Improvements on Cross-Validation: The .632+ Bootstrap Method

Bradley EFRON and Robert TIBSHIRANI

A training set of data has been used to construct a rule for predicting future responses. What is the error rate of this rule? This is an important question both for comparing models and for assessing a final selected model. The traditional answer to this question is given by cross-validation. The cross-validation estimate of prediction error is nearly unbiased but can be highly variable. Here we discuss bootstrap estimates of prediction error, which can be thought of as smoothed versions of cross-validation. We show that a particular bootstrap method, the .632+ rule, substantially outperforms cross-validation in a catalog of 24 simulation experiments. Besides providing point estimates, we also consider estimating the variability of an error rate estimate. All of the results here are nonparametric and apply to any possible prediction rule; however, we study only classification problems with 0–1 loss in detail. Our simulations include "smooth" prediction rules like Fisher's linear discriminant function and unsmooth ones like nearest neighbors.

KEY WORDS: Classification; Cross-validation bootstrap; Prediction rule.

1. INTRODUCTION

This article concerns estimating the error rate of a prediction rule constructed from a training set of data. The training set $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ consists of n observations $x_i = (\mathbf{t}_i, y_i)$, with $\mathbf{t}_i$ the predictor or feature vector and y_i the response. On the basis of $\mathbf{x}$, the statistician constructs a prediction rule $r_{\mathbf{x}}(\mathbf{t})$ and wishes to estimate the error rate of this rule when it is used to predict future responses from their predictor vectors.

Cross-validation, the traditional method of choice for this problem, provides a nearly unbiased estimate of the future error rate. However, the low bias of cross-validation is often paid for by high variability. Here we show that suitably defined bootstrap procedures can substantially reduce the variability of error rate predictions. The gain in efficiency in our catalog of simulations is roughly equivalent to a 60% increase in the size of the training set. The bootstrap procedures are nothing more than smoothed versions of cross-validation, with some adjustments made to correct for bias.

We are interested mainly in the situation when the response is dichotomous. This is illustrated in Figure 1, where the $n = 20$ observations, $x_i = (\mathbf{t}_i, y_i)$, in the training set $\mathbf{x}$ each consist of a bivariate feature vector $\mathbf{t}_i$ and a $0 - 1$ response y_i; 12 of the points are labeled 0 and 8 are labelled 1. Two different prediction rules are indicated. Figure 1a shows the prediction rule based on Fisher's linear discriminant function (LDF), following Efron (1983). The rule $r_{\mathbf{x}}(\mathbf{t})$ will predict $y = 0$ if $\mathbf{t}$ lies to the lower left of the LDF boundary and will predict $y = 1$ if $\mathbf{t}$ lies to the upper right of the LDF boundary. Figure 1b shows the nearest-neighbor (NN) rule, in which future $\mathbf{t}$ vectors will have y predicted according to the label of the nearest observation in the training set. We wish to estimate the error rates of the two prediction rules.

The data shown in Figure 1 were generated as part of the extensive simulation experiments described in Section 4. In this case the y_i were selected randomly and the $\mathbf{t}_i$ were bivariate normal vectors whose means depended on y_i,

$$y_i = \begin{cases} 0 & \Pr \frac{1}{2} \\ 1 & \Pr \frac{1}{2} \end{cases} \quad \text{and} \quad \mathbf{t}_i | y_i \sim N_2\left(\begin{pmatrix} y_i - \frac{1}{2} \\ 0 \end{pmatrix}, I \right), \quad (1)$$

independently for $i = 1, 2, \ldots, n = 20$.

Table 1 shows results from the simulations. Cross-validation is compared to the bootstrap-based estimator 632+ described in Section 3. Cross-validation is nearly unbiased as an estimator of the true error rate for both rules LDF and NN, but the bootstrap-based estimator has a root mean squared (RMS) error only 80% as large. These results are fairly typical of the 24 simulation experiments reported in Section 4. The bootstrap estimator in these experiments was run with only 50 bootstrap replications per training set, but this turns out to be sufficient for most purposes, as the *internal variance* calculations of Section 2 show.

The bootstrap has other important advantages besides providing more accurate point estimates for prediction error. The bootstrap replications also provide a direct assessment of variability for estimated parameters in the prediction rule. For example, Efron and Gong (1983) discussed the stability of the "significant" predictor variables chosen by a complicated stepwise logistic regression program. Section 5 provides another use for the bootstrap replications: to estimate the *variance* of a point estimate of prediction error.

Section 2 begins with a discussion of *bootstrap smoothing,* a general approach to reducing the variability of nonparametric point estimators. When applied to the prediction problem, bootstrap smoothing gives a smoothed version of cross-validation with considerably reduced variability but with an upward bias. A bias correction discussed in Section 3 results in the .632+ estimates of Table 1. The .632+ estimator is shown to substantially outperform ordinary cross-validation in the catalog of 24 sampling experi-

Bradley Efron is Professor, Department of Statistics, Stanford University, Stanford, CA 94305. Robert Tibshirani is Professor, Department of Preventive Medicine and Biostatistics, University of Toronto, Toronto, Ontario M5S 1A8, Canada. The authors thank Ronny Kohavi for reviving their interest in this problem, Jerry Friedman, Trevor Hastie, and Richard Olshen for enjoyable and fruitful discussions, and the associate editor and referee for helpful comments. Tibshirani was supported by the Natural Sciences and Engineering Research Council of Canada.

