J. R. Statist. Soc. B (2005)
67, *Part* 3, *pp.* 363–379

Properties of bagged nearest neighbour classifiers

Peter Hall

Australian National University, Canberra, Australia

and Richard J. Samworth

Australian National University, Canberra, Australia, and University of Cambridge, UK

[Received July 2003. Revised October 2004]

Summary. It is shown that bagging, a computationally intensive method, asymptotically improves the performance of nearest neighbour classifiers provided that the resample size is less than 69% of the actual sample size, in the case of with-replacement bagging, or less than 50% of the sample size, for without-replacement bagging. However, for larger sampling fractions there is no asymptotic difference between the risk of the regular nearest neighbour classifier and its bagged version. In particular, neither achieves the large sample performance of the Bayes classifier. In contrast, when the sampling fractions converge to 0, but the resample sizes diverge to ∞, the bagged classifier converges to the optimal Bayes rule and its risk converges to the risk of the latter. These results are most readily seen when the two populations have well-defined densities, but they may also be derived in other cases, where densities exist in only a relative sense. Cross-validation can be used effectively to choose the sampling fraction. Numerical calculation is used to illustrate these theoretical properties.

Keywords: Bayes risk; Bootstrap; Classification error; Cross-validation; Density; Discrimination; Error rate; Marked point process; Poisson process; Prediction; Regret; Statistical learning; Without-replacement sampling; With-replacement sampling

1. Introduction

Bagging, or bootstrap aggregation, was introduced by Breiman (1996, 1999) as a means for improving the performance of a 'predictor', e.g. a classifier, by combining the results of many empirically simulated predictions. Bagging a conventional classifier, in particular one based on nearest neighbours, can sometimes, although not always, reduce the error rate. See, for example, Bay (1998), whose 'multiple-feature subsets' approach is another method for improving nearest neighbour classifiers. Other recent contributions to bagging and to related methodology include those of Ho (1998a, b), Skurichina and Duin (1998), Bay (1999), Guerra-Salcedo and Whitley (1999), Zemke (1999), Francois *et al*. (2001), Kuncheva *et al*. (2002) and Skurichina *et al*. (2002).

Nearest neighbour methods are one of the oldest approaches to classification, dating from work of Fix and Hodges (1951). Nevertheless, they are constantly being adapted to new settings, e.g. by replacing 'prototypes' by 'representatives' to produce new criteria for classification (see for example Kuncheva and Bezdek (1998) and Mollineda *et al*. (2000)). A major attraction of nearest neighbour classifiers is their simplicity. For implementation they require only a measure of distance in the sample space, along with samples of training data; hence their popularity as a starting-point for refinement and improvement.

Address for correspondence: Richard J. Samworth, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, UK.
E-mail: r.j.samworth@statslab.cam.ac.uk

 1369–7412/05/67363

In this paper we show that, in circumstances where the relative densities of two populations can be meaningfully defined, a bagged nearest neighbour classifier can converge, as the training sample sizes increase, to the optimal Bayes classifier. However, this limit obtains if and only if the simulated training samples are of asymptotically negligible size relative to the respective actual training samples, and if the simulated training sample sizes diverge together with the actual training sample sizes. If, as is common in practice, the simulated training sample sizes are the same as those of the actual training samples, then the nearest neighbour classifier does not converge to the Bayes classifier. These results apply to both with- and without-replacement bagging, which are the two approaches that are most commonly used in practice.

The extent of the improvements can be determined by probability calculations, which we discuss theoretically and illustrate numerically. It is shown, for example, that in the case of with-replacement bagging the resample size should be at most 69% of the sample size if the bagged nearest neighbour classifier is to improve on the performance of its unbagged counterpart asymptotically. The ceiling is reduced to 50% in the case of without-replacement bagging. These results are of interest because the majority of bagging experiments employ relatively large resample sizes; much of the evidence against the performance of bagged nearest neighbour classifiers (e.g. Breiman (1996) and Bay (1998)) is for full size resamples.

An alternative way of enhancing the performance of nearest neighbour classifiers is to use k-nearest-neighbour methods. If k is permitted to increase with n, in such a manner that $k \to \infty$ but $k/n \to 0$, then k-nearest-neighbour methods, like bagged nearest neighbour techniques, asymptotically achieve the performance of the Bayes classifier. The fact that bagging is currently under such wide study and development, with relatively little theoretical underpinning, motivates the work in the present paper. We are not attempting to promote bagging against, for example, k-nearest-neighbour techniques, only to describe its properties.

Nevertheless, if attention is confined to classification problems for Euclidean data, rather than the more general setting that is treated in the present paper, then it is possible to develop higher order theory describing the relative performance of bagging and nearest neighbour methods. There it can be shown that, if a Poisson model for the sample size is assumed, and for the optimal choice of tuning parameters in either case, the bagged classifier is inferior to its nearest neighbour counterpart in one dimension. However, the two are asymptotically equivalent in two dimensions, and bagging is superior in higher dimensions. The theoretical arguments are particularly complex, however, and so will not be given here.

Bagging one-nearest-neighbour techniques is more computer intensive than implementing k-nearest-neighbour methods for general k. We could bag k-nearest-neighbour classifiers, and let k increase with n, although there seems little motivation for doing this. Moreover, an advantage of bagging is that, even in cases where we do not attempt to achieve optimality, performance can be described quite simply, in absolute terms, via the ratio of the resample size to sample size; see, for example, Section 3.1.

Bühlmann and Yu (2002) discussed the performance of bagging from the viewpoint of its ability to reduce instability, and Buja and Stuetzle (2000a, b) and Friedman and Hall (2000) explored it in terms of its success in accommodating difficulties that are caused by non-linear features of a statistical method. The present paper takes up a different angle, addressing the way that performance depends on the resample size.

In some respects the problem is a little like bootstrap inference in settings where the statistic of interest is not asymptotically normally distributed. There, the bootstrap gives consistency only if the resample size is an order of magnitude smaller than the sample size; see, for example, Mammen (1992) and Bickel *et al*. (1997). In each case it is necessary to select the resample size.

In the classification setting the leave-one-out method, or cross-validation, can be used to minimize risk empirically. We point out that this approach will lead to an asymptotically optimal choice of resample sizes, in the sense of minimizing large sample risk for either with- or without-replacement sampling, but that cross-validation usually will not produce asymptotic minimization of regret, i.e. of the distance between the actual risk and that of the Bayes classifier.

In most statistical work, the performance of classifiers is discussed in settings where population densities are well defined and estimable. However, to construct the Bayes classifier only the relative densities are needed, weighted by the prior probabilities. Indeed, the optimal Bayes classifier depends only on whether the weighted density ratio is greater than 1 or less than 1, not on its exact value. We show that our result about convergence of the bagged nearest neighbour classifier to the Bayes rule can be set up in this context, so that it is relevant to a relatively general class of problems.

There is an extensive literature on nonparametric methods for classification, including techniques which converge to the Bayes rule. An example in the univariate case is Stoller's (1954) approach, which is based on the empirical distribution function. In multivariate settings, methods founded on nonparametric density estimation are sometimes used; they have been discussed by Hand (1981), have been shown by Marron (1983) to enjoy optimality properties and are addressed in most recent monographs on nonparametric function estimation. Optimality theory for nonparametric classifiers has been explored by many researchers; we mention only the work of Lugosi and Nobel (1996), who proved the existence of a class of universally consistent classifiers, and of Devroye (1982), Devroye *et al.* (1996), chapter 7, and Yang (1999), who showed that optimal convergence rates can nevertheless be arbitrarily slow. Cover (1968), Fukunaga and Hummels (1987) and Psaltis *et al.* (1994) have shown that, in d-variate settings, the risk of nearest neighbour classifiers converges to its limit at rate $n^{-2/d}$. However, the limit here generally exceeds the risk of the Bayes classifier. Efron (1983) and Efron and Tibshirani (1997) have discussed the performance of bootstrap-based estimators of risk for general classification methods, and Steele and Patterson (2000) have shown how to make exact calculations of bootstrap estimators of the expected prediction error for nearest neighbour classifiers. Recent statistical work on classification in very high dimensional settings includes that of Breiman (2001), Schapire *et al.* (1998), Friedman *et al.* (2000), Kim and Loh (2001), Dudoit *et al.* (2002) and Jiang (2002). We mention also the method of boosting (e.g. Freund and Schapire (1997)) and the existence of a wide range of techniques that are based on combining the results of different classifiers (e.g. Larkey and Croft (1996)).

