Optimal bandwidth in nonparametric classification between two univariate densities

Peter Hall[1] **Kee-Hoon Kang**[2*]

ABSTRACT

We consider the problem of optimal bandwidth choice for nonparametric classification, based on kernel density estimators, where the problem of interest is distinguishing between two univariate distributions. When the densities intersect at a single point, optimal bandwidth choice depends on curvatures of the densities at that point. The problem of empirical bandwidth selection and classifying data in the tails of a distribution are also addressed.

KEY WORDS. *Bandwidth selection; Bayes risk; bootstrap; cross-validation; discrimination; classification error; error rate; kernel methods; nonparametric density estimation.*

1. Introduction

A common approach to nonparametric discrimination, based on data from training samples, is to construct nonparametric estimators of population densities and substitute them for the true densities in a theoretically optimal algorithm for minimising Bayes risk. Not only is this approach intuitively appealing and operationally straightforward, it is optimal in a minimax sense, as argued by Marron (1983). However, it is unclear how one might select a bandwidth that minimises risk. In particular, we might ask from a theoretical viewpoint what relationship exists between the sizes of bandwidth that are appropriate for pointwise density estimation and optimal classification, respectively. And even if we understand this connection, and have a theoretically optimal formula for bandwidth, how might we go about constructing empirical approximations to it?

In the present paper we show that the problem of optimal bandwidth choice for classification has unusual and intriguing aspects, which ironically are of more significance in simple problems than in complex ones. Indeed, if the distributions are multivariate or if there are more than two distributions among which to distinguish, then it is almost always the case that the optimal size of bandwidth (in the sense of minimising Bayes risk) is the same as that which would be used if we were constructing pointwise density estimators. The only exceptions to this rule concern very rare cases where the angles at which different densities cross one another (in the multiple-distribution case) are all identical, or where the "valley" formed by two intersecting densities (in a multivariate setting) has a constant curvatures on either side of its floor.

[1]Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia.

[2]Department of Statistics, Hankuk University of Foreign Studies, Yongin, Kyungki-Do 449-791, Korea.

By way of contrast, the relatively simple case where only two populations are involved, where the populations are univariate, and where the densities intersect at a single point, produces the following dichotomous result. If the curvatures of the two densities are of different sign at the crossing point then minimum Bayes risk is achieved using bandwidths that are of the same sizes as those which minimise pointwise estimation error. On the other hand, if the curvature signs are identical then quite different bandwidth sizes, in fact similar to those that would be employed if the kernel was of fourth (rather than second) order, are appropriate.

As these properties suggest, optimal bandwidth selection is driven by performance of the classification algorithm in places where values of two or more of the densities are similar. This comes as no surprise, since it is in those locations that discrimination is difficult, and we expect error rates to be governed by such instances. Similar issues also arise when all the densities are close to zero, and there they are more difficult to resolve. We shall pay particular attention to that case. It is not straightforward, since in the tails of a heavy-tailed distribution it is often the case that density estimators oscillate between zero and nonzero values. In particular, we may need to discriminate between two distributions when both the corresponding density estimates are zero. We suggest a method for doing this, and show that it performs well in cases where tail behaviours of the densities are not too similar.

There is an extensive literature on nonparametric methods for classification, much of it based on using an empirical version of the Bayes-optimal rule. Hall (1995) and Efron and Tibshirani (1997) discuss the performance of bootstrap-based estimators of error rate in general classification methods; Hall and Wand (1988) suggest a method for bandwidth choice when using discrimination based on kernel density estimators; Ancukiewicz (1998) introduces class-based classification rules founded on nonparametric density estimators; and Mammen and Tsybakov (1999) discuss nonparametric decision rules based on inference about places where densities cross.

2. Effect of bandwidth choice on Bayes risk

Suppose there are two populations, with distributions F and G and respective densities f and g. Let $0 < p < 1$ reflect the prior probability that a new, unclassified datum, x say, lying in a given interval $\mathcal{I}$, is drawn from F. (To avoid degeneracy we assume throughout that $0 < p < 1$.) Denote by $\mathcal{A}_0$ the "ideal" algorithm that classifies x as coming from F (or G) according as $\Delta(x) \equiv p\,f(x) - (1-p)\,g(x)$ is positive (or negative), respectively. (We may make the classification arbitrarily if $\Delta(x)$ vanishes.) Among all measurable algorithms $\mathcal{A}$ for classification on $\mathcal{I}$, $\mathcal{A}_0$ is optimal in the sense of minimising the Bayes risk:

$$\begin{aligned} \mathrm{err}_{\mathcal{A}}(f,g\,|\,\mathcal{I}) \;=\; & p \int_{\mathcal{I}} P\big(x \text{ is classified by } \mathcal{A} \text{ as coming from } g\big)\; f(x)\,dx \\ & +(1-p) \int_{\mathcal{I}} P\big(x \text{ is classified by } \mathcal{A} \text{ as coming from } f\big)\; g(x)\,dx\,. \end{aligned}$$

