

Some issues on supervised classification for Gaussian processes¹

Amparo Baíllo(*), Juan Cuesta-Albertos(**) and Antonio Cuevas²(*)

(*) Departamento de Matemáticas, Univ. Autónoma de Madrid

(**) Departamento de Matemáticas, Estadística y Computación
Universidad de Cantabria

Abstract: We consider the classical discrimination problem (supervised classification) between two “populations”, P_0 and P_1 , of curves on $[0, 1]$. The respective functions in P_0 and P_1 are the trajectories of two stochastic processes with Gaussian distributions, μ_0 and μ_1 . The present communication relies in some results (dating at least from the 1960’s) which provide sufficient conditions to ensure the absolute continuity $\mu_1 \ll \mu_0$ and, more importantly, the explicit expression of the corresponding Radon-Nikodym derivative $d\mu_1/d\mu_0$. We will show how these expressions can be used in two ways in the discrimination problem: (a) To guarantee the consistency of the nearest-neighbor classifier and (b) To suggest new plug-in classifiers based on the estimation of the Radon-Nikodym derivatives.

Keywords: Functional data, supervised classification, nearest-neighbours classifiers, triangular covariance functions.

1. Introduction

The statement of the discrimination (supervised classification) problem is well known: Suppose that a random variable X can be observed in two populations P_0 and P_1 . Denote by μ_i , for $i = 0, 1$ the distribution of X when observed in P_i . In general, μ_0 and μ_1 are not assumed to be completely known but it is usual to impose some restrictions on them; for example, it is quite common to assume that μ_i belong to some given family of distributions and, in any case, it is always assumed that $\mu_0 \neq \mu_1$. The available data consist of a sample of independent observations $\{(X_i, Y_i), 1 \leq i \leq n\}$ where $Y_i = 0$ if the i -th individual, with observation X_i , belongs to P_0 and $Y_i = 1$ if X_i has been observed in P_1 . The problem is to decide, from the information provided by the training sample, if a new observation $X = x$ (for which the value of Y is not known) has been taken in P_0 or in P_1 .

In the classical (multivariate) case, the *feature space* $\mathcal{X}$ where the random observations X_i take values is just the d -dimensional Euclidean space $\mathcal{X} = \mathbb{R}^d$. We are however concerned here with the (infinite-dimensional) functional situation in which $\mathcal{X}$ is a function

¹This note summarizes work in progress still unfinished at the moment of writing this note.

²Presenting author. Address for correspondence: Departamento de Matemáticas, Facultad de Ciencias, Universidad Autónoma de Madrid, 28049-Madrid, Spain. E-mail: antonio.cuevas@uam.es. This work has been partially supported by Spanish grants MTM2007-666032 (first and third author) and MTM2008-06067-C02-02 and the Consejería de Educación y Cultura de la Junta de Castilla y León, GR150 (second author).

space, for example $\mathcal{X} = \mathcal{C}[0, 1]$, the space of real continuous functions defined on the interval $[0, 1]$, endowed with the usual supremum norm $\|\cdot\|$.

The mathematical aim is to find a “classifier” (or “classification rule”) $g : \mathcal{X} \rightarrow \{0, 1\}$ that minimizes, at least asymptotically, the classification error (often called “risk”) $P(g(\mathbf{X}) \neq Y)$. It can be shown (e.g., Devroye et al., 1996, p. 11) that the optimal classification rule (or “Bayes rule”) is

$$g^*(x) = \mathbf{1}_{\{\eta(x) > 1/2\}}, \quad (1)$$

where $\mathbf{1}_A$ denotes the indicator function of the set A and $\eta(x) = \mathbb{E}(Y|X = x)$. This rule is usually unknown but it can be approximated, in different ways, from the training data.

The problem of multivariate (finite-dimensional) discrimination is a standard issue in statistical theory and practice since, at least, seventy years ago. The book by Devroye et al. (1996) is a good general reference with a clear and deep treatment of the mathematical aspects. So far the functional case has received much less attention but there is a current growing interest in the subject coinciding with the general development of the functional data analysis (FDA). The monographs by Ramsay and Silverman (2005) and Ferraty and Vieu (2006) include interesting accounts on supervised classification with functional data. A recent survey on the subject can be found in Baíllo et al. (2009).