2. Definitions of classifiers, and basic properties

2.1. Nearest neighbour, bagged and Bayes classifiers

Assume that there are two populations, Π_X and Π_Y, from which we have random samples $\mathcal{X}$ and $\mathcal{Y}$. Suppose that $\mathcal{X}$ is of size m and $\mathcal{Y}$ of size n. A nearest neighbour classifier, based on $\mathcal{X}$ and $\mathcal{Y}$, assigns a new datum z to Π_X or Π_Y according to whether z is nearest to an element of $\mathcal{X}$ or $\mathcal{Y}$ respectively. Now draw resamples $\mathcal{X}^*$ and $\mathcal{Y}^*$, of sizes $m_1 \leqslant m$ and $n_1 \leqslant n$, by resampling randomly, with or without replacement, from $\mathcal{X}$ and $\mathcal{Y}$ respectively. The bagged version of the nearest neighbour classifier allocates z to Π_X if the nearest neighbour classifier, based on $\mathcal{X}^*$ and $\mathcal{Y}^*$ rather than $\mathcal{X}$ and $\mathcal{Y}$, assigns it more often to Π_X than to Π_Y. For definiteness we shall always treat the version of the bagged nearest neighbour classifier which uses an infinite number of simulations in the 'majority vote' step.

Of course, nearest neighbour classification requires a measure of distance, which is generally supplied through a metric, or norm. Therefore the data in $\mathcal{X}$ and $\mathcal{Y}$ will be assumed to come

from a set on which a norm, $\|\cdot\|$, is defined; in mathematical terminology this set would be a Banach space and will be denoted by $\mathcal{B}$.

If well-defined densities, f and g say, exist for the populations Π_X and Π_Y respectively, and if the prior probabilities of these populations are p and $1-p$, then the 'ideal' Bayes rule classifier assigns z to Π_X or Π_Y according to whether $p\,f(z)-(1-p)\,g(z)$ is positive or negative. Equivalently, z is assigned to Π_X if the probability

$$q(z)=\frac{p\,f(z)}{p\,f(z)+(1-p)\,g(z)} \tag{2.1}$$

exceeds $\frac{1}{2}$, and to Π_Y if $q(z)<\frac{1}{2}$.

2.2. Error rates of Bayes and nearest neighbour classifiers

In cases where the population densities f and g exist, we define the average error rate, or risk, for a general classification rule as

$$p\int P(z \text{ is classified as coming from } \Pi_Y)\; f(z)\,\mathrm{d}z + (1-p)\int P(z \text{ is classified as coming from } \Pi_X)\; g(z)\,\mathrm{d}z. \tag{2.2}$$

Thus equal losses for each of the two types of error are assumed. The risk for the Bayes classifier is therefore

$$\mathrm{err}_{\mathrm{Bayes}}=\int \min\{p\,f(z),(1-p)\,g(z)\}\,\mathrm{d}z. \tag{2.3}$$

Still assuming the existence of f and g, the large sample limit of risk for the nearest neighbour classifier can be deduced from a standard point process approximation, which is given in Appendix A.1. This enables a range of properties of classifiers to be derived, including the theorem below, which were discussed by Cover and Hart (1967). For simplicity we shall state the theorem in the case of data from $\mathbb{R}^d$, although it has analogues in other settings, e.g. in functional data contexts such as that of theorem 3.

Theorem 1. Assume that the densities f and g are continuous and that m and n increase together in such a manner that $m/(m+n)\to p\in(0,1)$. Then, as $m,n\to\infty$, the risk of the nearest neighbour classifier converges to

$$\mathrm{err}_{\mathrm{NN}}=p\int\{1-q(z)\}\,f(z)\,\mathrm{d}z+(1-p)\int q(z)\,g(z)\,\mathrm{d}z.$$

Moreover, $\mathrm{err}_{\mathrm{NN}}\geqslant\mathrm{err}_{\mathrm{Bayes}}$.

3. Bagged nearest neighbour classifiers

3.1. Main results

We shall show that bagging the nearest neighbour classifier with relatively small resample sizes m_1 and n_1, satisfying $m_1/m\to 0$ and $n_1/n\to 0$ as $n\to\infty$, produces a classifier for which the risk converges to $\mathrm{err}_{\mathrm{Bayes}}$, which is given at equation (2.3). Our argument also demonstrates that this result holds for both with-replacement and without-replacement bagging. Furthermore, we shall show that the result fails, for both types of bagging, if m_1/m and n_1/n converge to non-zero limits; and we shall identify the asymptotic risk in the latter case, for both types of bagging.

Again, for simplicity, we shall initially assume that the data are d variate, although our main results are valid in cases where only 'relative densities' are available, and individual densities may not be well defined. See Section 4.

Theorem 2. Assume that m_1 and n_1 diverge, but that m_1/m converges to a limit, $l \geqslant 0$ say, as $n \to \infty$, and the ratio of the resample size to sample size is asymptotically the same for Π_X and Π_Y data, i.e.

$$m_1/m - n_1/n \to 0. \tag{3.1}$$

Define $\rho = \exp(-l)$ or $\rho = 1 - l$ in the cases of with-replacement bagging and without-replacement bagging respectively; thus, $0 \leqslant \rho \leqslant 1$. Suppose also that f and g are continuous and that $m/(m+n) \to p \in (0, 1)$. Then, the probability that a new data value z is identified by the bagged nearest neighbour classifier as coming from Π_X converges to

$$P(\rho, z) = P\left\{ \sum_{j=1}^{\infty} \rho^{j-1}(1-\rho) J_j > \tfrac{1}{2} \right\} \tag{3.2}$$

as $n \to \infty$, where $J_1, J_2, \ldots$ are independent and identically distributed 0–1 random variables with $P(J_j = 1) = q(z)$, the latter defined at equation (2.1). Furthermore, the risk of the bagged nearest neighbour classifier converges to

$$\mathrm{err}_{\mathrm{bagg}} = p \int \{1 - P(\rho, z)\}\, f(z)\, \mathrm{d}z + (1-p) \int P(\rho, z)\, g(z)\, \mathrm{d}z. \tag{3.3}$$

We shall show in Section 6.2 that ρ can be interpreted as the probability of thinning a two-type Poisson process, where the two types are X and Y. Assumption (3.1) implies that the probability ρ is identical for each of the two types, which greatly simplifies discussion. In without-replacement bagging, to obtain non-degenerate results we require $0 \leqslant l < 1$. In contrast, $l = 1$ gives non-degenerate results in the setting of with-replacement bagging.

A proof of theorem 2 is given in Appendix A.3, using properties that are developed in Appendices A.1 and A.2. It is tacitly assumed in theorem 2 that either with-replacement bagging or without-replacement bagging is used throughout; we do not, for example, use with replacement for one type of data and without replacement for another. Furthermore, the value $\rho = 1$ is permitted in theorem 1; it arises when the sampling ratios m_1/m and n_1/n both converge to 0.

The value of $P(\rho, z)$ when $\rho = 1$ is defined by taking the limit in equation (3.2): as $\rho \uparrow 1$, and for $q(z) \neq \frac{1}{2}$,

$$\begin{aligned} P(\rho, z) &= I\left\{ \sum_{j=1}^{\infty} \rho^{j-1}(1-\rho)\, E(J_j) > \tfrac{1}{2} \right\} + o(1) = I\{q(z) > \tfrac{1}{2}\} + o(1) \\ &\to \begin{cases} 1 & \text{if } q(z) > \frac{1}{2}, \\ 0 & \text{if } q(z) < \frac{1}{2}, \end{cases} \end{aligned} \tag{3.4}$$

where $I(\mathcal{E})$ denotes the indicator function of an event $\mathcal{E}$. Therefore, we take $P(1, z) = I\{q(z) > \frac{1}{2}\}$ if $q(z) \neq \frac{1}{2}$. To appreciate why the approximations in expression (3.4) are valid, it suffices to note that, for all ρ, the expected value of the infinite series within the probability on the right-hand side of equation (3.2) equals $q(z)$, and that, as $\rho \uparrow 1$, the variance of that series converges to 0.

For the bagged nearest neighbour classifier to converge to the Bayes classifier, it is necessary and sufficient that the probability at equation (3.2) equals 1 if $q(z) > \frac{1}{2}$ and equals 0 if $q(z) < \frac{1}{2}$; see Section 2.1. It is easy to see from equation (3.2), and from the argument in the previous paragraph, that this property holds if and only if $\rho = 1$, i.e. if and only if m_1/m and n_1/n both

converge to 0. Provided that this constraint holds, the risk of the bagged nearest neighbour classifier converges to $\mathrm{err}_{\mathrm{Bayes}}$, which is defined at equation (2.3). This is also the limit, as $\rho \uparrow 1$, of the risk at equation (3.3).

