Ideal Bootstrap Estimation of Expected Prediction Error for k-Nearest Neighbor Classifiers: Applications for Classification and Error Assessment

Brian M. Steele and David A. Patterson

Department of Mathematical Sciences, University of Montana, Missoula Montana, 59812-1032

Correspondence should be directed to the first author.

ABSTRACT

Euclidean distance k-nearest neighbor (k-NN) classifiers are simple nonparametric classification rules. Bootstrap methods, widely used for estimating the expected prediction error of classification rules, are motivated by the objective of calculating the ideal bootstrap estimate of expected prediction error. In practice, bootstrap methods use Monte Carlo resampling to estimate the ideal bootstrap estimate because exact calculation is generally intractable. In this article, we present analytic formulae for exact calculation of the ideal bootstrap estimate of expected prediction error for k-NN classifiers and propose a new weighted k-NN classifier based on resampling ideas. The resampling-weighted k-NN classifier replaces the k-NN posterior probability estimates by their expectations under resampling and predicts an unclassified covariate to belong to the group with the largest resampling expectation. A simulation study and an application involving remotely sensed data show that the resampling-weighted k-NN classifier compares favorably to unweighted and distance-weighted k-NN classifiers.

1. INTRODUCTION

Suppose that a population consists of g groups and that n independent training observations have been collected by probability sampling. Each observation in the training sample $\boldsymbol{x}$ is a pair consisting of a covariate vector and a label identifying group membership. The classification problem is to construct a rule $\eta_{\boldsymbol{x}}$ from $\boldsymbol{x}$ that will be used for predicting group membership of unclassified covariates. The prediction error of $\eta_{\boldsymbol{x}}$ is a binary random variable taking on the value 1 if the prediction is incorrect, and 0 if correct. The expected prediction error of $\eta_{\boldsymbol{x}}$, over all possible samples and covariates, is the probability of misclassifying an unclassified covariate.

The bootstrap (Efron 1982, Efron and Tibshirani 1993) is a popular method of estimating expected prediction error. The bootstrap idea is to replace the unknown distribution function F from which $\boldsymbol{x}$ has been sampled by the empirical distribution function $\widehat{F}$ and calculate expected prediction error over $\widehat{F}$ instead of F. This estimate is called the ideal bootstrap estimate of expected prediction error by Efron and Tibshirani (1993) because analytic calculation is almost always intractable. Consequently, bootstrap methods almost always estimate the ideal bootstrap estimate of prediction error by Monte Carlo resampling. There are a variety of Monte Carlo bootstrap methods, but this article concentrates on the leave-one-out bootstrap (Efron 1983, Efron and Tibshirani 1997) because it is particularly well-suited for estimating expected prediction error of k-nearest neighbor (k-NN) classifiers.

In this article, we derive analytic expressions for the leave-one-out ideal bootstrap estimator of expected prediction error for k-NN classifiers. While the practical applications of these analytic formulae are somewhat limited, they motivate a new k-NN classifier which replaces the k-NN posterior probability estimator by its expectation under resampling. This classifier, which we refer to as the resampling-weighted k-NN classifier, is easy and fast to compute and avoids the need for tie breaking. Comparisons of resampling-weighted and conventional k-NN classifiers, and Dudani's (1976) and Macleod, Luk and Titterington's (1987) distance-weighted k-NN classifiers, indicate that resampling-weighted k-NN classifiers may outperform these classifiers for some problems.

The article is organized as follows. Section 2 sets up the classification problem and notation. Section 3 discusses bootstrap estimation of expected prediction error and formulae for the ideal

bootstrap estimate of expected prediction error for conventional k-NN classifiers. Section 4 discusses distance-weighted k-NN classifiers and analytic calculation of the ideal estimate of expected prediction error for these classifiers. Section 5 introduces the resampling-weighted k-NN classifier and compares it to other k-NN classifiers via simulation. Section 6 provides an additional comparison involving a land cover classification problem and Section 7 concludes.

2. CLASSIFICATION AND ASSESSMENT OF PREDICTION ERROR

Suppose that a training sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ has been collected by random sampling of a population $\mathcal{P}$ consisting of g groups $G_1, \ldots, G_g$. Each observation $x_i = (t_i, y_i)$ consists of a covariate vector t_i and a group label $y_i \in \{1, \ldots, g\}$. The classification objective is to construct a classification rule $\eta_{\boldsymbol{x}}$ for predicting the membership of an unclassified covariate vector $t_0 \in \mathcal{P}$. Usually, a classification rule can be viewed as a method of estimating the posterior probability of membership in G_l. If so, then the rule predicts t_0 to belong to the group with the largest posterior probability estimate. We denote the posterior probability of membership in G_l by $\tau_l(t_0) = P(y_0 = l \mid t_0)$.