Journal of the American Statistical Association
June 1997, Vol. 92, No. 438, Theory and Methods

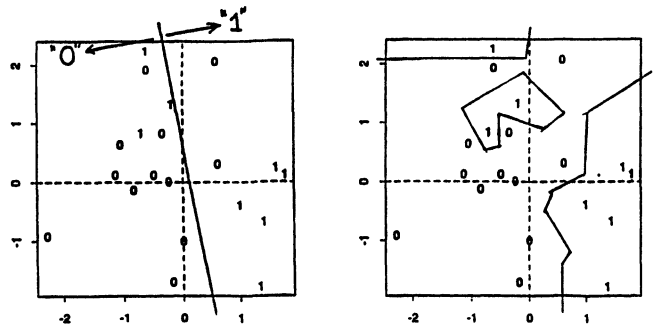

Figure 1. A Training Set Consisting of n = 20 Observations, 12 Labelled 0 and 8 Labelled 1. (a) The linear discriminant function predicts 0 to lower left of the solid line and 1 to upper right; (b) the nearest-neighbor rule predicts 1 in the three indicated islands and 0 elsewhere.

ments described in Section 4. Section 5 shows how the same bootstrap replications that provide a point estimate of prediction error can also provide an assessment of variability for that estimate. Section 6 presents the distance argument underlying the .632+ rule, along with other bias-correction techniques.

All of the results here are nonparametric and apply to any possible prediction rule; however, we study only classification problems with 0–1 loss in detail. Regression problems may exhibit qualitatively different behavior, and the statistical approach may also differ. "In-sample" prediction error is often the focus in regression, especially for model selection. In contrast, the error that we study here might be called "extra-sample" error. Efron (1986) studied estimates of the in-sample prediction error problem, including generalized cross-validation (Wahba 1980) and the Cp statistic of Mallows (1973). The note at the end of Section 2 clarifies the distinction between extra-sample and in-sample prediction errors.

Considerable work has been done in the literature on cross-validation and the bootstrap for error rate estimation. (A good general discussion can be found in McLachlan 1992; key references for cross-validation are Allen 1974 and Stone 1974, 1977.) Efron (1983) proposed a number of bootstrap estimates of prediction error, including the optimism and .632 estimates. The use of cross-validation and the bootstrap for model selection was studied by Breiman (1992), Breiman and Spector (1992), Shao (1993), and Zhang (1993). Breiman and Spector demonstrated that leave-one-out cross-validation has high variance if the prediction rule is unstable, because the leave-one-out training sets are too similar to the full dataset. Fivefold or tenfold cross-validation displayed lower variance in this case. A study of cross-validation and bootstrap methods for tree-structured models was carried out by Crawford (1989). Substantial work has also been done on the prediction error problem in the machine learning and pattern recognition fields; see, for example, the simulation studies of Chernick, Murthy, and Nealy (1985, 1986) and Jain, Dubes, and Chen (1987). Kohavi (1995) performed a particularly interesting study that renewed our interest in this problem.

2. CROSS-VALIDATION AND THE LEAVE-ONE-OUT BOOTSTRAP

This section discusses a bootstrap smoothing of cross-validation that reduces the variability of error-rate estimates. Here the notation $Q[y, r]$ indicates the discrepancy between a predicted value r and the actual response y. We are particularly interested in the dichotomous situation where both y and r are either 0 or 1, with

$$Q[y, r] = \begin{cases} 0 & \text{if } r = y \\ 1 & \text{if } r \neq y. \end{cases} \tag{2}$$

We also use the shorter notation

$$Q(x_0, \mathbf{x}) = Q[y_0, r_{\mathbf{x}}(\mathbf{t}_0)] \tag{3}$$

to indicate the discrepancy between the predicted value and response for a test point $x_0 = (\mathbf{t}_0, y_0)$ when using the rule $r_{\mathbf{x}}$ based on training set $\mathbf{x}$.

Suppose that the observations $x_i = (\mathbf{t}_i, y_i)$ in the training set are a random sample from some distribution F,

$$x_1, x_2, \ldots, x_n \overset{\text{iid}}{\sim} F, \tag{4}$$

and that $x_0 = (\mathbf{t}_0, y_0)$ is another independent draw from F, called a test point. The *true error rate* of the rule $r_{\mathbf{x}}$ is

$$\text{Err} = \text{Err}(\mathbf{x}, F) = E_{0F} Q(x_0, \mathbf{x}) = E_{0F} Q[y_0, r_{\mathbf{x}}(\mathbf{t}_0)], \tag{5}$$

with the notation E_{0F} indicating that only $x_0 = (\mathbf{t}_0, y_0)$ is random in (5), with $\mathbf{x}$ and $r_{\mathbf{x}}$ being fixed. Thus Err is the conditional error rate, conditional on the training set $\mathbf{x}$.

We compare error rate estimators in terms of their ability to predict Err. Section 4 briefly discusses estimating instead the *expected true error*,

$$\mu = \mu(F) = E_F\{\text{Err}\} = E_F E_{0F} Q(x_0, \mathbf{x}). \tag{6}$$

The results in this case are somewhat more favorable to the bootstrap estimator. Note, however, that although the conditional error rate is often what we would like to obtain, none of the methods correlates very well with it on a sample-by-sample basis (see Zhang 1995).

