-44).

Theorem 3 (Lebesgue-Besicovich differentiation theorem). *Let μ be a Radon measure on $\mathbb{R}^d$ and $f \in L^p_{\text{loc}}(\mathbb{R}^d)$, then*

$$\lim_{\delta \to 0} \frac{1}{\mu(B_{x,\delta})} \int_{B_{x,\delta}} |f - f(x)|^p \, d\mu = 0$$

for μ almost every x.

In our case, μ is a probability measure on $\mathbb{R}^d$ and η is bounded by 1, so this result can directly be applied. Devroye [5] already noticed that this is another way to prove Stone's theorem.

Corollary 1 (Stone's theorem). *In* $(\mathbb{R}^d, ||.||)$*, the* k*-nearest neighbor classifier is universally weakly consistent.*

Remark. If we wish to estimate the regression function η, our reasoning even shows that with the nearest neighbor method

$$\lim_{n\to\infty} \mathbb{E}\left[|\eta_n(x) - \eta(x)|^p\right] = 0$$

for μ almost every x.

Besicovich condition $\mathcal{H}$ appears also in recent papers on connected problems: Abraham, Biau and Cadre [1] use it for function classification with the kernel rule. Dabo-Niang and Rhomari [4] use it for nonparametric regression estimation in general metric spaces.

Now the question is: what about Besicovich density condition in general metric spaces[2] ?

This question has been studied by several authors in geometric measure theory, see for example [11], [12] and [9]. Unlike the situation in $\mathbb{R}^d$, this condition is no more automatically fulfilled in infinite dimension. In [11], Preiss introduces a rather technical notion, called the σ-finite dimensionality of a metric on a space. He shows that it is the *sine qua non* condition to get Besicovich property for all measures on a metric space. Without delving into the details of this notion, let us just mention that it is related to the σ-finite dimensionality of the space. In fact, reconsidering Poisson processes above, we were precisely in this situation.

Example. Fix $M \geq 0$ and denote $\mathcal{F}_M$ all the Poisson processes of $\mathcal{F}$ that have exactly M jumps. A process that has M jumps can be summarized in an M-dimensional vector, the times of jumps. Then it is obvious that the

[2] Of course, we still suppose that the metric space is separable.

metric space $(\mathcal{F}_M, \|.\|_1)$ is isometric to $([0,1]^M, \|.\|_1)$. So that we have the correspondence

$$(\mathcal{F}, \|.\|_1) = \bigcup_{M=0}^{+\infty} (\mathcal{F}_M, \|.\|_1) \sim \bigcup_{M=0}^{+\infty} ([0,1]^M, \|.\|_1),$$

and the σ-finite dimensionality is clear.

Let us focus now on the classical situation where $(\mathcal{F}, d)$ is a separable Hilbert space and μ a Gaussian measure. Let ν denote the centered and normalized Gaussian measure on $\mathbb{R}$, let (c_n) be a non-increasing sequence of positive numbers such that $\sum_{n=0}^{+\infty} c_n < +\infty$ and let $\ell_2(c)$ be the set of all sequences $x = (x_n)$ such that

$$|x|^2 = \sum_{n=0}^{+\infty} c_n x_n^2 < +\infty.$$

Then the measure $\mu = \nu^{\otimes \mathbb{N}}$ is a σ-additive measure in Hilbert space $\ell_2(c)$. More precisely, each Gaussian measure can be represented in this way.

Even in this rather comfortable context, one has to put conditions on the sequence (c_n) to get Besicovich property. Precisely, Preiss and Tišer [12] have shown the following result: if there exists $q < 1$ such that

$$\forall n \in \mathbb{N} \qquad \frac{c_{n+1}}{c_n} < q,$$

then Besicovich property is true for every function $f \in L^1(\mu)$. Roughly speaking, if we see (c_n) as the sequence of variances of μ along the dimensions, it means that these variances have to decay exponentially fast: this is a very strong condition.

Now let us see an example which shows that if Besicovich condition is not fulfilled, there is not much to hope about classification with the nearest neighbor rule. This example is due to Preiss [10].

Example: a problematic case for nearest neighbor classification
In this paper, Preiss constructs a Gaussian measure μ in a separable Hilbert

space $\mathcal{F}$ and a Borel set $M \subset F$ with $\mu(M) < 1$ such that

$$\lim_{\delta \to 0} \frac{\mu(M \cap B_{x,\delta})}{\mu(B_{x,\delta})} = 1$$

for μ almost every $x \in \mathcal{F}$.