Conventional Euclidean distance k-NN classifiers estimate $\tau_l(t_0)$ by the proportion of the k-nearest neighbors of t_0 belonging to G_l. To develop these ideas further, let $\boldsymbol{t}_{0,j}$ denote the jth closest observation to t_0 among $\{t_1, \ldots, t_n\}$, where the distance between covariate vectors is Euclidean distance, and let $\boldsymbol{y}_{0,j}$ denote the group label for $\boldsymbol{t}_{0,j}$. The conventional k-NN estimate of $\tau_l(t_0)$ is the sample proportion of observations belonging to G_l among the k nearest neighbors and can be expressed as

$$\tau_l^c(t_0) = \frac{1}{k}\sum_{j=1}^{k} \Psi(\boldsymbol{y}_{0,j} = l), \tag{1}$$

where $\Psi(P)$ is the indicator function of the event P. Ties among the maximum posterior probability estimates may be broken by randomly choosing among the tied groups or by increasing the neighborhood size and recomputing formula (1).

The following notation is based on Efron and Tibshirani (1993, 1997) although it is not entirely consistent with either. The prediction error of $\eta_{\boldsymbol{x}}$ in classifying t_0 is denoted by

$$Q(x_0, \boldsymbol{x}) = \begin{cases} 1, & \text{if } \eta_{\boldsymbol{x}}(t_0) \neq y_0, \\ 0, & \text{if } \eta_{\boldsymbol{x}}(t_0) = y_0, \end{cases}$$

where $x_0 = (t_0, y_0)$ and $\eta_{\boldsymbol{x}}(t_0)$ is the predicted group membership. The *conditional expected prediction error* of $\eta_{\boldsymbol{x}}$ is $\text{err}(\boldsymbol{x}, F) = E_{0F}Q(x_0, \boldsymbol{x})$ where the expectation is conditional on the sample $\boldsymbol{x}$ and over F, the distribution of x_0. The *expected prediction error* is the expectation of $\text{err}(\boldsymbol{x}, F)$ over the distribution of $\boldsymbol{x}$, and is denoted by $E_F\text{err}(\boldsymbol{x}, F)$. Let $\widehat{F}$ denote the empirical distribution function of $\boldsymbol{x}$ placing probability mass $1/n$ at each $x_i \in \boldsymbol{x}$. The apparent error rate $\text{err}(\boldsymbol{x}, \widehat{F}) = \sum_{i=1}^{n} Q(x_i, \boldsymbol{x})/n$ is a simple estimate of conditional expected prediction error. However, it is optimistically biased because each x_i is used both to construct the classification rule and to evaluate the prediction error of the rule.

3. BOOTSTRAP ESTIMATES OF EXPECTED PREDICTION ERROR

Let $\boldsymbol{x}^*$ denote a random bootstrap sample obtained by sampling $\widehat{F}$ with replacement. The ideal bootstrap estimator of the expected prediction error replaces the realized sample $\boldsymbol{x}$ by a random bootstrap sample $\boldsymbol{x}^*$ in $\text{err}(\boldsymbol{x}, \widehat{F})$ and takes the expectation over $\widehat{F}$. Hence, the ideal estimator is

$$E_{\widehat{F}}\text{err}(\boldsymbol{x}^*, \widehat{F}) = E_{\widehat{F}}\frac{1}{n}\sum_{i=1}^{n} Q(x_i, \boldsymbol{x}^*).$$

For nearly all practical problems, $E_{\widehat{F}}Q(x_i, \boldsymbol{x}^*)$ is very difficult to calculate analytically. The standard method of applying bootstrap ideas to estimation of expected prediction error uses Monte Carlo resampling to sample $\widehat{F}$ and approximate $E_{\widehat{F}}\text{err}(\boldsymbol{x}^*, \widehat{F})$. For instance, select B bootstrap samples $\boldsymbol{x}^{*1}, \ldots, \boldsymbol{x}^{*B}$ from $\widehat{F}$ and construct a classification rule from each sample. Each bootstrap rule is used to classify $\boldsymbol{x}$, and the estimate of $E_{\widehat{F}}\text{err}(\boldsymbol{x}^*, \widehat{F})$ is the average proportion of misclassified training observations given by

$$\widehat{E}_{\widehat{F}}\text{err}(\boldsymbol{x}^*, \widehat{F}) = \frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\sum_{i=1}^{n} Q(x_i, \boldsymbol{x}^{*b}).$$

As illustrated in the next section, $\widehat{E}_{\widehat{F}}\text{err}(\boldsymbol{x}^*, \widehat{F})$ may be is optimistically biased. Efron and Tibshirani (1997) discuss bias-corrected estimates of expected prediction error and related topics.