The *apparent error rate* (or *resubstitution rate*) is

$$\overline{\text{err}} = \text{Err}(\mathbf{x}, \hat{F}) = E_{0\hat{F}} Q(x_0, \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} Q[y_i, r_{\mathbf{x}}(\mathbf{t})], \tag{7}$$

with $\hat{F}$ indicating the *empirical distribution* that puts probability $1/n$ on each observation $x_1, x_2, \ldots x_n$; $\overline{\text{err}}$ tends to be

Table 1. Error Rate Estimation for Situation (1)

	LDF		NN	
	Exp	RMS	Exp	RMS
CV1	.362	.123	.419	.123
.632+	.357	.096	.380	.099
True	.357		.418	

NOTE: CV1 is the cross-validation estimate based on omitting one observation at a time from the training set; 632+ is the bootstrap-based estimator described in Section 3. The table shows the expectation and RMS error of the two estimates for both the LDF and NN prediction rules. In both cases, RMS(.632+)/RMS(CV1) is about 80%.

biased downward as an estimate of Err because the training set $\mathbf{x}$ has been used twice, both to construct the rule and to test it.

Cross-validation (Geisser 1975; Stone 1974) avoids this problem by removing the data point to be predicted from the training set. The ordinary cross-validation estimate of prediction error is

$$\widehat{\mathrm{Err}}^{(\mathrm{cv1})} = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, r_{\mathbf{x}_{(i)}}(\mathbf{t})] = \frac{1}{n}\sum_{i=1}^{n} Q(x_i, \mathbf{x}_{(i)}), \quad (8)$$

where $\mathbf{x}_{(i)}$ is the training set with the ith observation removed. $\widehat{\mathrm{Err}}^{(\mathrm{cv1})}$ is *leave-one-out* cross-validation; the k-fold version $\widehat{\mathrm{Err}}^{(\mathrm{cv}k)}$ partitions the training set into k parts, predicting in turn the observations in each part from the training sample formed from all of the remaining parts.

The statistic $\widehat{\mathrm{Err}}^{(\mathrm{cv1})}$ is a discontinuous function of the training set $\mathbf{x}$ when $Q[y, r]$ itself is discontinuous as in (2). *Bootstrap smoothing* is a way to reduce the variance of such functions by averaging. Suppose that $Z(\mathbf{x})$ is an unbiased estimate of a parameter of interest, say

$$\zeta(F) = E_F\{Z(\mathbf{x})\}. \quad (9)$$

By definition, the nonparametric maximum likelihood estimate (MLE) of the same parameter ζ is

$$\hat{\zeta} = \zeta(\hat{F}) = E_{\hat{F}}\{Z(\mathbf{x}^*)\}. \quad (10)$$

Here $\mathbf{x}^*$ is a random sample from $\hat{F}$,

$$x_1^*, x_2^*, \ldots, x_n^* \overset{\text{iid}}{\sim} \hat{F}; \quad (11)$$

that is, a bootstrap sample. The bootstrap expectation in (10) smooths out discontinuities in $Z(\mathbf{x})$, usually reducing its variability. However, $\hat{\zeta}$ may now be biased as an estimate of ζ. Breiman (1994) introduced a very similar idea under the sobriquet "bagging."

Now consider applying bootstrap smoothing to $Z_i(\mathbf{x}) = Q(x_i, \mathbf{x}_{(i)})$, with x_i fixed. The nonparametric MLE of $E_F Z_i(\mathbf{x})$ is $\hat{\zeta}_i = E_{\hat{F}_{(i)}}\{Q(x_i, \mathbf{x}^*_{(i)})\}$, where $\mathbf{x}^*_{(i)}$ is a bootstrap sample from the empirical distribution on $\mathbf{x}_{(i)}$,

$$\hat{F}_{(i)} : \Pr \frac{1}{n-1} \text{ on } x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n. \quad (12)$$

It might be argued that the bootstrap samples $\mathbf{x}^*_{(i)}$ should be of size $n-1$ instead of n, but there is no advantage to this. In what follows we take bootstrap samples from $\hat{F}_{(i)}$ to be of size n and indicate them by $\mathbf{x}^*$ rather than $\mathbf{x}^*_{(i)}$, so

$$\hat{\zeta}_i = E_{\hat{F}_{(i)}}\{Q(x_i, \mathbf{x}^*)\}. \quad (13)$$

Notice that an $\mathbf{x}^*$ sample drawn from $\hat{F}_{(i)}$ never contains the point x_i.

Applying (11) and (12) to each case i in turn leads to the *leave-one-out bootstrap*,

$$\widehat{\mathrm{Err}}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} E_{\hat{F}_{(i)}}\{Q(x_i, \mathbf{x}^*)\}, \quad (14)$$

a smoothed version of $\widehat{\mathrm{Err}}^{(\mathrm{cv1})}$. This estimate predicts the error at point i only from bootstrap samples that do not contain the point i.