Cases where $m/(m+n)$ does not converge to p can be handled by choosing the sampling fractions to represent the prior probabilities. In particular, one should select m_1 and n_1 so that $m_1/n_1 \to p/(1-p)$. However, to obtain the Bayes classifier as the limit of the nearest neighbour rule one should still ensure that m_1 and n_1 both diverge, and m_1/m and n_1/n both converge to 0.

Several properties can be deduced from expressions (3.2) and (3.4). For example, if $0 \leqslant \rho \leqslant \frac{1}{2}$, then, by equation (3.2), $P(\rho, z) = P(J_1 = 1) = q(z)$. It follows that the asymptotic limit of risk for the bagged nearest neighbour classifier will be the same as that for the regular nearest neighbour classifier if $0 \leqslant \rho \leqslant \frac{1}{2}$ and will generally be reduced if $\rho > \frac{1}{2}$. Since, in the cases of with- and without-replacement bagging, the respective values of ρ are the limits of $\exp(-m_1/m)$ and $1 - m_1/m$, then bagging the nearest neighbour classifier will asymptotically improve performance if $m_1 < m \log(2) \approx 0.69m$ in the with-replacement case, and if $m_1 < \frac{1}{2} m$ in the without-replacement setting, but not otherwise. Therefore, reducing the sampling fraction m_1/m does not immediately lead to a reduction in risk; it must be reduced below the threshold, 0.69 or 0.5, in the cases of with- or without-replacement bagging respectively.

3.2. Numerical values of error

Recall that $P(\rho, z)$ depends on z only through the quantity $q(z)$ that is defined at equation (2.1). In a slight abuse of notation, and in this section only, we shall write $P(\rho, q)$ for the value of $P(\rho, z)$ when $q(z) = q$, for any $q \in [0, 1]$. In this notation, Fig. 1(a) shows $P(\rho, q)$ as a function of q, for $\rho = 0.5, 0.7, 0.9$. The last two curves were obtained by simulation and show the convergence of the large sample approximation of the bagged nearest neighbour classifier to the Bayes classifier that was defined in the last paragraph of Section 2.1. Fig. 1(b) gives the complementary plot of $P(\rho, q)$ as a function of ρ, for $q = \frac{1}{8}, \frac{1}{4}, \frac{3}{8}, \ldots, \frac{7}{8}$.

To demonstrate the asymptotic improvement in risk of the bagged nearest neighbour classifier over its nearest neighbour counterpart, we study two examples. In the first, we choose $f = \phi$, the standard normal density, and set $g(z) = \phi(z - \mu)$, for $\mu \in [0, 3]$. In Fig. 2(a) we plot $\mathrm{err}_{\mathrm{bagg}}$, given by equation (3.3), as a function of μ, for $\rho = 0.5, 0.7, 0.9, 1$. Throughout, where

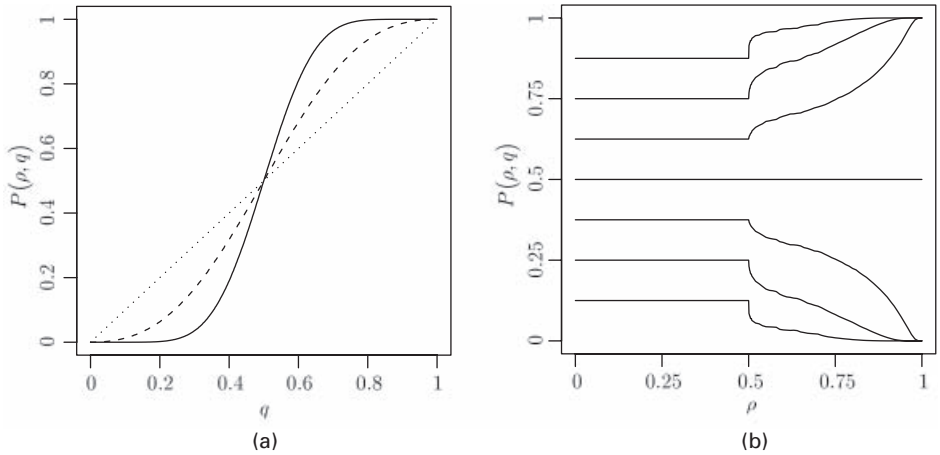

Fig. 1. Plots of $P(\rho, q)$ (·····, $\rho = 0.5$; -------, $\rho = 0.7$; ———, $\rho = 0.9$): (a) $P(\rho, q)$ as a function of q for fixed ρ; (b) $P(\rho, q)$ as a function of ρ for fixed q, the latter ranging from $\frac{1}{8}$ (bottom line) to $\frac{7}{8}$ (top line) in steps of $\frac{1}{8}$

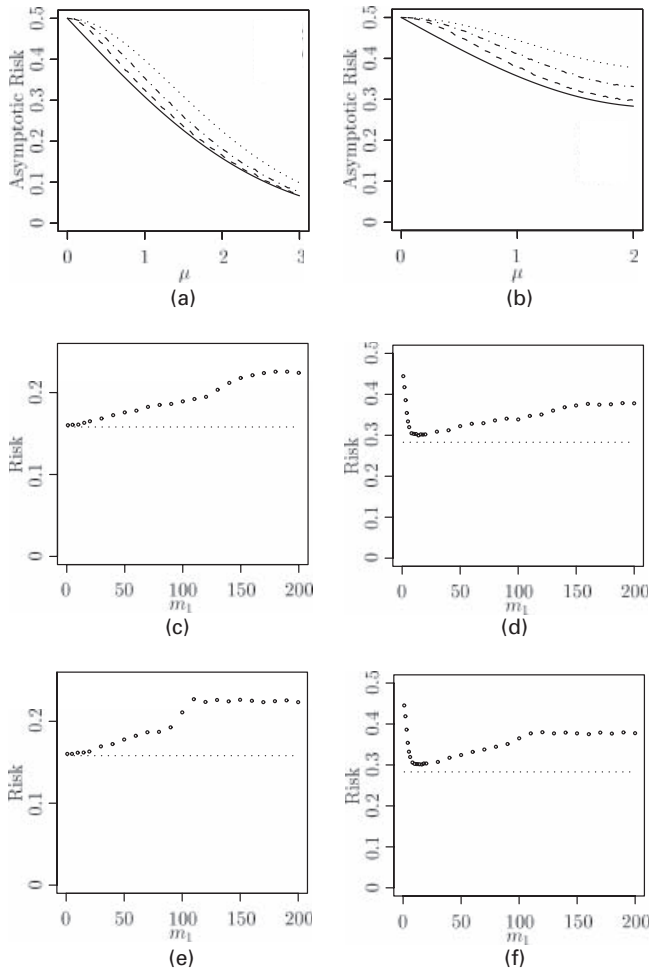

Fig. 2. Plots of err_{bagg} for the bagged nearest neighbour classifier when $m = n = 200$: (a), (b) err_{bagg} as a function of μ, the difference between the means of the distributions with respective densities f and $g = f(\cdot - \mu)$: in (a) $f = \phi$, i.e. the standard normal density, and in (b) it is the density at equation (3.5) (·····, $\rho = 0.5$; - · - · - ·, $\rho = 0.7$; - - - - - -, $\rho = 0.9$; ———, $\rho = 1$); (c), (d), (e), (f) err_{bagg} as a function of m_1 ($= n_1$), for various values of m_1 (in (c) and (e), the densities are the same as in (a), but with $\mu = 2$; likewise, in (d) and (f), the densities are the same as in (b), but again with $\mu = 2$; resampling was done with replacement in the case of (c) and (d), and without replacement for (d) and (f); ·····, Bayes risks)

a finite number of bagged samples needed to be drawn, we took $B = 199$. Recall that the cases $\rho = 0.5$ and $\rho = 1$ correspond to err_{NN} and $\text{err}_{\text{Bayes}}$ respectively. In this problem, the graphs of the functions $y = f(z)$ and $y = g(z)$ cross at only one point, so we expect classification to be relatively straightforward, provided that μ is not too small.

In the second of these two examples we choose f to have the mixture density

$$f(z) = \tfrac{1}{5} \sum_{i=0}^{4} \phi(z - 4i), \tag{3.5}$$

and we set $g(z) = f(z - \mu)$ for $\mu \in [0, 2]$. For these versions of f and g, Fig. 2(b) again plots err_{bagg} as a function of μ, for the same values of ρ. Here the densities cross at nine points, and classification is therefore much more difficult.