3.1. Bootstrap Estimates of Expected Prediction Error for k-NN Classification Rules

Suppose that each observation in $\boldsymbol{x}$ is unique and consider the 1-NN Euclidean distance classifier. If x_i is in $\boldsymbol{x}^{*b}$, then the bootstrap rule constructed from $\boldsymbol{x}^{*b}$ is certain to correctly classify x_i. Furthermore, the probability that x_i is in $\boldsymbol{x}^{*b}$ is $1 - P(\, x_i \notin \boldsymbol{x}^{*b}) = 1 - [1 - 1/n]^n$, and it is easy to show that this probability tends to .632 as n tends to infinity. Especially for small k, $\widehat{E}_{\widehat{F}}\text{err}(\boldsymbol{x}^*, \widehat{F})$ may be badly and optimistically biased. The *leave-one-out* bootstrap (Efron 1983, Efron and Tibshirani 1997) avoids optimistic bias at the cost of a small amount of negative bias. The leave-one-out idea is to remove one observation from $\boldsymbol{x}$, choose a bootstrap sample from the remaining observations, compute a classification rule from the bootstrap sample, and then evaluate the rule using the left-out observation. To develop this idea further, let $\widehat{F}_{(i)}$ denote an empirical distribution function placing probability mass $1/(n-1\,)$ at each $x_j \in \boldsymbol{x}\,, x_j \neq x_i$, and let $\boldsymbol{x}^*_{(i)}$ denote a bootstrap sample from $\widehat{F}_{(i)}$. The leave-one-out ideal bootstrap estimator of expected prediction error is the average of the n expected prediction errors $E_{\widehat{F}_{(i)}} Q(x_i, \boldsymbol{x}^*_{(i)})$ given by

$$E_{\widehat{F}_{(\cdot)}}\text{err}(\boldsymbol{x}^*, \widehat{F}) = \frac{1}{n}\sum_{i=1}^{n} E_{\widehat{F}_{(i)}} Q(x_i, \boldsymbol{x}^*_{(i)}).$$

For simplicity, we write $\text{Err}^{(1)}$ in place of $E_{\widehat{F}_{(\cdot)}}\text{err}(\boldsymbol{x}^*, \widehat{F})$. Efron and Tibshirani (1997) provide a computationally efficient estimator of $\text{Err}^{(1)}$ given by

$$\widehat{\text{Err}}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_b \Psi(x_i \notin \boldsymbol{x}^{*b}) Q(x_i, \boldsymbol{x}^{*b})}{\sum_b \Psi(x_i \notin \boldsymbol{x}^{*b})}\,.$$

3.2. The Ideal Estimate of Expected Prediction Error for k-NN Classifiers

For k-NN Euclidean distance classifiers, $\text{Err}^{(1)}$ can be expressed analytically. In this section, we demonstrate how to derive expressions for $\text{Err}^{(1)}$ involving 2^{k-1} equations and given a value of k. Consider $\text{Err}^{(1)}$ when $k = 1$. Because $\boldsymbol{x}^*_{(i)}$ is a bootstrap sample from $\widehat{F}_{(i)}$, there are only $n-1$ possible neighbors of x_i in $\boldsymbol{x}^*_{(i)}$. Let $\boldsymbol{t}_{i,j}$ denote the jth nearest covariate to t_i in $\boldsymbol{x}$ besides t_i and let

$\boldsymbol{y}_{i,j}$ denote the membership label of $\boldsymbol{t}_{i,j}$. Also, let $\boldsymbol{t}^*_{i,1}$ denote the covariate nearest to t_i in $\boldsymbol{x}^*_{(i)}$. Because t_i will be misclassified whenever $\boldsymbol{t}^*_{i,1} = \boldsymbol{t}_{i,j}$ and $\boldsymbol{y}_{i,j} \neq y_i$, the resampling expectation of prediction error in classifying t_i is the sum over all possible nearest neighbors of t_i given by

$$E_{\widehat{F}_{(i)}} Q(x_i, \boldsymbol{x}^*_{(i)}) = \sum_{j \neq i} P_{\widehat{F}_{(i)}}(\boldsymbol{t}^*_{i,1} = \boldsymbol{t}_{i,j}) \Psi(\boldsymbol{y}_{i,j} \neq y_i). \tag{2}$$

To derive a computational formula for $P_{\widehat{F}_{(i)}}(\boldsymbol{t}^*_{i,1} = \boldsymbol{t}_{i,j})$, let A_j denote the event that $\boldsymbol{x}^*_{(i)}$ is drawn from $\{(\boldsymbol{t}_{i,j}, \boldsymbol{y}_{i,j}), \ldots, (\boldsymbol{t}_{i,n-1}, \boldsymbol{y}_{i,n-1})\}$ and note that $P(A_j) = [(n-j)/(n-1)]^{n-1}$. Because $\boldsymbol{t}^*_{i,1} = \boldsymbol{t}_{i,j}$ if and only if both A_j and A^c_{j+1} occur, and because $A_{j+1} \subset A_j$,

$$\begin{aligned} P_{\widehat{F}_{(i)}}(\boldsymbol{t}^*_{i,1} = \boldsymbol{t}_{i,j}) &= P(A_j) - P(A_{j+1}) \\ &= \left(\frac{n-j}{n-1}\right)^{n-1} - \left(\frac{n-j-1}{n-1}\right)^{n-1}. \end{aligned} \tag{3}$$