The actual calculation of $\widehat{\mathrm{Err}}^{(1)}$ is a straightforward bootstrap exercise. Ordinary bootstrap samples $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ are generated as in (11), so $\mathbf{x}^*$ is a random draw of size n, with replacement, from $\{x_1, x_2, \ldots, x_n\}$. A total of B such samples are independently drawn, say $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$, with $B = 50$ in our simulations, as discussed later. Let N_i^b be the number of times that x_i is included in the bth bootstrap sample and define

$$I_i^b = \begin{cases} 1 & \text{if } N_i^b = 0 \\ 0 & \text{if } N_i^b > 0. \end{cases} \quad (15)$$

Also define

$$Q_i^b = Q(x_i, \mathbf{x}^{*b}) = Q[y_i, r_{\mathbf{x}^{*b}}(\mathbf{t}_i)]. \quad (16)$$

Then

$$\widehat{\mathrm{Err}}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{E}_i, \quad \text{where} \quad \hat{E}_i = \sum_b I_i^b Q_i^b \Big/ \sum_b I_i^b. \quad (17)$$

This definition agrees with (14) because a bootstrap sample that has $I_i^b = 1$ is the same as a bootstrap sample from $\hat{F}_{(i)}$ (see Efron 1992). A slightly different definition was given by Efron (1983) (where $\widehat{\mathrm{Err}}^{(1)}$ is $\hat{\varepsilon}^{(0)}$), namely $\sum_i \sum_b I_i^b Q_i^b / \sum_i \sum_b I_i^b$, but the two definitions agree as $B \to \infty$ and produced nearly the same results in our simulations.

There is another way to view cross-validation and $\widehat{\mathrm{Err}}^{(1)}$: as estimates of the average error $\mu(F)$. Direct application of the bootstrap gives the plug-in estimate $\mu(\hat{F}) = E_{\hat{F}} E_{0\hat{F}} Q(x_0, \mathbf{x})$. This estimate, discussed in Section 6, tends to be biased downward. The reason is that $\hat{F}$ is being used twice: as the population, say F_0, from which bootstrap training sets $\mathbf{x}^*$ are drawn, and as the population F_1 from which test points X_0 are drawn. Let us write $\mu(F)$ explicitly as a function of both F_1 and F_0:

$$\begin{aligned} \mu(F_1, F_0) &= E_{F_1}\{E_{0F_0} Q[Y_0, r_{\mathbf{x}}(T_0)]\} \\ &= E_{0F_0} E_{F_1} Q[Y_0, r_{\mathbf{x}}(T_0)]), \end{aligned} \quad (18)$$

where, for convenience, we have switched the order of expectation in the second expression. We assume that in the unknown true state of affairs, $F_1 = F_0 = F$. Plugging in $\hat{F}$ for the test distribution F_0 gives

$$\mu(F_1, \hat{F}) = \frac{1}{n}\sum_{i=1}^{n} E_{F_1} Q[Y_i, r_{\mathbf{x}}(T_i)]. \quad (19)$$

The remaining task is to estimate the training sample distribution F_1. Ideally, we would take $F_1 = F$. Notice that for continuous populations F, the probability of the test point $X_0 = x_i$ appearing in a training sample drawn from $F_1 = F$ is 0. The plug-in estimate $\mu(\hat{F}) = \mu(\hat{F}, \hat{F})$ uses $\hat{F}$ for F_1. With this choice, the probability that $X_0 = x_i$ appears in the training sample is $1 - (1 - 1/n)^n \approx .632$. Hence $\mu(\hat{F})$

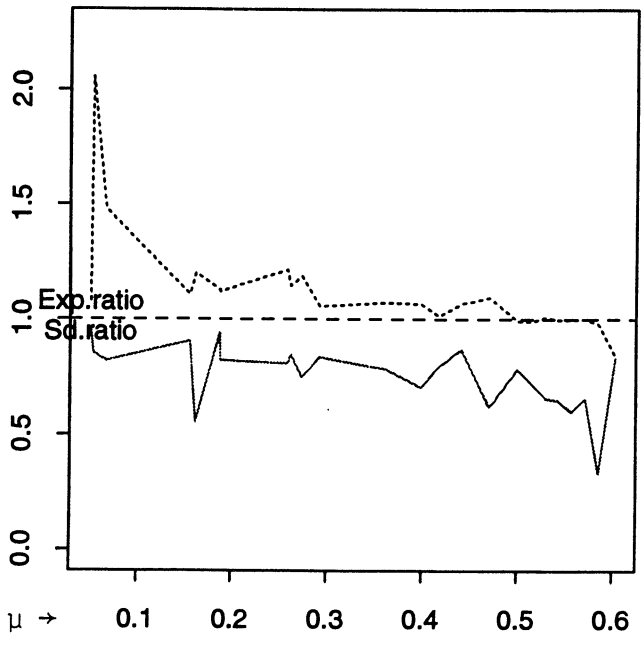

Figure 2. Ratio of Standard Deviations and Expectations for the Leave-One-Out Bootstrap $\widehat{\text{Err}}^{(1)}$ Compared to Cross-Validation $\widehat{\text{Err}}^{(\text{cv}1)}$ for the 24 sampling experiments described in Section 4, plotted versus expected true error μ, (6). The median SD ratio for the 24 experiments was .79; the median expectation ratio was 1.07.

uses training samples that are too close to the test points, leading to potential underestimation of the error rate. Cross-validation uses the leave-one-out training samples to ensure that the training samples do not contain the test point; that is, cross-validation estimates $E_{F_1}Q[Y_i, r_{\mathbf{x}}(T_i)]$ by

$$\hat{E}_{F_1}Q[Y_i, r_{\mathbf{x}}(T_i)] = Q[Y_i, r_{\mathbf{x}_{(i)}}(T_i)]. \tag{20}$$

On the other hand, using the estimate $\hat{F}_1 = \hat{F}_{(i)}$ in each term $E_{F_1}Q[Y_i, r_{\mathbf{x}}(T_i)]$ gives the leave-one-out bootstrap estimate $\widehat{\text{Err}}^{(1)}$.