In both examples, we observe that the bagged nearest neighbour classifier can provide considerable asymptotic improvement in risk over the nearest neighbour classifier. To see whether these gains are manifested in practice, we studied both the examples above with $\mu = 2$ and simulated training samples of size $m = n = 200$. The risks of the bagged nearest neighbour classifier as functions of m_1 are given in Figs 2(c), 2(d), 2(e) and 2(f). Figs 2(c) and 2(e) refer to the first example, whereas Figs 2(d) and 2(f) refer to the second. Resampling was done with replacement in Figs 2(c) and 2(d) and without replacement in Figs 2(e) and 2(f) to show that the risk becomes flat at different points, specifically at $m_1 \approx 0.7m$ and $m_1 \approx 0.5m$ respectively. The behaviour for relatively small m_1 is virtually identical for either resampling scheme. Two further points are worthy of note:

(a) in all four cases, the optimal choice of m_1 is much smaller than m;
(b) whereas the optimal choice appears to be $m_1 = 1$ in the first example under both types of resampling, this is a poor choice in the more complicated second example. There, $m_1 = 14$ and $m_1 = 16$ are optimal for with- and without-replacement sampling respectively.

The next two examples concern functional data sets, one real and one simulated. The real data are the mean monthly temperatures recorded at 27 different weather-stations, averaged over the years 1960–1994. The stations are classified according to their geographical climate, there being $m = 13$ continental stations and $n = 14$ Atlantic stations. These data are represented by their first 13 Fourier coefficients, yielding the curves in Fig. 3(a). The square of the distance between two curves is taken to be the sum of the squares of the differences between these Fourier coefficients. Nearest neighbour and bagged nearest neighbour classification was performed by leaving out one curve at a time, and attempting to classify the missing curve by using the remaining data. Taking $p = 13/27$, the risk of the nearest neighbour classifier was 0.185. Using with-replacement resampling, and $n_1 = m_1$, the risk of the bagged nearest neighbour classifier was 0.111 for $m_1 \leqslant 8$, 0.148 for $m_1 = 9$ and 0.185 for $m_1 \geqslant 10$.

Simulated functional data were obtained by generating $m = 100$ and $n = 100$ temperature curves, the joint distributions of the temperatures over the 12 months being Gaussian and having the same mean and covariance as those of the continental and Atlantic weather data

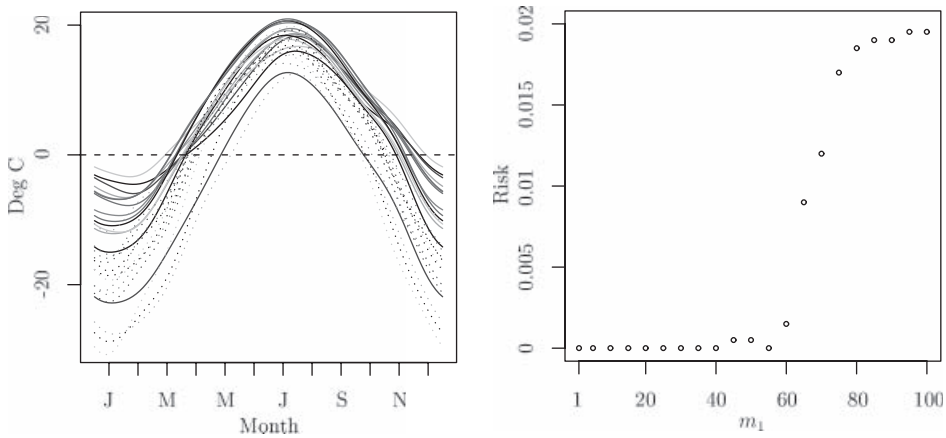

Fig. 3. Weather data plots and bagged nearest neighbour classifier risks for simulated weather data: (a) temperature curves generated from the monthly records at 27 weather-stations (·····, continental stations; ———, Atlantic stations); (b) risks of the bagged nearest neighbour classifier at various resample sizes for the simulated weather data, for with-replacement resampling (the sample sizes were $m = n = 100$)

respectively. The data were again represented by their first 13 Fourier coefficients. The reason for choosing such large sample sizes is that, since virtually any classifier can be expected to perform very well in such cases, the effects of choosing different resample sizes might be expected to be non-observable for these values of m and n. Risks of the nearest neighbour and bagged nearest neighbour classifiers were computed over 1000 test data and are given in Fig. 3(b). They are indeed quite low, but the benefits that are gained by bagging with a relatively small resample size are nevertheless clear.

4. Surrogates for densities of *X* and *Y*

In general circumstances, extending beyond Euclidean data models, the bagged nearest neighbour classifier will tend to assign a new data value z to X or Y depending on which of these two distributions has 'greater likelihood' in the vicinity of z. Following convention, in Section 2 we argued that the distribution of X has greater likelihood if $p\,f(z) > (1-p)\,g(z)$, where f and g denote densities of X and Y, and p and $1-p$ are the respective prior probabilities. However, standard definitions of f and g, requiring differentiable distribution functions, are not meaningful in some cases, e.g. when X and Y are random functions. Significant progress can nevertheless be made under less restrictive assumptions, involving little more than smoothness, as a function of ball radius, of the probability that X or Y lies in a ball, and properties of approximations to relative densities, rather than existence of actual densities.

To appreciate how this is done, let us assume that X and Y take values in a common sample space $\mathcal{B}$, which we take to be a Banach space equipped with a norm, $\|\cdot\|$. Our smoothness assumption on balls is

$$\begin{aligned}&\text{for both } Z = X \text{ and } Z = Y, \text{ and for each } z \in \mathcal{B}, \text{ the function } \pi_Z(\delta|z)\\ &\qquad = P(\|Z - z\| \leqslant \delta) \text{ is continuous in } \delta \in [0, \infty), \text{ with } \pi_Z(0|z) = 0.\end{aligned} \tag{4.1}$$

We define relative density in terms of ratios of probabilities that X and Y lie in balls. Specifically, given $\eta > 0$ let $\mathcal{S}_X(\eta)$ or $\mathcal{S}_Y(\eta)$ denote the sets of $z \in \mathcal{B}$ such that, for all $\delta \in (0, \eta)$,

$$\frac{p\,\pi_X(\delta|z)}{(1-p)\,\pi_Y(\delta|z)} \geqslant 1 + \eta \text{ or } \frac{(1-p)\,\pi_Y(\delta|z)}{p\,\pi_X(\delta|z)} \geqslant 1 + \eta \tag{4.2}$$

respectively. Thus, if $z \in \mathcal{S}_X(\eta)$ then we can fairly say that the distribution of X has greater density than that of Y in the neighbourhood of z, weighted by the prior probabilities p and $1-p$, without having to specify what we mean by 'the' densities of the distributions of X or Y.

It is straightforward to construct realistic, non-Euclidean, examples, for instance in function spaces, where assumption (4.1) holds and the definition of relative density at expression (4.2) is meaningful. See the next paragraph for discussion. Of course, such constructions depend on the norm, $\|\cdot\|$, that is chosen for the sample space. This dependence is to be expected; even for finite dimensional, d-variate Euclidean data, where $\mathcal{B} = \mathbb{R}^d$ and suitable norms for the distance between $(x^{(1)}, \ldots, x^{(d)})$ and $(y^{(1)}, \ldots, y^{(d)})$ include $(\Sigma_i\, |x^{(i)} - y^{(i)}|^2)^{1/2}$, $\Sigma_i\, |x^{(i)} - y^{(i)}|$ and $\max_i |x^{(i)} - y^{(i)}|$, each choice gives rise to a different definition of densities f and g, defined in each case by

$$\begin{aligned}\delta^{-d}\, P(\|X - z\| \leqslant \delta) &\to c_d\, f(z),\\ \delta^{-d}\, P(\|Y - z\| \leqslant \delta) &\to c_d\, g(z)\end{aligned}$$

as $\delta \to 0$, where c_d denotes the d-variate content of a d-dimensional sphere of unit radius. Our argument in Sections 2 and 3 would usually be interpreted in the context of the traditional

Euclidean norm, $\|x-y\| = (\Sigma_i |x^{(i)} - y^{(i)}|^2)^{1/2}$, but it can also be validly understood in the setting of the other two norms that were mentioned above.