2. Gaussian processes with triangular covariance functions

In this communication we are particularly concerned with functional discrimination between two populations P_0 and P_1 whose corresponding distributions, μ_0 and μ_1 are Gaussian processes on $[a, b]$, with mean functions denoted by $m_j(t)$, for $j = 0, 1$, and with covariance functions $\Gamma_j(s, t) = u_j(\min(s, t)) v_j(\max(s, t))$, for $s, t \in [a, b]$, where $u_j > 0, v_j > 0$, and $u_j, v_j \in C^2[a, b]$ for $j = 0, 1$. In other words, the available data are of type (X_i, Y_i) where the X_i are random trajectories $X_i = X_i(t) = X_i(t, w)$ with $t \in [a, b]$, coming either from P_0 or P_1 , as indicated by the value of Y_i ; that is, the conditional distribution $X|Y = j$ is μ_j for $j = 0, 1$.

Covariance functions of type $\Gamma(s, t) = u(\min(s, t)) v(\max(s, t))$, are called *triangular*. Some important processes, as the Brownian motion or the Ornstein-Uhlenbeck process, have a covariance structure of this type. Moreover, Varberg (1964) has derived an interesting integral representation for processes with a triangular covariance function.

The reason of considering here this kind of processes is that there exist some results for them which ensure the mutual absolute continuity $\mu_0 \ll \mu_1, \mu_1 \ll \mu_0$ and, more importantly, provide the explicit expressions of the respective Radon-Nikodym derivatives $d\mu_0/d\mu_1$ and $d\mu_1/d\mu_0$. It turns out that these expressions are simple enough to be used for different practical purposes related to the discrimination problem between P_0 and P_1 . In particular, we will show (Section 2) the consistency of the k -NN classifier in the discrimination problem between P_0 and P_1 and we will consider (Section 3) the construction of new plug-in classifiers based on the estimation of $d\mu_0/d\mu_1$ and $d\mu_1/d\mu_0$.

3. k -NN classifiers

A simple, intuitive way to classify a datum x is given by the following k -NN rule: given a integer $k = k_n$ we look at the values of X_i in the training sample. Then x is assigned to

P_1 if and only if the majority of the k training data closest to x correspond to observations coming from P_1 . In formal terms, define

$$\eta_n(x) = \frac{1}{k} \sum_{i=1}^n \mathbf{1}_{\{X_i \in k(x)\}} Y_i \quad (2)$$

where “ $X_i \in k(x)$ ” means that X_i is one of the k nearest neighbours of x . Then, the k -NN classifier is given by $g_n(x) = \mathbf{1}_{\{\eta_n(x) > 1/2\}}$, where $g_n(x) = j$ means that x is assigned to P_j for $j = 0, 1$ and $\mathbf{1}_A$ denotes the indicator of A .

In the finite dimensional case (where X takes values in $\mathbb{R}^d$) the k -NN classifier is known to be *universally consistent* (Stone, 1977) in the sense that the probability of classification error (conditional to the sample) converges (in probability) as $n \rightarrow \infty$ to that of the optimal rule (1), provided that $k_n \rightarrow \infty$ and $k_n/n \rightarrow 0$, with no restriction at all on the distribution of (X, Y) . See Devroye et al. (1996) for details.

The interesting fact is that in the infinite dimensional case such universal consistency does not longer hold: However, in Cérou and Guyader (2006, Th. 2) it is shown that under an appropriate assumption (called *Besicovich condition*) on $\eta(x) = E(Y|X = x)$ the consistency (in probability) of the k -NN classifier holds whenever $k_n \rightarrow \infty$ and $k_n/n \rightarrow 0$.