Given training datasets $\mathcal{X} = \{X_1, \ldots, X_m\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ drawn from F and G,

respectively, an empirical version of $\mathcal{A}_0$ may be based on nonparametric density estimators, $\hat{f}$ and $\hat{g}$ say, computed from $\mathcal{X}$ and $\mathcal{Y}$. Specifically, given a nonnegative kernel K and bandwidths $h_1, h_2 > 0$, let

$$\hat{f}(x) = \frac{1}{mh_1} \sum_{i=1}^{m} K\Big(\frac{x - X_i}{h_1}\Big), \quad \hat{g}(x) = \frac{1}{nh_2} \sum_{i=1}^{n} K\Big(\frac{x - Y_i}{h_2}\Big), \tag{1}$$

and let $\mathcal{A}_1$ be the rule that classifies x as coming from F (or G) according as $\widehat{\Delta}(x) \equiv p\,\hat{f}(x) - (1-p)\,\hat{g}(x)$ is positive (or negative), respectively.

If x is a point at which $p\,f(x)$ and $(1-p)\,g(x)$ differ by a substantial amount, then empirical methods will generally be able to determine whether $\Delta(x)$ is positive or negative. In such places the empirical rule $\mathcal{A}_1$ will give results almost identical to those obtained using $\mathcal{A}_0$. Employing arguments such as this one can deduce that overall performance of the empirical rule, and in particular, choice of bandwidths h_1 and h_2 that give good performance, is determined by properties of f and g in neighbourhoods of points y where discrimination is particularly difficult, i.e. where $\Delta(y) = 0$.

The range of different modes of behaviour can be very broad, but fortunately many of them would seldom arise in practice. To describe the theory underlying optimality properties, let us assume for simplicity that the sample sizes m and n are asymptotically similar; formally,

$$m/n \text{ is bounded away from zero and infinity as } n \to \infty\,.$$

Then, provided there is just one point y in $\mathcal{I}$ at which the graphs of $p\,f$ and $(1-p)\,g$ cross, the optimal bandwidth, in the sense of minimising $\mathrm{err}_{\mathcal{A}_1}(f, g\,|\,\mathcal{I})$, is generally of size either $n^{-1/5}$ or $n^{-1/9}$. The first case arises when the curvatures of f and g are of opposite signs at y, and the second when the curvatures have the same signs.

Theorem 2.1 describes the size of the additional classification error, over and above that for the optimal algorithm $\mathcal{A}_0$, that is incurred through using the empirical algorithm $\mathcal{A}_1$ rather than $\mathcal{A}_0$. Put $h = n^{-\rho}$, where $0 < \rho < 1$.

Theorem 2.1. *Assume $0 < p < 1$ and $\mathcal{I}$ is a compact interval, and that Δ vanishes at just $\nu \geq 1$ points, $y_1, \ldots, y_\nu$, in $\mathcal{I}$, all of them interior points and at each of which $\Delta'(y_j) \neq 0$. In addition, some regularity conditions on f and g are assumed. Then,*

$$\mathrm{err}_{\mathcal{A}_1}(f, g\,|\,\mathcal{I}) - \mathrm{err}_{\mathcal{A}_0}(f, g\,|\,\mathcal{I}) = \tfrac{1}{2} \sum_{j=1}^{\nu} |\Delta'(y_j)|^{-1}\, E\big\{p\,\hat{f}(y_j) - (1-p)\,\hat{g}(y_j)\big\}^2 + o\big\{(nh)^{-1} + h^4\big\} \tag{2}$$

If in addition $\nu = 1$, $f''(y_1)\,g''(y_1) > 0$,

$$\frac{h_2}{h_1} = \left\{\frac{p\,f''(y_1)}{(1-p)\,g''(y_1)}\right\}^{1/2} + o(h^2)$$

and f and g each have four continuous derivatives in a neighbourhood of y_1, then (2) continues to hold if the remainder there is replaced by $o\{(nh)^{-1} + h^8\}$.

In many circumstances the discussion of classification given following Theorem 2.1 applies in a general, global sense, to an empirical algorithm $\widehat{\mathcal{A}}$ applied to any new datum $x \in \mathbb{R}$, rather than only to the algorithm $\mathcal{A}_1$ restricted to $\mathcal{I}$. But, there are methodological as well as technical difficulties in using nonparametric density estimators for classification in the tails of distributions. In particular, if both $\hat{f}$ and $\hat{g}$ vanish at a point x, how do we classify a new datum that takes the value x? One approach, which we explore here, is to implement classification using a global bandwidth, but employ a new algorithm for classifying data x for which $\hat{f}(x) = \hat{g}(x) = 0$. Instances where $\hat{f}(x)$ and $\hat{g}(x)$ take equal but nonzero values can be resolved by classifying "at random", using the prior probabilities.