When $\mathcal{B}$ denotes a function space, $\|\cdot\|$ might be an L_r-norm for some $r \geqslant 1$. However, it is easier to construct examples in the case of a componentwise supremum norm, as follows. Assume that X- and Y-distributions may be represented as $X = \Sigma_{j \geqslant 1} \alpha_j U_j \psi_j$ and $Y = \Sigma_{j \geqslant 1} \alpha_j V_j \psi_j$, where $\alpha_1, \alpha_2, \ldots$ is a sequence of non-negative constants satisfying $\Sigma_{j \geqslant 1} j\alpha_j < \infty$, $U_1, U_2, \ldots$ and $V_1, V_2, \ldots$ are sequences of independent and identically distributed random variables with finite variance and $\psi_1, \psi_2, \ldots$ is a sequence of bounded orthogonal functions. In this model the distance between the distributions of X and Y expresses the distance between the distributions of U_j and V_j. Given functions $u = \Sigma_{j \geqslant 1} \alpha_j u_j \psi_j$ and $v = \Sigma_{j \geqslant 1} \alpha_j v_j \psi_j$ in the common sample space of X and Y, define $\|u - v\| = \max_j (\alpha_j |u_j - v_j|)$. Then, for example,

$$P(\|u - X\| \leqslant \delta) = \prod_j P(|u_j - U_j| \leqslant \delta/\alpha_j).$$

Using this formula, its analogue for $P(\|u - Y\| \leqslant \delta)$ and the definition (4.2) of $\mathcal{S}_X(\eta)$ and $\mathcal{S}_Y(\eta)$, it may be determined whether $u \in \mathcal{S}_X(0)$ or $u \in \mathcal{S}_Y(0)$, where $\mathcal{S}_Z(0)$ denotes the limit, as $\eta \downarrow 0$, of $\mathcal{S}_Z(\eta)$.

We conclude this section by illustrating theory that can be developed when the distributions of X and Y are smooth in the sense of assumption (4.1), and relative density is defined by using expression (4.2). Theorem 3 below asserts that the bagged nearest neighbour classifier, applied to z, converges to the generalized Bayes classifier which assigns z to the population with greater relative density in the sense at expression (4.2). Under an additional assumption it may be proved that the risk of the bagged nearest neighbour classifier converges to that of the Bayes classifier.

Define $\mathcal{S}_Z(\eta, \varepsilon)$ to be the set of $z \in \mathcal{S}_Z(\eta)$ for which $P(\|Z - z\| \leqslant \eta) > \varepsilon$. Restricting attention to $z \in \mathcal{S}_Z(\eta, \varepsilon)$ for some $\varepsilon > 0$, rather than just to $z \in \mathcal{S}_Z(\eta)$, amounts to asking that the density of the distribution of Z is not too small in the neighbourhood of z.

Theorem 3. Making assumption (4.1), and that

$$\begin{aligned} &\text{the resample sizes } m_1 \text{ and } n_1 \text{ satisfy } \min(m_1, n_1) \to \infty, \max(m_1/m, n_1/n) \\ &\qquad \to 0 \text{ and } m_1/n_1 \to p/(1-p) \text{ as } m \to \infty, \end{aligned} \tag{4.3}$$

then, for each $\eta, \varepsilon > 0$,

$$\begin{aligned} &\inf_{z \in \mathcal{S}_X(\eta, \varepsilon)} \{P(\text{bagged nearest neighbour classifier assigns } z \text{ to } \Pi_X)\} \to 1, \\ &\inf_{z \in \mathcal{S}_Y(\eta, \varepsilon)} \{P(\text{bagged nearest neighbour classifier assigns } z \text{ to } \Pi_Y)\} \to 1 \end{aligned} \tag{4.4}$$

as $m \to \infty$.

Condition (4.3) reflects remarks that were made in the second-last paragraph of Section 3.1.

5. Choice of sampling fraction by cross-validation

5.1. Methodology

Let $\mathcal{X}_i = \mathcal{X} \setminus \{X_i\}$ and $\mathcal{Y}_i = \mathcal{Y} \setminus \{Y_i\}$ denote the two data sets after the ith data value has been dropped, where $1 \leqslant i \leqslant m$ or $1 \leqslant i \leqslant n$ in the respective cases. Write $\mathcal{C}_{-i,X}$ and $\mathcal{C}_{-i,Y}$ for the bagged nearest neighbour classifiers based on the sample pairs $(\mathcal{X}_i, \mathcal{Y})$ and $(\mathcal{X}, \mathcal{Y}_i)$ respectively, rather than on $(\mathcal{X}, \mathcal{Y})$. The classifier $\mathcal{C}_{-i,X}$ is constructed by sampling m_1 data from $\mathcal{X}_i$ and n_1 data from $\mathcal{Y}$, using either with- or without-replacement sampling, and analogously for

$\mathcal{C}_{-i,Y}$. To simplify optimization we shall put $r = m_1/m$ and take $n_1 = [rn]$, i.e. the integer part of rn, so that optimization is over only a single parameter. Both with- and without-replacement resampling could be used, and leave-one-out methods employed to minimize risk over both approaches as well as over r. However, for simplicity we shall assume that just one of the two types of resampling is employed, and that optimization over r is attempted for just that type.

A leave-one-out, or cross-validation-based, estimator of risk is

$$\widehat{\text{err}}(r) = \frac{p}{m} \sum_{i=1}^{m} I\{\mathcal{C}_{-i,X}(X_i) = Y\} + \frac{1-p}{n} \sum_{i=1}^{n} I\{\mathcal{C}_{-i,Y}(Y_i) = X\},$$

where '$\mathcal{C}_{-i,X}(X_i) = Y$' means that the classifier $\mathcal{C}_{-i,X}$, when applied to X_i, misclassifies the latter as coming from the Y-population. It is suggested that r be chosen to minimize $\widehat{\text{err}}(r)$.

5.2. *Large sample properties*

The expected value of $\widehat{\text{err}}(r)$ equals p times the average error of the bagged nearest neighbour classifier when the training sample sizes are $m-1$ and n, and data come from Π_X, plus $1-p$ times the average error when the training sample sizes are m and $n-1$, and data come from Π_Y. Therefore, assuming for the moment that

$$\max_{1 \leqslant m_1 \leqslant m-2,\, 1 \leqslant n_1 \leqslant n-2} \left| \widehat{\text{err}}(r) - E\{\widehat{\text{err}}(r)\} \right| \to 0 \tag{5.1}$$

in probability, where $m_1 = rm$ and $n_1 = [rn]$, we would expect $\widehat{\text{err}}(r)$ to represent accurately the risk of bagged nearest neighbour classifiers in cases where reducing the training sample sizes by 1 does not appreciably affect the risk, given by expression (2.2). Moreover, as argued in Sections 3 and 4, when m_1 and n_1 diverge but m_1/m and n_1/n converge to 0, the risk of the bagged nearest neighbour classifier is less than the risk when this condition fails.

It follows that the value, $\hat{r}$ say, of r which minimizes $\widehat{\text{err}}(r)$ must satisfy $\hat{r} \to 0$ and $m\hat{r} \to \infty$, both convergences occurring in probability. In consequence, it may be proved that the risk of the empirical classifier, constructed by choosing r to minimize $\widehat{\text{err}}(r)$, converges to the risk of the Bayes classifier under regularity conditions. An outline derivation of expression (5.1) is given in Appendix A.5.

Importantly, this is not the same as saying that cross-validation leads to asymptotic minimization of regret, i.e. of the difference between the risk of the bagged nearest neighbour classifier and that of the Bayes rule to which the bagged nearest neighbour classifier converges. Indeed, it can be shown that, in many cases, the regret depends on local rather than global properties of the two sampling distributions. For example, if the populations Π_X and Π_Y are univariate with densities f and g, and if the graphs of $y = p\, f(z)$ and $y = (1-p)\, g(z)$ cross, in the (z, y) plane, at only a finite number of points, say $z_1, \ldots, z_\nu$, then the values of m_1 and n_1 which minimize regret depend, to first order, on the behaviour of f and g in neighbourhoods of the points z_i. In the nomenclature of statistical classification, these points form the 'margin' of the present problem, i.e. the parts of the sample space where classification is difficult. It is known that standard cross-validation algorithms generally fail to produce optimality in such local, as distinct from global, settings and require additional smoothing, or regularization, to perform well at that level; see, for example, Hall and Schucany (1989) and Mielniczuk *et al.* (1989). In the present problem the smoothing that would be necessary to enable cross-validation to minimize regret asymptotically would depend intimately on the problem, e.g. on whether the data were d variate or functional.