Now suppose that X is distributed with respect to μ and its label Y is deterministic

$$Y = \mathbb{1}_M(X)$$

As usual the Bayes risk is equal to 0. Nevertheless, if we try to apply the nearest neighbor rule to this example, it is intuitively clear that we have some problems to classify elements $x \in \overline{M}$. Indeed, one can easily prove that

$$\underline{\lim}_{n \to \infty} L_n^* \geq \frac{1}{2}\mu(\overline{M}) > L^* = 0.$$

It is worth precising that this result is not in contradiction with the one of Biau *et al.* [2]. In this paper, they consider a random variable X taking values in a separable Hilbert space $\mathcal{X}$, with label $Y \in \{0, 1\}$. They establish the universal weak consistency of the a neighbor-**type** classifier, but not of the **classical** nearest neighbor classifier. More precisely, they reduce the infinite dimension of $\mathcal{X}$ by considering only the first d coefficients of the Fourier series expansion of each X_i, and then perform nearest neighbor classification in $\mathbb{R}^d$. In fact, their result and the example above suggest that in infinite dimension, the classical nearest neighbor classification is not the good way to proceed.

A Technical lemma

In this section we use the notations of the proof of Theorem 2. And for all x in $\mathcal{F}$ and $r \geq 0$ we denote respectively by $B_{x,r}$, $U_{x,r}$ and $S_{x,r}$ the closed ball, the open ball and the sphere centered at x and of radius r. We recall that in case of equality, the ties are broken by comparing auxiliary i.i.d. variables $\beta_1, \ldots, \beta_n$, independent of all the other random objects, and uniformly distributed in $(0, 1)$.

Lemma 2. *Let F be a μ-integrable real function on $\mathcal{F}$. For all x in the support of μ,*

$$\mathbb{E}[\frac{1}{k}\sum_{j=1}^{k} F(X_{(j)})|\, d_{(k+1)}] = C\int_{B_{x,d_{(k+1)}}} F(x')\, d\tilde{\mu}(x'),$$

where $\tilde{\mu} = (\mathbb{1}_{U_{x,d_{(k+1)}}} + \frac{1}{2}\mathbb{1}_{S_{x,d_{(k+1)}}})\,\mu$, and C is a normalizing $d_{(k+1)}$-measurable constant

$$C = \left(\tilde{\mu}(B_{x,d_{(k+1)}})\right)^{-1}.$$

Proof. Let $\mathcal{Q}(n)$ be the set of all the n-permutations, Q denote the random permutation given by the ordering of the nearest neighbor, and $\tilde{\mathcal{Q}}(n,k)$ all the subsets of k elements in $\{1,\ldots,n\}$. C will denote either a deterministic constant, or a $d_{(k+1)} = d_{Q(k+1)}$ measurable random variable, which may change from line to line, but stays uniform in F. We have

$$\begin{aligned}
&\mathbb{E}[\sum_{j=1}^{k} F(X_{(j)})|\, d_{(k+1)}] \\
&= \mathbb{E}[\sum_{j=1}^{k} F(X_{Q(j)}) \sum_{q\in\mathcal{Q}(n)} \mathbb{1}_{Q=q}|\, d_{Q(k+1)}] \\
&= \sum_{q\in\mathcal{Q}(n)} \mathbb{E}[\sum_{j=1}^{k} F(X_{q(j)})\mathbb{1}_{Q=q}|\, d_{Q(k+1)}] \\
&= \sum_{\{q(1),\ldots,q(k)\}\in\tilde{\mathcal{Q}}(n,k)} \mathbb{E}[\sum_{j=1}^{k} F(X_{(j)})\mathbb{1}_{\{Q(1),\ldots,Q(k)\}=\{q(1),\ldots,q(k)\}}|\, d_{Q(k+1)}] \\
&= C\,\mathbb{E}[\sum_{j=1}^{k} F(X_j)\mathbb{1}_{\{Q(1),\ldots,Q(k)\}=\{1,\ldots,k\}}|\, d_{Q(k+1)}].
\end{aligned}$$

The last two equalities come from reordering the terms in the summation, and the fact that all the orderings have the same probability. Then we decompose

the event:

$$\mathbb{1}_{\{Q(1),\ldots,Q(k)\}=\{1,\ldots,k\}} = \prod_{j=1}^{k}(\mathbb{1}_{d_j<d_{Q(k+1)}} + \mathbb{1}_{\beta_j<\beta_{Q(k+1)}}\mathbb{1}_{d_j=d_{Q(k+1)}}) \times \prod_{h=k+1}^{n}(\mathbb{1}_{d_h>d_{Q(k+1)}} + \mathbb{1}_{\beta_h\geq\beta_{Q(k+1)}}\mathbb{1}_{d_h=d_{Q(k+1)}}).$$