Combining equations (2) and (3) yields the following ideal bootstrap estimate of expected prediction error for $k = 1$:

$$\begin{aligned} \text{Err}^{(1)} &= \frac{1}{n}\sum_{i=1}^{n} E_{\widehat{F}_{(i)}} Q(x_i, \boldsymbol{x}^*_{(i)}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j \neq i}\left\{\left(\frac{n-j}{n-1}\right)^{n-1} - \left(\frac{n-j-1}{n-1}\right)^{n-1}\right\}\Psi(\boldsymbol{y}_{i,j} \neq y_i). \end{aligned}$$

Besides improving on the accuracy of $\widehat{\text{Err}}^{(1)}$, $\text{Err}^{(1)}$ is easy to compute.

We illustrate our general approach for deriving analytic expressions for $\text{Err}^{(1)}$ using the case $k = 3$. Let $\boldsymbol{t}^*_{i,j}$ denote the jth closest covariate to t_i among $\boldsymbol{x}^*_{(i)}$, let $B_{h,j,l}$ denote the event $(\boldsymbol{t}^*_{i,1}, \boldsymbol{t}^*_{i,2}, \boldsymbol{t}^*_{i,3}) = (\boldsymbol{t}_{i,h}, \boldsymbol{t}_{i,j}, \boldsymbol{t}_{i,l})$, and let $Q(x_i, B_{h,j,l})$ denote the prediction error when classifying t_i, given that $B_{h,j,l}$ has occurred. The resampling expectation of prediction error in classifying t_i is the sum over all possible three closest neighbors given by

$$E_{\widehat{F}_{(i)}} Q(x_i, \boldsymbol{x}^*_{(i)}) = \sum_{1 \leq h \leq j \leq l \leq n-1} P_{\widehat{F}_{(i)}}(B_{h,j,l}) Q(x_i, B_{h,j,l}).$$

As before, $\text{Err}^{(1)}$ is the average of $E_{\widehat{F}_{(i)}} Q(x_i, \boldsymbol{x}^*_{(i)})$ over $i = 1, \ldots, n$. Formulae for computing $P_{\widehat{F}_{(i)}}(B_{h,j,l})$ are derived in the Appendix for each of the relations $h = j = l$,$h = j < l$,$h < j = l$, and $h < j < l$. For large n, it is neither practical nor necessary to calculate these probabilities for all values of h, j, and l. Instead, $P_{\widehat{F}_{(i)}}(B_{h,j,l})$ can be approximated by zero whenever one or more of h, j, and l are greater than some minimum value, say 15, because it is very unlikely that any of the three closest neighbors to t_i among a bootstrap sample will be more distant than the original 15th nearest neighbor. For instance, the probability of selecting at least one of $\boldsymbol{t}^*_{0,1}$, $\boldsymbol{t}^*_{0,2}$ or $\boldsymbol{t}^*_{0,3}$ from $\{\boldsymbol{t}_{0,15}, \ldots, \boldsymbol{t}_{0,n}\}$ is 4×10^{-6}, 21×10^{-6}, and 32×10^{-6}, when n is to equal 15, 150 and 450, respectively. A lower bound on the ideal bootstrap estimate of expected prediction *accuracy*, $1 - \text{Err}^{(1)}$, is given by the partial sum

$$\frac{1}{n}\sum_{i=1}^{n} \sum_{1 \le h \le j \le l \le n'} P_{\widehat{F}_{(i)}}(B_{h,j,l}) \Psi\{Q(x_i, \boldsymbol{x}^*_{(i)}) = 0\} \le 1 - \text{Err}^{(1)}.$$

where $n' < n$ is chosen to expedite computational effort. An upper bound on $\text{Err}^{(1)}$ can be calculated by subtracting the lower bound on expected prediction accuracy from 1.

There are several problems that limit the practical utility of the analytic formulae for computing upper bounds on $\text{Err}^{(1)}$. The computation may be time-consuming even when k is smaller than 5 because of the large number of terms in the partial sum approximation. Furthermore, implementation of our approach is burdensome when $k > 5$ because 2^{k-1} distinct formulae are needed to compute the resampling expectation of prediction error for k nearest neighbors. If the classification rule breaks ties by increasing the neighborhood size, then additional formulae and computational effort will be needed.