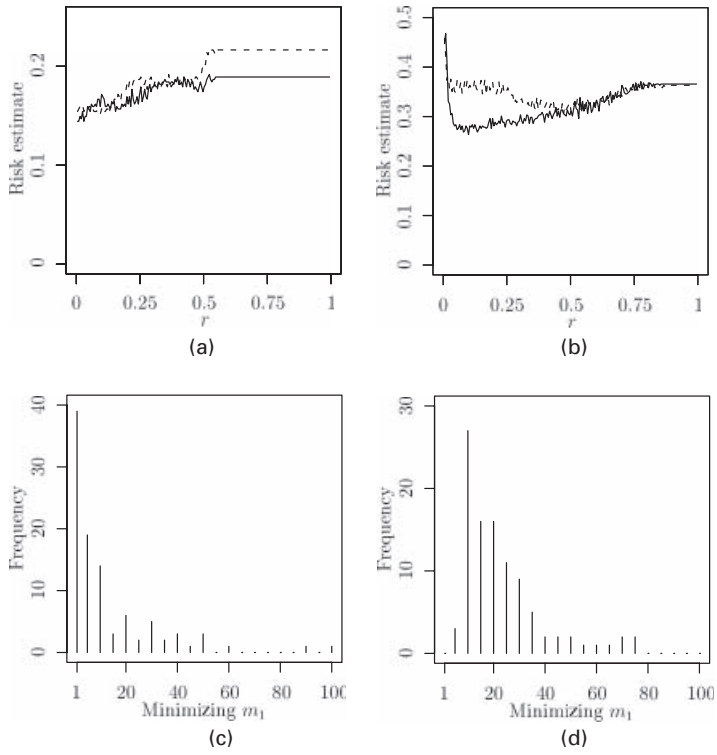

Fig. 4. Typical plots of cross-validation criterion and the frequencies with which different values of m_1 are selected by cross-validation: the sample sizes are $m = n = 200$; the densities in (a) and (c) are the same as those in Fig. 2(a), but with $\mu = 2$, whereas in (b) and (d) they are the same as those in Fig. 2(b), again with $\mu = 2$; in (a) and (b), results for two typical samples are shown, indicated by the broken and full lines; the horizontal and vertical axes graph $r = m_1/m$ $(= n_1/n)$ and $\widehat{\mathrm{err}}(r)$ respectively; (c) and (d) give the relative frequencies, obtained from 100 simulations, with which cross-validation selects different values of m_1; resampling is done without replacement in (a) and (c), and with replacement in (b) and (d)

5.3. Numerical properties

Figs 4(a) and 4(b) show, for the same distribution pairs that were used in Figs 2(a) and 2(b) respectively, plots of $\widehat{\mathrm{err}}(r)$ against $r = m_1/m$ for two typical data sets. The sample size is $m = n = 200$, as in Figs 2(c), 2(d), 2(e) and 2(f). Figs 4(c) and 4(d) give the frequencies with which different values of m_1 are selected by cross-validation. Resampling is without replacement for Figs 4(a) and 4(c) and with replacement for Figs 4(b) and 4(d). The results reflect very closely the fact, indicated in Fig. 2, that $m_1 = 1$ and $m_1 = 14$ are optimal in the respective cases of Figs 4(a) and 4(b).

Not only does cross-validation lead to an appropriate choice of m_1, it also produces significant reductions in risk, as we point out next. Of course, we expect that

$$\begin{aligned}\text{Bayes risk} &< \text{risk of bagged nearest neighbour classifier with } m_1 \text{ chosen by cross-validation}\\ &< \text{risk of regular nearest neighbour classifier.}\end{aligned}$$

The values of these three risks, in the contexts of the first two numerical examples that were treated in Section 3.2, are $0.158 < 0.163 < 0.224$ and $0.283 < 0.311 < 0.379$.

As an alternative to leave-one-out cross-validation, tenfold cross-validation could be used. Here, the training data are divided randomly into 10 equal parts and the classifier is based on

the data in all except one of the parts. The risk is estimated by attempting to classify the data in the remaining part. The risks of the bagged nearest neighbour classifier with m_1 chosen by tenfold cross-validation were 0.163 and 0.307 in the two examples.

Acknowledgements

The authors are grateful to three reviewers for constructive criticism, and to Jim Ramsay for making the weather data available. The second author was supported by an Engineering and Physical Sciences Research Council studentship and is grateful for a Visiting Fellowship from the Australian National University.

Appendix A: Technical details

A.1. Poisson approximations to spatial distributions

As the density of points in the sample space increases, and provided that $m/(m+n) \to p$, the distribution of data in the neighbourhood of each $z \in \mathbb{R}^d$ converges to that for a marked Poisson process $\mathcal{P}$, in which each point has one of two marks, X and Y, chosen independently of the marks for all other points. See for example Daley and Vere-Jones (1988), page 205, for discussion of marked point processes.

Indeed, let $\mathcal{P}$ denote a homogeneous Poisson process with intensity $pf + (1-p)g$ in $\mathbb{R}^d$, and let each of its points be marked as X or Y, with respective probabilities $q(Z)$ and $1-q(Z)$ (given that the point occurred at Z), independently of all other points. Assume that f and g are continuous and that $m/(m+n) \to p$. Let $\mathcal{R} \subseteq \mathbb{R}^d$ be a compact set on which $f+g>0$. Given $z \in \mathcal{R}$, let $T(z) = (T_1, \ldots, T_k)$ denote the vector of marks, each either X or Y, for the (nearest, $\ldots$, kth-nearest) respectively data in $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$, or in $\mathcal{P}$, to z. Denote by P the probability measure for the finite marked point process that is created by the points in $\mathcal{Z}$, with point marks X and Y ascribed in the obvious way, and let P^{Pois} be the analogous probability measure for the marked point process $\mathcal{P}$. Let $\mathcal{T}$ be the set of all 2^k vectors $t = (t_1, \ldots, t_k)$ of k marks, each either X or Y. Then,

$$\sup_{z \in \mathcal{R}} \sup_{t \in \mathcal{T}} |P\{T(z) = t\} - P^{\text{Pois}}\{T(z) = t\}| \to 0 \tag{A.1}$$

as $n \to \infty$. Result (A.1) implies that the probability that a new datum at z is classified as coming from X converges to $q(z)$. This property directly gives theorem 1.

A.2. Bagging a marked Poisson point process

Bagging applied to $\mathcal{P}$ amounts to the majority vote rule for classifiers based on resamples $\mathcal{P}^*$ drawn independently from $\mathcal{P}$. In the case of sampling with replacement, $\mathcal{P}^*$ will typically involve repeated data, but the number of repeats of any given data value is not used by the bagged classifier. Therefore we may disregard repeats and view $\mathcal{P}^*$ as simply a randomly chosen subset of $\mathcal{P}$. In the case of without-replacement resampling it is clear that $\mathcal{P}^*$ is a randomly chosen subset of $\mathcal{P}$.

Hence, we may view $\mathcal{P}^*$ as having been obtained from $\mathcal{P}$ by standard point process 'thinning', i.e. each point in $\mathcal{P}$ is 'killed or kept' with probability κ or $1-\kappa$ respectively, where $0 \leqslant \kappa \leqslant 1$ will depend on the type of resampling, and each point will be thinned independently of all other points.

Let $T_j = T_j(z)$ denote the type, either X or Y, of the point in $\mathcal{P}$ that is jth nearest to $z \in \mathcal{B}$. It follows from the definition of the thinned point process $\mathcal{P}^*$ that

$$\pi(\mathcal{P}) \equiv P(\text{the point in } \mathcal{P}^* \text{ that is nearest to } z \text{ has mark } X | \mathcal{P}) = \sum_{j=1}^{\infty} \kappa^{j-1}(1-\kappa)\, I(T_j = X). \tag{A.2}$$

Since the marks of the points of $\mathcal{P}$ are independent and identically distributed as X or Y, the former having probability $q(z)$ at equation (2.1), then the 0–1 variables $J_j = I(T_j = X)$ are independent and identically distributed, taking the value 1 with probability $q(z)$.

Suppose that we create $\mathcal{P}^*$ a total of B independent times, on each occasion starting from the same $\mathcal{P}$. Then, using a weak law of large numbers for a sum of B independent and identically distributed 0–1 variables, it may be proved that, excepting the case $\pi(\mathcal{P}) = \frac{1}{2}$,

the limit as $B \to \infty$ of the conditional probability, given $\mathcal{P}$, that for the majority of resampled point processes $\mathcal{P}^*$ the nearest point in $\mathcal{P}^*$ to z is from the X-population equals $I\{\pi(\mathcal{P}) > \frac{1}{2}\}$.