It turns out that when $\eta(x)$ is continuous, this Besicovich condition is valid. We will first establish, as a useful auxiliary result, the following inequality

$$|\eta(z) - \eta(x)| \leq \frac{p}{1-p} \left| \frac{d\mu_0}{d\mu_1}(x) - \frac{d\mu_0}{d\mu_1}(z) \right|, \quad (3)$$

where $p = P\{Y = 0\} \in (0, 1)$. Then, we will show (as a direct consequence from some results by Varberg (1961) and Jørsboe (1968)) that for some Gaussian processes with distributions μ_0 and μ_1 and triangular covariance functions, the Radon-Nikodym derivative $d\mu_0/d\mu_1(x)$ is continuous so that, from (3) we conclude the continuity of $\eta(x)$ and therefore the consistency of the k -NN classifier.

4. Plug-in classifiers

Since, under some broad conditions the regression function can be expressed by

$$\eta(x) = \frac{1-p}{p \frac{d\mu_0}{d\mu_1}(x) + 1-p}, \quad (4)$$

for x in the common support of μ_0 and μ_1 , we may think of estimating $d\mu_0/d\mu_1(x)$ and plugging-in the result in the expression (1) of the optimal classifier.

As mentioned above, in the case of Gaussian processes with a triangular function the expression of this Radon-Nikodym derivative is explicitly known under some additional assumptions. For example, Jørsboe (1968, p. 61) considers the case where $m_1 = 0$, and the other mean function $m_0 = m$ is a function in $C^2[a, b]$, such that $m(a) = 0$ whenever $u(a) = 0$. Also both covariance functions are equal to $\Gamma(s, t) = u(\min(s, t)) v(\max(s, t))$, for $s, t \in [a, b]$, where $u, v \in C^2[a, b]$ are positive functions and v and $vu' - uv'$ are bounded away from zero on $[a, b]$. Under these conditions, μ_0 and

μ_1 are equivalent measures with the following Radon-Nikodym derivative:

$$\frac{d\mu_0}{d\mu_1}(x) = \exp \left\{ D_1 + D_2 x(a) + \frac{1}{2} \int_a^b g(t) d \left(\frac{2x(t) - m(t)}{v(t)} \right) \right\} \quad (5)$$

where D_1 and D_2 are some constants depending on $u(a)$, $v(a)$ and $m(a)$ and $f(t)$ is a function depending on $u(t)$, $v(t)$, $m(t)$ and their first derivatives.

The relative simplicity of expression (5), which can be handled integrating by parts, suggests the idea of replacing m , u and v (and their relevant derivatives) with suitable estimators thus obtaining, via (1) and (4), a plug-in version of the optimal classifier. This can be done in several ways, including the possibility of using some extra parametric assumptions on the structure of Γ . These topics are currently under study.

5. Practical aspects

We will also comment the results of some simulation studies and real data examples which complement those given in Cuevas et al. (2007). They are aimed to compare the efficiency of several standard functional classifiers with that of the simple k -NN rule. The results are not by any means conclusive. In general terms they suggest a quite reliable behavior of the k -NN classifier which appears as a reasonable all-purposes rule with a reasonable behavior when compared with other more sophisticated alternatives. At the time of writing this summary we don't have yet the outputs corresponding to the implementation of the plug-in classifiers mentioned in the previous section.

REFERENCES

- Baíllo, A., Cuevas, A. and Fraiman, R. (2009). Classification methods for functional data. In *Oxford Handbook on Statistics and FDA*, F. Ferraty and Y. Romain editors. Oxford University Press, Oxford.
- Cérou, F. and Guyader, A. (2006). Nearest neighbor classification in infinite dimension. *ESAIM: Probability and Statistics* 10, 340-355.
- Cuevas, A., Febrero, M and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics* 22, 481-496.
- Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag.
- Ferraty, F. and Vieu, P. (2006). *Nonparametric Modelling for Functional Data*. Springer.
- Jørsboe, O. G. (1968). Equivalence or Singularity of Gaussian Measures on Function Spaces. Various Publications Series, No. 4, Matematisk Institut, Aarhus Universitet, Aarhus.
- Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*. Second edition. Springer.
- Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics* 5, 595-645.
- Varberg, D.E. (1961). On equivalence of Gaussian measures. *Pacific Journal of Mathematics* 11, 751-762.
- Varberg, D.E. (1964). On Gaussian measures equivalent to Wiener measure. *Transactions of the American Mathematical Society* 113, 262-273.