4. DISTANCE-WEIGHTED *k*-NN CLASSIFIERS

Before turning to other applications of resampling ideas to k-NN classifiers besides calculating $\text{Err}^{(1)}$, we review distance-weighted k-NN classifiers and show that analytic formulae for $\text{Err}^{(1)}$ can be derived for these classifiers. Equation (1) shows that the k-NN Euclidean distance classifier assigns weight $1/k$ to each of the k nearest neighbors. Because nearer neighbors may be more informative

than distant neighbors, Dudani (1976) proposed a distance-weighted classifier. Dudani's (1976) weight for the jth closest neighbor is $w_{j,k}^{d} = (d_k - d_j)/(d_k - d_1)$ if $d_k \neq d_1$, and $w_{j,k}^{d} = 1$ if $d_k = d_1$, where d_j is the Euclidean distance from t_0 to $\boldsymbol{t}_{0,j}$, $j = 1, \ldots, k$. An unclassified covariate t_0 is assigned to G_l if $\sum_{j=1}^{k} w_{j,k}^{d} \Psi(\boldsymbol{y}_{0,j} = l)$ is largest. Macleodet al. (1987) proposed a more general distance-weighting function and demonstrated improvement over Dudani's (1976) weighted rule. In Section 5, we use a form of Macleodet al.'s (1987) generalized weight given by $w_{j,n}^{m} = (d_n - d_j)$ $/(d_n - d_1)$ where d_n is the maximum distance between t_0 and any training covariate. The simulation studies of Dudani (1976) and Macleod et al. (1987) demonstrated that weighted rules may improve on the unweighted rule for finite n although Bailey and Jain (1978) showed that the asymptotic performance of a weighted rule is no better than the unweighted rule.

Analytic formulae for $\text{Err}^{(1)}$ can be determined for distance-weighted k-NN classifiers as follows. For arbitrary k, let $B_{h,\ldots,l}$ denote the event $(\boldsymbol{t}_{i,1}^{*}, \ldots, \boldsymbol{t}_{i,k}^{*}) = (t_{i,h}, \ldots, t_{i,l})$, $1 \leq h \leq \cdots \leq l$ $\leq n-1$, and let $Q(x_i, B_{h,\ldots,l})$ denote the prediction error of a distance-weighted rule when classifying t_i given that $B_{h,\ldots,l}$ has occurred. Then, $\text{Err}^{(1)}$ is the average of the n resampling expectations

$$E_{\widehat{F}_{(i)}} Q(x_i, \boldsymbol{x}_{(i)}^{*}) = \sum_{1 \leq h \leq \cdots \leq l \leq n-1} P_{\widehat{F}_{(i)}}(B_{h,\ldots,l}) Q(x_i, B_{h,\ldots,l}).$$

5. RESAMPLING WEIGHTED k-NN CLASSIFIERS

The k-NN estimate of $\tau_l(t_0)$ is the sample proportion of observations in a k nearest-neighborhood of t_0 belonging to G_l. This estimator may be criticized for two reasons. Unless k is large, it is not smooth because there are only $k+1$ possible realizations. Secondly, each of the k neighbors has equal weight in determining the estimate even though the informativeness of the neighbors may decrease as the distance between t_0 and its neighbors increases. In some instances, Euclidean distance weights, such as those proposed by Dudani (1976) and Macleod et al. (1987), may not be an effective scale for measuring the value of each neighbor for classifying t_0. Furthermore, weights based on the distance ranks seem truer to the nonparametric nature of the k-NN classifier than Euclidean distances. In the

next section, a smoother estimator of $\tau_l(t_0)$ is proposed which gives each training observation a weight determined by its distance rank to t_0.

5.1 The Resampling-Weighted k-NN Classifier

We propose to smooth the k-NN estimator of $\tau_l(t_0)$ taking its resampling expectation. This is accomplished by replacing the sample $\boldsymbol{x}$ in $\tau_l^c(t_0)$ (equation [1]) by a random bootstrap sample $\boldsymbol{x}^*$, and calculating the expectation of this function with respect to $\widehat{F}$. Thus, the resampling-weighted k-NN estimator of $\tau_l(t_0)$ is

$$\begin{aligned}\tau_l^r(t_0) &= \frac{1}{k}\sum_{j=1}^{k} E_{\widehat{F}}\Psi(\boldsymbol{y}_{0,j}^* = l) \\ &= \frac{1}{k}\sum_{j=1}^{k}\sum_{i=1}^{n} P_{\widehat{F}}(\boldsymbol{t}_{0,j}^* = \boldsymbol{t}_{0,i})\Psi(\boldsymbol{y}_{0,i} = l). \end{aligned} \tag{4}$$

Rearranging equation (4) shows that $\tau_l^r(t_0)$ is a weighted average over all n neighbors where the weight assigned to $\boldsymbol{t}_{0,i}$ is $w_{i,k}^r = \sum_{j=1}^{k} P_{\widehat{F}}(\boldsymbol{t}_{0,j}^* = \boldsymbol{t}_{0,i})/k$. In comparison, the conventional k-NN estimator assigns the weight $\Psi(i \le k)/k$ to $\boldsymbol{t}_{0,i}$. The resampling weighted k-NN classifier predicts t_0 to belong to the group with the largest resampling expectation $\tau_l^r(t_0)$.