The efficacy of bootstrap smoothing is shown in Figure 2, where the solid line plots the standard deviation ratio for the 24 sampling experiments of Section 4. The horizontal axis is the expected true error μ for each experiment, (6). We see that $\widehat{\text{Err}}^{(1)}$ always has smaller standard deviation than $\widehat{\text{Err}}^{(\text{cv}1)}$, with the median ratio over the 24 experiments being .79. Going from $\widehat{\text{Err}}^{(\text{cv}1)}$ to $\widehat{\text{Err}}^{(1)}$ is roughly equivalent to multiplying the size n of the training set by $1/.79^2 = 1.60$. The improvement of the bootstrap estimators over cross-validation is due mainly to the effect of smoothing. Cross-validation and the bootstrap are closely related, as Efron (1983, Sec. 2) has shown. In smoother prediction problems—for example, when y and r are continuous and $Q[y, r] = (y - r)^2$—we would expect to see little difference between $\widehat{\text{Err}}^{(\text{cv}1)}$ and $\widehat{\text{Err}}^{(1)}$.

The dotted curve in Figure 2 is the expectation ratio $E\{\widehat{\text{Err}}^{(1)}\}/E\{\widehat{\text{Err}}^{(\text{cv}1)}\}$. We see that $\widehat{\text{Err}}^{(1)}$ is biased *upward* relative to the nearly unbiased estimate $\widehat{\text{Err}}^{(\text{cv}1)}$. This is not surprising. Let μ_n indicate the expected true error (6) when $\mathbf{x}$ has sample size n. Ordinary cross-validation produces an unbiased estimate of μ_{n-1}, whereas k-fold cross-validation estimates μ_{n-k}. Because smaller training sets produce bigger prediction errors, larger k gives bigger upward bias $\mu_{n-k} - \mu_n$. The amount of bias depends on the slope of the error curve μ_n at sample size n. Bootstrap samples are typically supported on about $.632n$ of the original sample points, so we might expect $\widehat{\text{Err}}^{(1)}$ to be estimating $\mu_{.632n}$. The more precise calculations of Efron (1983, Sec. 8) show that $\widehat{\text{Err}}^{(1)}$ closely agrees with *half-sample cross-validation* (where $\mathbf{x}$ is repeatedly split into equal-sized training and test sets), and that the expectation of $\widehat{\text{Err}}^{(1)}$ is $\mu_{n/2}$ to second order. The next section concerns a bias-corrected version of $\widehat{\text{Err}}^{(1)}$ called the .632+ rule, that reduces the upward bias.

The choice of $B = 50$ bootstrap replications in our simulation experiments was based on an assessment of *internal error*, the Monte Carlo error due to using B instead of infinity replications (Efron 1992). The same bootstrap replications that give $\widehat{\text{Err}}^{(1)}$ also give a jackknife estimate of its internal error. Let $q_i^b = I_i^b Q_i^b$, $q_i^+ = \sum_{b=1}^B q_i^b$, and $I_i^+ = \sum_{b=1}^B I_i^+$. Then the estimate of $\widehat{\text{Err}}^{(1)}$ with the bth bootstrap replication removed is

$$\widehat{\text{Err}}_{(b)}^{(1)} = \frac{1}{n}\sum_{i=1}^n \hat{E}_{i(b)}, \quad \text{where} \quad \hat{E}_{i(b)} = \frac{q_i^+ - q_i^b}{I_i^+ - I_i^b}. \tag{21}$$

The jackknife estimate of internal standard deviation for $\widehat{\text{Err}}^{(1)}$ is then

$$\widehat{\text{SD}}_{\text{int}} = \left[\frac{B-1}{B}\sum_b (\widehat{\text{Err}}_{(b)}^{(1)} - \widehat{\text{Err}}_{(\cdot)}^{(1)})^2\right]^{1/2}, \tag{22}$$

$\widehat{\text{Err}}_{(\cdot)}^{(1)} = \sum_b \widehat{\text{Err}}_{(b)}^{(1)}/B$.

In our simulations $\widehat{\text{SD}}_{\text{int}}$ was typically about .02 for $B = 50$. The *external* standard deviation of $\widehat{\text{Err}}^{(1)}$ (i.e., the standard deviation due to randomness of $\mathbf{x}$) was typically .10. (Sec. 5 discusses the estimation of external error, also using the same set of bootstrap replications.) This gives corrected external standard deviation $[.10^2 - .02^2]^{1/2} = .098$, indicating that $B = 50$ is sufficient here.

Note that definition (5) of prediction error, $\text{Err} = E_{0F}Q[y_0, r_{\mathbf{x}}(x_0)]$, might be called *extra-sample error*, because the test point $(\mathbf{t}_0, y_0)$ is chosen randomly from F without reference to the training sample $\mathbf{x}$. Efron (1986) investigated a more restrictive definition of prediction error. For dichotomous problems, let $\pi(\mathbf{t}) = \Pr_F\{y = 1|\mathbf{t}\}$ and $\pi_i = \pi(\mathbf{t}_i)$. The *in-sample error* of a rule $r_{\mathbf{x}}$ is defined to be

$$\text{err} = \frac{1}{n}\sum_{i=1}^n E_{0\pi_i}\{Q[y_{0i}, r_{\mathbf{x}}(\mathbf{t}_i)]\}, \tag{23}$$

where the notation $E_{0\pi_i}$ indicates that only $y_{0i} \sim$ binomial $(1, \pi_i)$ is random, with $\mathbf{x}$ and $r_{\mathbf{x}}(\mathbf{t}_i)$ being fixed. This situation is similar to a standard regression problem in that the predictors $\mathbf{t}_i$ are treated as fixed at their observed values, rather than as random. In-sample error prediction is mathematically simpler than the extra-sample case and leads to quite different solutions for the error rate prediction problem (see Efron 1986).