Let us assume for definiteness that the supports of both f and g are intervals, that neither density vanishes in the interior of its support, and that a discrimination rule is sought in the upper tail. In this instance our algorithm will be based on the assumption that, sufficiently far to the right, the tail of f exceeds that of g, or vice versa. Formally, we ask that either $f(x) > g(x)$ for all $x \in (x_0, x_{\text{supp}})$, or $g(x) > f(x)$ for all $x \in (x_0, x_{\text{supp}})$, where x_0 is strictly less than the right-hand end, x_{supp}, of the support of f or g, respectively; and we seek a means of classifying new data $x > x_0$. Of course, x_{supp} may be infinite.

If $x > x_0$ and $\hat{f}(x) = \hat{g}(x) = 0$, let $\hat{x}$ denote the infimum of values of $y \le x$ such that $\hat{f}(z) = \hat{g}(z) = 0$ for all $z \in [y, x]$. Our algorithm, to which we refer below as $\mathcal{A}_{\text{R}}$ where the subscript indicates the right-hand tail, consists of classifying x as coming from f or g according as $\hat{f}(\hat{x}-) > 0$ or $\hat{g}(\hat{x}-) > 0$. (With probability 1, exactly one of $\hat{f}(\hat{x}-)$ and $\hat{g}(\hat{x}-)$ will be nonzero.) A definition of $\hat{x}$ which gives asymptotically equivalent results, is that $\hat{x}$ is the nearest value to x such that at least one of the four values of $\hat{f}(\hat{x}\pm)$ and $\hat{g}(\hat{x}\pm)$ is strictly positive.

3. Empirical choice of bandwidth and numerical properties

An effective approach to get estimates of bandwidths, which we shall consider in detail, is based on using the bootstrap to estimate $\text{err}_{\mathcal{A}_1}(f, g \,|\, \mathcal{I})$ and thereby to select the optimal bandwidths. Specifically, let $\tilde{f}$ and $\tilde{g}$ be the versions of $\hat{f}$ and $\hat{g}$, defined at (1), that arise if we use respective bandwidths h_3 and h_4 (instead of h_1 and h_2). Conditional on $\mathcal{X}$ (or on $\mathcal{Y}$), draw m data $\mathcal{X}^* = \{X_1^*, \ldots, X_m^*\}$ independently and uniformly from the distribution with density $\tilde{f}$ (or, respectively, n data $\mathcal{Y}^* = \{Y_1^*, \ldots, Y_n^*\}$ independently and uniformly from the distribution with density $\tilde{g}$), and let

$$\hat{f}^*(x) = \frac{1}{mh_1} \sum_{j=1}^{m} K\Big(\frac{x - X_j^*}{h_1}\Big), \quad \hat{g}^*(x) = \frac{1}{nh_2} \sum_{j=1}^{n} K\Big(\frac{x - Y_j^*}{h_2}\Big).$$

Put $\widehat{\Delta}^*(x) = p\,\hat{f}^*(x) - (1-p)\,\hat{g}^*(x)$ and

$$\widehat{\text{err}}_{\mathcal{A}_1}(h_1, h_2) = p \int P\{\widehat{\Delta}^*(x) \;<\; 0 \,|\, \mathcal{X} \cup \mathcal{Y}\}\, \tilde{f}(x)\, dx + (1-p) \int P\{\widehat{\Delta}^*(x) > 0 \,|\, \mathcal{X} \cup \mathcal{Y}\}\, \tilde{g}(x)\, dx.$$

Then, choose $(h_1, h_2) = (\hat{h}_1, \hat{h}_2)$ to minimise $\widehat{\text{err}}_{A_1}(h_1, h_2)$.

We have done a simulation study addressing properties of the empirical bandwidth selector introduced in the previous paragraph. Recall from section 2 that there are two main classes of problem, respectively characterised by the property that the densities, f and g, intersect at a point where the curvatures have different signs or the same sign. Call these classes 1 and 2; they correspond to the optimal bandwidth being of size $n^{-1/5}$ or $n^{-1/9}$, respectively. We examine for two examples in each class. Throughout, the distribution with density f was standard normal, $p = \frac{1}{2}$ and $m = n$. In the tails of the distributions, in any cases of ambiguity we classified using the method suggested in section 2.

From some plots of the simulation results, we can say practical aspects fit the theory. The agreement between theory and numerical simulation is somewhat better in the case of class 1, but in the second class the numerical results also clearly reflect the theory.

REFERENCES

ANCUKIEWICZ, M. (1998). An unsupervised and nonparametric classification procedure based on mixtures with known weights. *J. Classification* **15**, 129–141.

EFRON, B. AND TIBSHIRANI, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *J. Amer. Statist. Assoc.* **92**, 548–560.

HALL, P. AND WAND, M.P. (1988). *On nonparametric discrimination using density differences. Biometrika* **75**, 541–547.

HALL, P. (1995). On the biases of error estimators in prediction problems. *Statist. Probab. Lett.* **24**, 257–262.

MAMMEN, E. AND TSYBAKOV, A.B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27**, 1808–1829.

MARRON, J.S. (1983). Optimal rates on convergence to Bayes risk in nonparametric discrimination. *Ann. Statist.* **11**, 1142–1155.