To calculate $P_{\widehat{F}}(\boldsymbol{t}_{0,j}^* = \boldsymbol{t}_{0,i})$, note that $\boldsymbol{t}_{0,j}^* = \boldsymbol{t}_{0,i}$ will occur if $\boldsymbol{t}^* = \{\boldsymbol{t}_{0,1}^*, \ldots, \boldsymbol{t}_{0,n}^*\}$ contains at most $j-1$ copies of $t_{0,1}, \ldots, t_{0,i-1}$ and at least j copies of $t_{0,1}, \ldots, t_{0,i}$. If a ranges over the number of copies of $t_{0,1}, \ldots, t_{0,i-1}$ and is constrained to $\{0, \ldots, j-1\}$, and b ranges over the number of copies of $t_{0,i}$ while constrained $\{j-a, \ldots, n-a\}$, so that $\boldsymbol{t}_{0,j}^* = \boldsymbol{t}_{0,i}$, then there must be $n-a-b$ copies of $t_{0,i+1}, \ldots, t_{0,n}$ to fill out $\boldsymbol{t}^*$. Hence,

$$P_{\widehat{F}}(\boldsymbol{t}_{0,j}^* = t_{0,i}) = \frac{1}{n^n}\sum_{a=0}^{j-1}\sum_{b=j-a}^{n-a}\binom{n}{a}(i-1)^a\binom{n-a}{b}(n-i)^{n-a-b}.$$

To illustrate the behavior of this weighting scheme, Table 1 shows $P_{\widehat{F}}(\boldsymbol{t}_{0,j}^* = t_{0,i})$ for $i \le 10$ and $j \le 5$ and the resampling weights $w_{1,5}^r, \ldots, w_{10,5}^r$ for the resampling-weighted 5-NN classifier when $n = 150$. In general, the resampling k-NN classifier places greater weight on the neighbors closest to t_0

and extends non-zero weights beyond the first k neighbors. Finally, there does not appear to be a practical method of computing the ideal bootstrap estimate of expected prediction error of $\tau_l^r(t_0)$.

Another resampling estimator of $\tau_l(t_0)$ can be constructed by selecting bootstrap samples $\boldsymbol{x}^{*b}$, $b = 1, \ldots, B$ from $\boldsymbol{x}$ and classifying t_0 using each of B k-NN classification rules $\eta_{\boldsymbol{x}^{*b}}$. The proportion of times that t_0 is predicted to belong to G_l is an estimate of $P_{\widehat{F}}\{\eta_{\boldsymbol{x}^*}(t_0) = l\}$, and also of $\tau_l(t_0)$. Furthermore, an analytic expression for this probability can be derived by enumerating and calculating the probabilities of events $B_{h,\ldots,l}$, $1 \leq h \leq \cdots \leq l \leq n$, defined in Section 4. However, the usefulness of this estimator, which we call the resampling k-set estimator, is greatly limited by the effort of computing these probabilities.

We used one of the simulated data sets discussed in the following section to compare the root mean square error (RMSE) of the conventional k-NN, resampling-weighted k-NN, and resampling k-set estimators of expected prediction error. Because the simulated data were generated from a mixture of normals, the exact posterior probabilities and the RMSE's of the estimators can be calculated. In particular, the RMSE of the conventional 3-NN, resampling-weighted 3-NN, and resampling 3-set estimators were .060, .043, and .051, respectively. For $k = 5$, the corresponding RMSE's were .042, .031, and .052. The optimal expected prediction error (McLachlan 1992, Chap. 3) was 0.322, and the sample means of expected prediction error were .383, .379, and .380, for the conventional 3-NN, resampling-weighted 3-NN, and resampling 3-setclassifiers, respectively. For $k = 5$, the corresponding sample means were .379, .369, and .370. Based on these comparisons and computational effort, we conclude that resampling k-set estimator is not useful, and the remainder of this article concentrates on the resampling-weighted k-NN classifier.

5.2. A Comparison of Weighted and Unweighted k-NN Classifiers

Weighted and conventional k-NN classifiers were compared by replicating the simulation study used by Bailey and Jain (1978), Dudani (1976), and Macleod et al. (1987). The simulated data consist 3 sets of 50 observations randomly sampled, respectively, from the bivariate distributions: $\frac{2}{5}N(\mu_{11}, \sigma_1 \boldsymbol{I}_2) + \frac{3}{5}N(\mu_{12}, \sigma_1 \boldsymbol{I}_2)$, $N(\mu_2, \sigma_2 \boldsymbol{I}_2)$, and $N(\mu_3, \sigma_3 \boldsymbol{I}_2)$ where $\mu_{11}^T = (3,3)$, $\sigma_1 = 1.5$,

$\mu_{12}^T = (7,7)$, $\mu_2^T = (4,6)$, $\sigma_2 = 2$, $\mu_3^T = (7.5,3.5)$, and $\sigma_3 = 3$. Six training sets of 150 observations and a test set of 3000 observations were generated. Each training set was used to classify the test set. The estimated expected prediction error for a classifier was the proportion of incorrectly classified test observations.