3. THE .632+ ESTIMATOR

Efron (1983) proposed the *.632 estimator*,

$$\widehat{\mathrm{Err}}^{(.632)} = .368 \cdot \overline{\mathrm{err}} + .632 \cdot \widehat{\mathrm{Err}}^{(1)}, \tag{24}$$

designed to correct the upward bias in $\widehat{\mathrm{Err}}^{(1)}$ by averaging it with the downwardly biased estimate $\widehat{\mathrm{Err}}^{(.632)}$. The coefficients $.368 = e^{-1}$ and .632 were suggested by an argument based on the fact that bootstrap samples are supported on approximately $.632n$ of the original data points. In Efron's article $\widehat{\mathrm{Err}}^{(.632)}$ performed better than all competitors, but the simulation studies did not include highly overfit rules like nearest-neighbors, where $\overline{\mathrm{err}} = 0$. Such statistics make $\widehat{\mathrm{Err}}^{(.632)}$ itself downwardly biased. For example, if y equals 0 or 1 with probability 1/2, independently of the (useless) predictor vector t, then Err = .50 for any prediction rule, but the expected value of $\widehat{\mathrm{Err}}^{(.632)}$ for the nearest-neighbor rule is $.632 \cdot .5 = .316$. Both $\widehat{\mathrm{Err}}^{(1)}$ and $\widehat{\mathrm{Err}}^{(\mathrm{cv1})}$ have the correct expectation .50 in this case. Breiman, Friedman, Olshen, and Stone (1984) suggested this example.

This section proposes a new estimator $\widehat{\mathrm{Err}}^{(.632+)}$, designed to be a less-biased compromise between $\overline{\mathrm{err}}$ and $\widehat{\mathrm{Err}}^{(1)}$. The .632+ rule puts greater weight on $\widehat{\mathrm{Err}}^{(1)}$ in situations where the amount of overfitting, as measured by $\widehat{\mathrm{Err}}^{(1)} - \overline{\mathrm{err}}$, is large. To correctly scale the amount of overfitting, we first need to define the *no-information error rate*, γ, that would apply if $\mathbf{t}$ and y were independent, as in the example of the previous paragraph.

Let F_{ind} be the probability distribution on points $x = (\mathbf{t}, y)$ having the same $\mathbf{t}$ and y marginals as F, but with y independent of $\mathbf{t}$. As in (5), define

$$\gamma = E_{0F_{\mathrm{ind}}} Q(x_0, \mathbf{x}) = E_{0F_{\mathrm{ind}}} Q[y_0, r_{\mathbf{x}}(\mathbf{t}_0)], \tag{25}$$

the expected prediction error for rule $r_{\mathbf{x}}$ given a test point $x_0 = (\mathbf{t}_0, y_0)$ from F_{ind}. An estimate of γ is obtained by permuting the responses y_i and predictors $\mathbf{t}_j$,

$$\hat{\gamma} = \sum_{i=1}^{n} \sum_{j=1}^{n} Q[y_i, r_{\mathbf{x}}(\mathbf{t}_j)]/n^2. \tag{26}$$

For the dichotomous classification problem (2), let $\hat{p}_1$ be the observed proportion of responses y_i equalling 1, and let $\hat{q}_1$ be the observed proportion of predictions $r_{\mathbf{x}}(\mathbf{t}_j)$ equalling 1. Then

$$\hat{\gamma} = \hat{p}_1(1 - \hat{q}_1) + (1 - \hat{p}_1)\hat{q}_1. \tag{27}$$

With a rule like nearest-neighbors for which $\hat{q}_1 = \hat{p}_1$, the value of $\hat{\gamma}$ is $2\hat{p}_1(1 - \hat{p}_1)$. The multicategory generalization of (27) is $\hat{\gamma} = \sum_l \hat{p}_l(1 - \hat{q}_l)$.

The *relative overfitting rate* is defined as

$$\hat{R} = \frac{\widehat{\mathrm{Err}}^{(1)} - \overline{\mathrm{err}}}{\hat{\gamma} - \overline{\mathrm{err}}}, \tag{28}$$

a quantity that ranges from 0 with no overfitting ($\widehat{\mathrm{Err}}^{(1)} = \overline{\mathrm{err}}$) to 1 with overfitting equal to the no-information value $\hat{\gamma} - \overline{\mathrm{err}}$. The "distance" argument of Section 6 suggests a less-biased version of (24) where the weights on $\overline{\mathrm{err}}$ and $\widehat{\mathrm{Err}}^{(1)}$ depend on $\hat{R}$,

$$\widehat{\mathrm{Err}}^{(.632+)} = (1 - \hat{w}) \cdot \overline{\mathrm{err}} + \hat{w} \cdot \widehat{\mathrm{Err}}^{(1)},$$

$$\left[\hat{w} = \frac{.632}{1 - .368\hat{R}}\right]. \tag{29}$$

The weight w ranges from .632 if $\hat{R} = 0$ to 1 if $\hat{R} = 1$, so $\widehat{\mathrm{Err}}^{(.632+)}$ ranges from $\widehat{\mathrm{Err}}^{(.632)}$ to $\widehat{\mathrm{Err}}^{(1)}$. We can also express (29) as

$$\widehat{\mathrm{Err}}^{(.632+)} = \widehat{\mathrm{Err}}^{(.632)} + (\widehat{\mathrm{Err}}^{(1)} - \overline{\mathrm{err}}) \frac{.368 \cdot .632 \cdot \hat{R}}{1 - .368\hat{R}}, \tag{30}$$

emphasizing that $\widehat{\mathrm{Err}}^{(.632+)}$ exceeds $\widehat{\mathrm{Err}}^{(.632)}$ by an amount depending on $\hat{R}$.