Therefore, the probability that the bagged nearest neighbour classifier applied to $\mathcal{P}$ assigns z to Π_X converges to $P\{\pi(\mathcal{P}) > \frac{1}{2}\}$, provided that we take $\kappa = \rho$ at result (A.2). Now, $P\{\pi(\mathcal{P}) > \frac{1}{2}\}$ is exactly the probability $P(\rho, z)$ at equation (3.2). It follows that the risk of the bagged nearest neighbour classifier converges to that given by equation (3.3). Recall that we always treat the infinite simulation case of bagging.

A.3. Approximating finite sample bagging by Poisson process bagging

Assume initially that we are conducting with-replacement resampling with a resample of size $m_1 \leqslant m$, and define $\rho_n = \exp(-m_1/m)$. Let $j \geqslant 1$ denote a fixed integer. Then the probability that a resample of size m_1 drawn from $\mathcal{X}$ excludes j specified data in $\mathcal{X}$ equals

$$(1 - jm^{-1})^{m_1} = \rho_n^j\{1 + O(m_1/m^2)\} \sim \rho_n^j,$$

since m_1/m is bounded and $m \to \infty$. Likewise, if assumption (3.1) holds then the probability that a resample of size n_1 drawn from $\mathcal{Y}$ excludes j specified data in $\mathcal{Y}$ is asymptotic to ρ_n^j as $n \to \infty$.

If, in contrast, resampling is without replacement, if we redefine $\rho_n = 1 - m_1/m$ and if $m - m_1 \to \infty$, then the probability that a resample of size m_1, drawn from $\mathcal{X}$ without replacement, excludes j specified data in $\mathcal{X}$ equals

$$\left(1 - \frac{j}{m}\right)\left(1 - \frac{j}{m-1}\right) \ldots \left(1 - \frac{j}{m - m_1 + 1}\right) = \rho_n^j\left[1 + O\left\{\frac{m_1}{m(m - m_1)}\right\}\right] \sim \rho_n^j.$$

Provided that assumption (3.1) holds, the probability that a resample of size n_1 from $\mathcal{Y}$ excludes j specified data in $\mathcal{Y}$ is also asymptotic to ρ_n^j.

Combining the results in the previous two paragraphs we deduce that for either with- or without-replacement resampling, provided that we take $\rho_n = \exp(-m_1/m)$ or $\rho_n = 1 - m_1/m$ respectively, the probabilities that resamples exclude j specific data are, in each case, asymptotic to ρ_n^j. Combining this property with the results in Appendix A.2 we deduce that, if $\rho_n \to \rho$, the probability that a new datum z is identified by the bagged nearest neighbour classifier, applied to the original finite data sets $\mathcal{X}$ and $\mathcal{Y}$, as coming from Π_X, converges to $P\{\pi(\mathcal{P}) > \frac{1}{2}\}$ as $n \to \infty$, where $\pi(\mathcal{P})$ is defined at equation (A.2), with $\kappa = \rho$. This proves the first part of theorem 2. The second part is a direct consequence of the first.

A.4. Proof of theorem 3

Order the values of $U(z) = \|z - Z\|$, for $Z \in \mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$, as $U_1(z) < \ldots < U_{m+n}(z)$, where $U_j(z) = \|z - Z_j\|$ and $Z_j = Z_j(z) \in \mathcal{Z}$. Define $\mathcal{Z}^* = \mathcal{X}^* \cup \mathcal{Y}^*$, let Z^* denote a generic value in $\mathcal{Z}^*$ and put

$$\begin{aligned}
\rho_j &= P(Z_j \notin \mathcal{Z}^* | Z_1, \ldots, Z_{j-1} \notin \mathcal{Z}^*; \mathcal{X}, \mathcal{Y}), \\
\pi(z|\mathcal{Z}) &= P(\text{nearest } Z^* \text{ to } z \text{ is in } \mathcal{X} | \mathcal{X}, \mathcal{Y}) \\
&= \sum_{j=1}^{m+n} \rho_1 \ldots \rho_{j-1}(1 - \rho_j)\, I(Z_j \in \mathcal{X}).
\end{aligned} \tag{A.3}$$

In this notation,

$$P(\text{bagged nearest neighbour classifier assigns } z \text{ to } \Pi_X) = P\{\pi(z|\mathcal{Z}) > \tfrac{1}{2}\}.$$

It may be proved from equation (A.3) that $\text{var}\{\pi(z|\mathcal{Z})\} \to 0$ uniformly in $z \in \mathcal{S}_X(\eta, \varepsilon)$, for each $\eta, \varepsilon > 0$; compare the argument below expression (3.4). From this result and Chebyshev's inequality we see that, to prove the first part of expression (4.4), it suffices to show that, for each $\eta, \varepsilon > 0$,

$$\liminf_{m \to \infty} \inf_{z \in \mathcal{S}_X(\eta, \varepsilon)} [E\{\pi(z|\mathcal{Z})\}] > \tfrac{1}{2}. \tag{A.4}$$

The second part of expression (4.4) follows analogously.

Finally we derive inequality (A.4). Let M and N equal the numbers of distinct values in $\mathcal{X}^*$ and $\mathcal{Y}^*$ respectively. For example, in the case of without-replacement sampling, $M = m_1$ and $N = n_1$. Now,

$$
\begin{aligned}
Q(z|M,N) &\equiv P(\text{nearest } Z^* \text{ to } z \text{ is in } \mathcal{X}|M,N) \\
&= \int_{\delta>0} P(\|Y-z\|>\delta)^N \mathrm{d}_\delta\{1-P(\|X-z\|>\delta)^M\} \\
&= \tfrac{1}{2} + Q_1(z|M,N),
\end{aligned}
$$

where

$$
Q_1(z|M,N) = M\int_{\delta>0}\{1-\pi_X(\delta|z)\}^{2M-1}\left[\frac{\{1-\pi_Y(\delta|z)\}^N}{\{1-\pi_X(\delta|z)\}^M}-1\right]\mathrm{d}_\delta\pi_X(\delta|z).
$$

Given $\eta\in(0,\infty)$, put $Q_1 = Q_2 + Q_3$ where

$$
Q_2(z|M,N) = M\int_{\delta\leqslant\eta}\{1-\pi_X(\delta|z)\}^{2M-1}\left[\frac{\{1-\pi_Y(\delta|z)\}^N}{\{1-\pi_X(\delta|z)\}^M}-1\right]\mathrm{d}_\delta\pi_X(\delta|z), \tag{A.5}
$$

$$
\begin{aligned}
|Q_3(z|M,N)| &\leqslant M\int_{\delta>\eta}\{1-\pi_X(\delta|z)\}^{2M-1}\left[\frac{\{1-\pi_Y(\delta|z)\}^N}{\{1-\pi_X(\delta|z)\}^M}+1\right]\mathrm{d}_\delta\pi_X(\delta|z) \\
&\leqslant 2\{1-\pi_X(\eta|z)\}^M. \qquad (A.6)
\end{aligned}
$$

The distributions of M and N depend only on m, n, m_1 and n_1, not on z or on the distributions of X or Y; and $M=(1-\Delta_1)m_1$ and $N=(1-\Delta_2)n_1$, where Δ_j denotes a random variable satisfying $P(0\leqslant\Delta_j\leqslant 1)=1$ and $P(\Delta_j>\gamma)\to 0$, as $m\to\infty$, for each $\gamma>0$. Call this property (P_1). It follows from property (P_1) and the definition of $\mathcal{S}_X(\eta,\varepsilon)$ that $E[\{1-\pi_X(\eta|z)\}^M]\to 0$ uniformly in $z\in\mathcal{S}_X(\eta,\varepsilon)$. Hence, by inequality (A.6), $E|Q_3(z\,|\,M,N)|\to 0$ uniformly in the same sense. Call this property (P_2). Note also that, by assumption (4.3), $m_1/n_1\to p/(1-p)$. Using this property, property (P_1) and equation (A.5) we deduce that

$$
\liminf_{m\to\infty}\inf_{z\in\mathcal{S}_X(\eta,\varepsilon)}[E\{Q_2(z|M,N)\}]>0. \tag{A.7}
$$

Combining this formula with property (P_2) we deduce that inequality (A.7) continues to hold if Q_2 there is replaced by Q_1. Since $E\{\pi(z|\mathcal{Z})\}=E\{Q(z|M,N)\}=\frac{1}{2}+E\{Q_1(z|M,N)\}$ then this result is equivalent to inequality (A.4).