We used two conventional k-NN classifiers: one which broke ties randomly and one which broke ties by increasing k. The resampling k-NN classifier and the distance-weighted classifiers proposed by Dudani (1976) and Macleod et al. (1987) were also used. We used the version of Macleod et al.'s (1987) distance-weighted classifier which appeared to perform best in their simulation study (Macleod et al. 1987). The distance weights were $w_{j,n}^m = (d_n - d_j)/(d_n - d_1)$. Figure 1 shows only small differences in expected prediction error among the weighted k-NN classifiers and poorer performance by the conventional k-NN classifier over the range $k = 1,\ldots,20$.

- Figure 1 about here -

6. POLYGON CLASSIFICATION USING REMOTELY SENSED DATA

The performance of the k-NN classifiers is illustrated by a training set provided by the Wildlife Spatial Analysis Laboratory, Montana Cooperative Wildlife Research Unit, University of Montana. The training set consisted of 3184 observations on reflectance intensity for 7 spectral bands sampled by the Landsat Thematic Mapper (TM) satellite, and elevation, and had been used to classify TM scene P41/R28, an area of 3.5 million hectares of remote, rugged terrain in western Montana and northern Idaho. The purpose of classification was to construct a map predicting land cover type for a set of contiguous polygons covering the TM scene. Fifteen land cover types were identified; eight were forest types, two each were shrub and grassland types, and one type each of subalpine meadow, riparian shrub, and barren ground.

We compared the conventional k-NN classifier with ties broken randomly, the conventional k-NN classifier with ties broken by increasing k, the resampling k-NN classifier, the distance-weighted k-NN classifiers proposed by Dudani (1976) and Macleod et al. (1987), and the linear discriminant classifier. Leave-one-out Monte Carlo bootstrap estimates of expected prediction error were computed from 50

bootstrap samples. Figure 2 plots leave-one-out bootstrap estimates of expected prediction error against k over the range $k = 1, \ldots, 20$. Classifier performance is similar to the simulation results of Section 5.2 as the expected prediction error estimates are smallest for the resampling-weighted k-NN classifier for fixed k, with relatively good results from the conventional classifier with ties broken by increasing k, and Macleod et al.'s (1987) distance-weighted k-NN classifier. The leave-one-out Monte Carlo bootstrap estimate of expected prediction error for the linear discriminant classifier was $.477$, substantially worse than almost all of the k-NN estimates.

- Figure 2 about here -

7. CONCLUSION

We have derived analytic expressions for the ideal bootstrap estimate of expected prediction error for unweighted and distance-weighted k-NN classifiers. For values of k less than 5, analytic calculation of $\mathrm{Err}^{(1)}$ is a feasible alternative to the Monte Carlo bootstrap approximation. For values of k greater than 5, Monte Carlo approximation usually will be preferable because of excessive computational demands.

Analytic formulae for the ideal bootstrap estimate of expected prediction error motivated the resampling-weighted k-NN classifier. This classifier is similar to the combination classifiers discussed by LeBlanc and Tibshirani (1996) and Mojirsheibani (1999), among others, because the resampling-weighted k-NN classifier combines the indicator functions $\Psi(\boldsymbol{y}_{0,j} = l), j = 1, \ldots, n$, which are elements of the $k = 1, \ldots, n$ conventional k-NN classifiers. However, inspection of equation (4) reveals that the resampling-weighted k-NN estimator of $\tau_l(t_0)$ cannot be expressed explicitly as a combination of the conventional k-NN classifiers.

The resampling ideas motivated by ideal bootstrap estimate of expected prediction error can be exploited for other purposes. For example, an l-estimator of the population mean can be constructed by calculating the resampling expectation of the sample median. We have not yet examined the properties these estimators.

ACKNOWLEDGMENTS

We are grateful for the valuable comments and suggestions provided by the referees. We thank the Wildlife Spatial Analysis Laboratory, Montana Cooperative Wildlife Research Unit, The University of Montana, Missoula MT, for use of the P41/R28 training data set.

REFERENCES

Bailey, T. and Jain, A.K. (1978) A note on distance-weighted k-nearest neighbor rules. *IEEE Transactions on Systems, Man and Cybernetics*, **8**, 311-313.

Dudani, S.A. (1976) The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man and Cybernetics*, **6**, 325-327.

Efron, B. (1982) *The Jackknife, the Bootstrap, and Other Resampling Plans*, Volume 38 of CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM.

Efron, B. (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, **78**, 316-331.

Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap*, Chapman and Hall, London.

Efron, B. and Tibshirani, R. (1997) Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, **92**, 548-560.

LeBlanc, M. and Tibshirani, R. (1996) Combining estimates in regression and classification. *Journal of the American Statistical Association*, **91**, 1641-1658.