It may happen that $\hat{\gamma} \leq \overline{\mathrm{err}}$ or $\overline{\mathrm{err}} < \hat{\gamma} \leq \widehat{\mathrm{Err}}^{(1)}$, in which case $\hat{R}$ can fall outside of [0, 1]. To account for this possibility, we modify the definitions of $\widehat{\mathrm{Err}}^{(1)}$ and $\hat{R}$:

$$\widehat{\mathrm{Err}}^{(1)'} = \min(\widehat{\mathrm{Err}}^{(1)}, \hat{\gamma})$$

and

$$\hat{R}' = \begin{cases} (\widehat{\mathrm{Err}}^{(1)} - \overline{\mathrm{err}})/(\hat{\gamma} - \overline{\mathrm{err}}) & \text{if } \widehat{\mathrm{Err}}^{(1)}, \hat{\gamma} > \overline{\mathrm{err}} \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

The .632+ rule used in the simulation experiments of Section 4 was

$$\widehat{\mathrm{Err}}^{(.632+)} = \widehat{\mathrm{Err}}^{(.632)} + (\widehat{\mathrm{Err}}^{(1)'} - \overline{\mathrm{err}}) \frac{.368 \cdot .632 \cdot \hat{R}'}{1 - .368\hat{R}'}. \tag{32}$$

Figure 3 shows that $\widehat{\mathrm{Err}}^{(.632+)}$ was a reasonably successful compromise between the upwardly biased $\widehat{\mathrm{Err}}^{(1)}$ and the downwardly biased $\widehat{\mathrm{Err}}^{(.632)}$. The plotted values are the relative bias in each of the 24 experiments, measured by

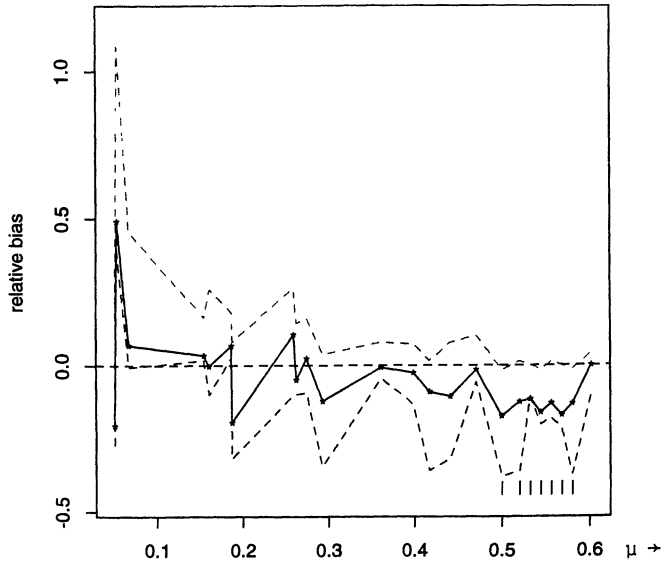

Figure 3. Relative Bias of $\widehat{\mathrm{Err}}^{(.632+)}$ for the 24 Experiments (Solid Curve), Compared to $\widehat{\mathrm{Err}}^{(1)}$ (Top Curve) and $\widehat{\mathrm{Err}}^{(.632)}$ (bottom curve). Plotted values are (mean − μ)/μ. Dashes indicate the seven no-information experiments, where Err = .50.

Table 2. The 24 Sampling Experiments Described in Text

Experiment	n	p	Class	μ	Rule	F
1	14	5	2	.24	LDF	$\text{Normal}_5 \pm (1, 0, 0, 0, 0)$
2	14	5	2	.50	LDF	$\text{Normal}_5 \pm (0, 0, 0, 0, 0)$
3*	20	2	2	.34	LDF	$\text{Normal}_2 \pm (.5, 0)$
4	20	2	2	.50	LDF	$\text{Normal}_2 \pm (0, 0)$
5	14	5	2	.28	NN	$\text{Normal}_5 \pm (1, 0, 0, 0, 0)$
6	14	5	2	.50	NN	$\text{Normal}_5 \pm (0, 0, 0, 0, 0)$
7*	20	2	2	.41	NN	$\text{Normal}_2 \pm (.5, 0)$
8	20	2	2	.50	NN	$\text{Normal}_2 \pm (0, 0)$
9	14	5	2	.26	3NN	$\text{Normal}_5 \pm (1, 0, 0, 0, 0)$
10	14	5	2	.50	3NN	$\text{Normal}_5 \pm (0, 0, 0, 0, 0)$
11	20	2	2	.39	3NN	$\text{Normal}_2 \pm (.5, 0)$
12	20	2	2	.50	3NN	$\text{Normal}_2 \pm (0, 0)$
13	100	10	2	.15	LDF	$N_{10}(0, I)$ versus
14				.18	NN	$N_{10}\left(\frac{\sqrt{j}}{2}, \frac{1}{j}\right)$
15				.17	TREES	
16				.05	QDF	
17	20	2	2	.18	LDF	$N_2 \pm (1, 0)$
18	14	12	2	.50	LDF	$N_{12}(0, I)$
19	20	2	4	.60	LDF	$\text{Normal}_2 \pm (.5, 0)$
20	100	19	4	.26	LDF	Vehicle
21				.07	NN	data
22	36	10	2	.07	LDF	Breast cancer
23				.04	NN	data
24	80	15	15	.47	3NN	Soybean data