A.5. Derivation of assumption (5.1)

Assume that n/m is bounded. Moment methods may be used to prove that, for sufficiently large integers $k\geqslant 1$,

$$
E[\widehat{\mathrm{err}}(r)-E\{\widehat{\mathrm{err}}(r)\}]^{2k}=o(m^{-1})
$$

uniformly in r, as $m\to\infty$. Therefore, by Markov's inequality,

$$
P[|\widehat{\mathrm{err}}(r)-E\{\widehat{\mathrm{err}}(r)\}|>\varepsilon]=o(m^{-1}),
$$

uniformly in r, for each $\varepsilon>0$. This implies assumption (5.1); note that there are no more than m distinct values of r.

References

Bay, S. D. (1998) Combining nearest classifiers through multiple feature subsets. In *Proc. 15th Int. Conf. Machine Learning*, pp. 37–45. San Francisco: Morgan Kaufmann.

Bay, S. D. (1999) Nearest neighbor classification from multiple feature subsets. *Intell. Data Anal.*, **3**, 191–209.

Bickel, P. J., Götze, F. and Van Zwet, W. R. (1997) Resampling fewer than n observations: gains, losses, and remedies for losses. *Statist. Sin.*, **7**, 1–31.

Breiman, L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 123–140.

Breiman, L. (1999) Using adaptive bagging to debias regressions. *Technical Report 547*. Department of Statistics, University of California, Berkeley.

Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *Ann. Statist.*, **30**, 927–961.

Buja, A. and Stuetzle, W. (2000a) The effect of bagging on variance, bias, and mean squared error. *Manuscript*. University of Pennsylvania, Philadelphia.

Buja, A. and Stuetzle, W. (2000b) Smoothing effects of bagging. *Manuscript*. University of Pennsylvania, Philadelphia.
Cover, T. M. (1968) Rates of convergence for nearest neighbor procedures. In *Proc. Hawaii Int. Conf. System Sciences* (eds B. K. Kinariwala and F. F. Kuo), pp. 413–415. Honolulu: University of Hawaii Press.
Cover, T. M. and Hart, P. E. (1967) Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory*, **13**, 21–27.
Daley, D. J. and Vere-Jones, D. (1988) *An Introduction to the Theory of Point Processes*. New York: Springer.
Devroye, L. (1982) Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Trans. Pattn Anal. Mach. Intell.*, **4**, 154–157.
Devroye, L., Györfi, L. and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition*. New York: Springer.
Dudoit, S., Fridlyand, J. and Speed, T. P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Statist. Ass.*, **97**, 77–87.
Efron, B. (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Statist. Ass.*, **78**, 316–331.
Efron, B. and Tibshirani, R. (1997) Improvements on cross-validation: the .632+ bootstrap method. *J. Am. Statist. Ass.*, **92**, 548–560.
Fix, E. and Hodges, J. (1951) Discriminatory analysis, nonparametric discrimination: consistency properties. *Technical Report 4, Project 21–49–004*. US Air Force School of Aviation Medicine, Randolph Field.
Francois, J., Grandvalet, Y., Denouex, T. and Roger, J.-M. (2001) Bagging belief structures in Dempster-Shafer k-NN rule. In *Information Processing and Management of Uncertainty in Knowledge-based Systems*, vol. I, pp. 111–118. Madrid: Universidad Politecnica de Madrid.
Freund, Y. and Schapire, R. E. (1997) A decision-theoretic generalization of online learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**, 119–139.
Friedman, J. H. and Hall, P. (2000) On bagging and nonlinear estimation. *Manuscript*. Stanford University, Stanford.
Friedman, J. H., Hastie, T. and Tibshirani, R. (2000) Additive logistic regression: a statistical view of boosting. *Ann. Statist.*, **28**, 337–374.
Fukunaga, K. and Hummels, D. M. (1987) Bias of nearest neighbor error estimates. *IEEE Trans. Pattn Anal. Mach. Intell.*, **9**, 103–112.
Guerra-Salcedo, C. and Whitley, D. (1999) Genetic approach to feature selection for ensemble creation. In *Proc. Genetic and Evolutionary Computation Conf.* (eds W. Banzhaf, J. Daida, A. E. Eiben, M. H. Garzon, V. Honavar, M. Jakiela and R. E. Smith), vol. 1, pp. 236–243. San Francisco: Morgan Kaufmann.
Hall, P. and Schucany, W. R. (1989) A local cross-validation algorithm. *Statist. Probab. Lett.*, **8**, 109–117.
Hand, D. J. (1981) *Discrimination and Classification*. New York: Wiley.
Ho, T. K. (1998a) Decision forests. In *Proc. 14th Int. Conf. Pattern Recognition, Brisbane,* pp. 545–549.
Ho, T. K. (1998b) The random subspace method for constructing decision forests. *IEEE Trans. Pattn Anal. Mach. Intell.*, **20**, 832–844.
Jiang, W. X. (2002) On weak base hypotheses and their implications for boosting regression and classification. *Ann. Statist.*, **30**, 51–73.
Kim, H. and Loh, W. Y. (2001) Classification trees with unbiased multiway splits. *J. Am. Statist. Ass.*, **96**, 589–604.
Kuncheva, L. I. and Bezdek, J. C. (1998) Nearest prototype classification: clustering, genetic algorithms, or random search? *IEEE Trans. Syst. Man Cyber.* C, **28**, 160–164.
Kuncheva, L. I., Skurichina, M. and Duin, R. P. W. (2002) An experimental study on diversity for bagging and boosting with linear classifiers. *Inform. Fusion*, 245–258.
Larkey, L. S. and Croft, W. B. (1996) Combining classifiers in text categorization. In *Proc. 19th A. Int. Conf. Research and Development in Information Retrieval* (eds H.-P. Frei, D. Harman, D. Schäuble and R. Wilkinson), pp. 289–297. New York: Association for Computing Machinery.
Lugosi, G. and Nobel, A. (1996) Consistency of data-driven histogram methods for density estimation and classification. *Ann. Statist.*, **24**, 687–706.
Mammen, E. (1992) When does bootstrap work?: asymptotic results and simulations. *Lect. Notes Statist.*, **77**.
Marron, J. S. (1983) Optimal rates on convergence to Bayes risk in nonparametric discrimination. *Ann. Statist.*, **11**, 1142–1155.
Mielniczuk, J., Sarda, P. and Vieu, P. (1989) Local data-driven bandwidth choice for density estimation. *J. Statist. Planng Inf.*, **23**, 53–69.
Mollineda, R. A., Ferri, F. J. and Vidal, E. (2000) Merge-based prototype selection for nearest-neighbor classification. In *Proc. 4th Wrld Multiconf. Systemics, Cybernetics and Informatics*, vol. 7 (eds Z. Huang, C. Sun and I. McDonald), pp. 640–645. Piscataway: Institute of Electrical and Electronic Engineers.
Psaltis, D., Snapp, R. R. and Venkatesh, S. S. (1994) On the finite sample performance of the nearest neighbour classifier. *IEEE Trans. Inform. Theory*, **40**, 820–837.
Schapire, R. E., Freund, Y., Bartlett, P. and Lee, W. D. (1998) Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Statist.*, **26**, 1651–1686.
Skurichina, M. and Duin, R. P. W. (1998) Bagging for linear classifiers. *Pattn Recogn*, **31**, 909–930.
Skurichina, M., Kuncheva, L. I. and Duin, R. P. W. (2002) Bagging and boosting for the nearest mean classifier: effects of sample size on diversity and accuracy. *Lect. Notes Comput. Sci.*, **2364**, 62–71.

Steele, B. M. and Patterson, D. A. (2000) Ideal bootstrap estimation of expected prediction error for k-nearest neighbor classifiers: applications for classification and error assessment. *Statist. Comput.*, **10**, 349–355.

Stoller, D. S. (1954) Univariate two-population distribution-free discrimination. *J. Am. Statist. Ass.*, **49**, 770–777.

Yang, Y. H. (1999) Minimax nonparametric classification—part I: rates of convergence. *IEEE Trans. Inform. Theory*, **45**, 2271–2284.

Zemke, S. (1999) Bagging imperfect predictors. In *Proc. Artificial Neural Networks in Engineering* (eds A. Buczak, C. Dagli, M. Embrechts, O. Ersoy and J. Ghosh), pp. 1067–1072. St Louis: American Society of Mechanical Engineers.