Macleod, J.E.S., Luk, A., and Titterington, D.M. (1987) A re-examination of the distance-weighted k-nearest-neighbor classification rule. *IEEE Transactions on Systems, Man and Cybernetics*, **17**, 689-696.

Mojirsheibani, M. (1999) Combining classifiers via discretization. *Journal of the American Statistical Association*, **94**, 600-609.

APPENDIX: FORMULAE FOR COMPUTING RESAMPLING PROBABILITIES

Let $m = n - 1$ where n is the number of observations in $\boldsymbol{x}_{(i)}$, and let Ω denote the sample space generated by sampling $\boldsymbol{x}_{(i)}$ with replacement. We can calculate the probability of $B_{h,j,l} = \{\boldsymbol{x}^*_{(i)} \in \Omega \mid (\boldsymbol{t}^*_{0,1}, \boldsymbol{t}^*_{0,2}, \boldsymbol{t}^*_{0,3}) = (t_{0,h}, t_{0,j}, t_{0,l})\}$ separately for each of the cases $h = j = l$, $h = j < l$, $h < j = l$, and $h < j < l$ as follows. For $h = j = l$, the probability of $B_{h,j,l}$ is the sum of the probabilities of getting a copies of $t_{0,h}$ and $m - a$ observations from $\{t_{0,h+1}, \ldots, t_{0,m}\}$ where $3 \leq a \leq m$. Hence,

$$P_{\widehat{F}_{(i)}}\left(B_{h,j,l} \mid h = j = l\right) = m^{-m} \sum_{a=3}^{m} \binom{m}{a} (m-h)^{m-a}.$$

For $h = j < l$, $B_{h,j,l}$ will occur if and only if exactly two copies of $t_{0,h}$ appear in $\boldsymbol{x}^*_{(i)}$, $1 \leq a \leq m - 2$ copies of $t_{0,l}$ appear in $\boldsymbol{x}^*_{(i)}$, and the remaining $m - a - 2$ observations are selected from $\{t_{0,h+1}, \ldots, t_{0,m}\}$. Hence,

$$\begin{aligned} P_{\widehat{F}_{(i)}}\left(B_{h,j,l} \mid h = j < l\right) &= m^{-2} \binom{m}{2} \left(\frac{m-h}{m}\right)^{m-2} \sum_{a=1}^{m-2} \binom{m-2}{a} \left(\frac{1}{m-h}\right)^{a} \left(\frac{m-l}{m-h}\right)^{m-a-2} \\ &= m^{-m} \binom{m}{2} \sum_{a=1}^{m-2} \binom{m-2}{a} (m-l)^{m-a-2}. \end{aligned}$$

Similarly, for $h < j = l$,

$$\begin{aligned} P_{\widehat{F}_{(i)}}\left(B_{h,j,l} \mid h < j = l\right) &= m^{-1} \binom{m}{1} \left(\frac{m-h}{m}\right)^{m-1} \sum_{a=2}^{m-1} \binom{m-1}{a} \left(\frac{1}{m-h}\right)^{a} \left(\frac{m-j}{m-h}\right)^{m-a-1} \\ &= m^{-(m-1)} \sum_{a=2}^{m-1} \binom{m-1}{a} (m-j)^{m-a-1}. \end{aligned}$$

After some reduction,

$$P_{\widehat{F}_{(i)}}\left(B_{h,j,l} \mid h < j < l\right) = \frac{m-1}{m^{(m-1)}} \sum_{a=1}^{m-2} \binom{m-2}{a} (m-l)^{m-a-2}.$$

Table 1. Resampling probabilities $P(\boldsymbol{t}^*_{0,j} = \boldsymbol{t}_{0,i})$ for $j \le k = 5$, $i = 1, \ldots, 10$, and $n = 150$. Also shown are the resampling weights $w^r_{i,5} = \sum_{j=1}^{5} P_{\widehat{F}}(\boldsymbol{t}^*_{0,j} = \boldsymbol{t}_{0,i})/5$.

	j					
i	1	2	3	4	5	$w^r_{i,5}$
1	.633	.264	.080	.019	.004	.200
2	.233	.332	.244	.123	.048	.196
3	.085	.208	.256	.211	.132	.178
4	.031	.108	.187	.216	.188	.146
5	.011	.050	.114	.171	.191	.107
6	.004	.022	.062	.114	.158	.072
7	.001	.009	.031	.069	.113	.045
8	.001	.004	.015	.038	.073	.026
9	.000	.002	.007	.020	.044	.014
10	.000	.001	.003	.010	.024	.008

FIGURE LEGENDS

Figure 1. Estimated expected prediction error for five k-NN classifiers. Six training data sets consisting of 150 observations were sampled from a mixture of three normal distributions. The test data consisted of 3000 observations independently sampled from the same mixture distribution.

Figure 2. Estimated expected prediction error for five k-NN classifiers, $n = 3184$ and $g = 15$.