It is quite obvious that the two products are independent conditionally to $d_{Q(k+1)}$, so we put the expectation of the second one into the C. Thus we have:

$$\begin{aligned}
&\mathbb{E}[\sum_{j=1}^{k} F(X_{(j)})|\, d_{(k+1)}] \\
&= C\,\mathbb{E}[\sum_{j=1}^{k} F(X_j)\prod_{h=1}^{k}(\mathbb{1}_{d_h<d_{Q(k+1)}} + \mathbb{1}_{\beta_h<\beta_{Q(k+1)}}\mathbb{1}_{d_h=d_{Q(k+1)}})|\, d_{Q(k+1)}] \\
&= C\sum_{j=1}^{k}\mathbb{E}[F(X_j)\prod_{h=1}^{k}(\mathbb{1}_{d_h<d_{Q(k+1)}} + \mathbb{1}_{\beta_h<\beta_{Q(k+1)}}\mathbb{1}_{d_h=d_{Q(k+1)}})|\, d_{Q(k+1)}] \\
&= k\,C\,\mathbb{E}[F(X_1)(\mathbb{1}_{d_1<d_{Q(k+1)}} + \mathbb{1}_{\beta_1<\beta_{Q(k+1)}}\mathbb{1}_{d_1=d_{Q(k+1)}})|\, d_{Q(k+1)}] \\
&= k\,C\,\mathbb{E}[F(X_1)(\mathbb{1}_{d_1<d_{Q(k+1)}} + \frac{1}{2}\mathbb{1}_{d_1=d_{Q(k+1)}})|\, d_{Q(k+1)}] \\
&= k\,C\int_{B_{x,d_{(k+1)}}} F(x')\,d\tilde{\mu}(x'),
\end{aligned}$$

where we used the following facts: the samples X_i are i.i.d., the other terms in the product are conditionally independent of X_1, and the β_i are i.i.d. and independent of the other random variables.
The value of C is then easily computed by taking F constant equal to 1. Also note $C \neq 0$ because x is in the support of μ.

■

Remark. If the probability μ does not put mass on the spheres, the proof is much simpler and the result of the Lemma is merely

$$\mathbb{E}[\frac{1}{k}\sum_{j=1}^{k} F(X_{(j)})|\, d_{(k+1)}] = \frac{1}{\mu(B_{x,d_{(k+1)}})}\int_{B_{x,d_{(k+1)}}} F(x')\,d\mu(x').$$

This can be seen as a particular case of the following general decorrelation result: if $\varphi : \mathbb{R}^k \to \mathbb{R}$ is a test function, symmetric in all its entries, then

$$\mathbb{E}[\varphi(X_{(1)}, \ldots, X_{(k)}) | \, d_{(k+1)}] = \mathbb{E}[\varphi(Z_1, \ldots, Z_k) | \, d_{(k+1)}]$$

with the $(Z_i)_{1 \leq i \leq k}$ i.i.d. random variables distributed according to the restriction of μ on the ball $B_{x, d_{(k+1)}}$.

Acknowledgments The authors would like to thank Bernard Delyon and Samy Abbes for valuable discussions during this work.

References

[1] Christophe Abraham, Gérard Biau, and Benoît Cadre. On the kernel rule for function classification. *submitted*, 2003.

[2] Gérard Biau, Florentina Bunea, and Marten H. Wegkamp. On the k-ernel rule for function classification. *IEEE Transactions on Information Theory, to appear*, 2005.

[3] Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, IT-13(1):21–27, January 1967.

[4] Sophie Dabo-Niang and Noureddine Rhomari. Nonparametric regression estimation when the regressor takes its values in a metric space. *submitted*, 2001.

[5] Luc Devroye. On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Statist.*, 9(6):1310–1319, 1981.

[6] Luc Devroye, László Györfi, Adam Krzyżak, and Gábor Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *Ann. Statist.*, 22(3):1371–1385, 1994.

[7] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

[8] Lawrence C. Evans and Ronald F. Gariepy. *Measure theory and fine properties of functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1992.

[9] Herbert Federer. *Geometric measure theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969.

[10] David Preiss. Gaussian measures and the density theorem. *Comment. Math. Univ. Carolin.*, 22(1):181–193, 1981.

[11] David Preiss. Dimension of metrics and differentiation of measures. In *General topology and its relations to modern analysis and algebra, V (Prague, 1981)*, volume 3 of *Sigma Ser. Pure Math.*, pages 565–568. Heldermann, Berlin, 1983.

[12] David Preiss and Jaroslav Tišer. Differentiation of measures on Hilbert spaces. In *Measure theory, Oberwolfach 1981 (Oberwolfach, 1981)*, volume 945 of *Lecture Notes in Math.*, pages 194–207. Springer, Berlin, 1982.

[13] Charles J. Stone. Consistent nonparametric regression. *Ann. Statist.*, 5(4):595–645, 1977. With discussion and a reply by the author.

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399