*Experiments #3 and #7 appear in Table 1 and Figure 1.

NOTE: Results for experiments #1–4 in Table 3; # 5–8 in Table 4; #4–12 in Table 5; #13–16 in Table 6; #17–19 in Table 7; #20–24 in Table 8.

(mean $-\ \mu)/\mu$. The dashes in Figure 3 indicate the seven "no-information" experiments, where y was independent of $\mathbf{t}$ and Err $\equiv .50$. (The μ values for these seven experiments have been spread out for clarity.) In these cases the definitions in (31) effectively truncate $\widehat{\text{Err}}^{(.632+)}$ at or near .50. This almost always gives a more accurate estimate of Err on a case-by-case basis but yields a downward bias overall. To put things another way, we could also improve the accuracy of $\widehat{\text{Err}}^{(\text{cv1})}$ by truncating it at $\hat{\gamma}$, but then $\widehat{\text{Err}}^{(\text{cv1})}$ would no longer be nearly unbiased.

Better bias adjustments of $\widehat{\text{Err}}^{(1)}$ are available, as discussed in Section 6. However, in reducing the bias of $\widehat{\text{Err}}^{(1)}$, they lose about half of the reduction in variance enjoyed by $\widehat{\text{Err}}^{(.632+)}$ and so offer less dramatic improvements over $\widehat{\text{Err}}^{(\text{cv1})}$.

Note that $\widehat{\text{Err}}^{(.632+)}$ requires no additional applications of the prediction rule after computation of $\widehat{\text{Err}}^{(1)}$. Because 50 bootstrap samples are often sufficient, both $\widehat{\text{Err}}^{(1)}$ and $\widehat{\text{Err}}^{(.632+)}$ can be less costly than $\widehat{\text{Err}}^{(\text{cv1})}$ if n is large.

4. SAMPLING EXPERIMENTS

Table 2 describes the 24 sampling experiments performed for this study. Each experiment involved the choice of a training set size n, a probability distribution F giving $\mathbf{x}$ as in (4), a dimension p for the prediction vectors $\mathbf{t}$, and a prediction rule $r_{\mathbf{x}}$. Twenty-one of the experiments were dichotomous, and three involved four or more classes; 0–1 error (2) was used in all 24 experiments. The first 12 experiments each comprised 200 Monte Carlo simulations (i.e., 200 independent choices of the training set $\mathbf{x}$) and the last 12 experiments each comprised 50 simulations. The bootstrap samples were *balanced* in the sense that the indices of the bootstrap data points were obtained by randomly permuting a string of B copies of the integers 1 to n (see Davison, Hinkley, and Schechtman (1986), but the balancing had little effect on our results.

This simulation study might seem excessive, but it must be large to investigate the effects of different classifiers, training set sizes, signal-to-noise ratio, number of classes, and number of observations. Here are some explanatory comments concerning the 24 experiments:

- In experiments #1–18, the response y_i equals 0 or 1 with probability .50, and the conditional distribution of $\mathbf{t}_i|y_i$ is multivariate normal. For example, experiment #3 is as described in (1), $\mathbf{t}_i|y_i \sim N_2\ ((y_i - .5, 0), I)$, but #4 has the no-information form $\mathbf{t}_i|y_i \sim N_2((0, 0), I)$, so that y_i is independent of $\mathbf{t}_i$ and every prediction rule has Err $= .50$. In experiment #19 the response y_i equals 1, 2, 3, or 4 with probability .25, and $\mathbf{t}_i|y_i \sim N_i(\xi_i, I)$ with $\xi_1 = (-.5, -.5), \xi_2 = (-.5, .5), \xi_3 = (.5 - .5)$, or $\xi_4 = (.5, .5)$.
- Experiments #13–16 are taken from Friedman (1994). There are two classes in 10 dimensions and 100 training observations; tors in class 1 are independent standard normal, and those in class 2 are independent normal with mean $\sqrt{j}/2$ and variance $1/j$, for $j = 1, 2, \ldots, 10$. All predictors are useful here, but the ones with higher index j are more so.

Table 3. Results for the LDF Classifier

	1: (14, 5)			2: (14, 5, ind)			3: (20, 2)			4: (20, 2, ind)		
	Exp	*SD*	*RMS*	*Exp*	*SD*	*RMS*	*Exp*	*SD*	*RMS*	*Exp*	*SD*	*RMS*
Err	.259	.063	0	.501	.011	0	.357	.051	0	.500	.010	0
Errhat1	.327	.116	.147	.500	.115	.115	.388	.101	.104	.502	.087	.088
632	.232	.095	.117	.393	.106	.150	.343	.093	.093	.448	.081	.097
632+	.286	.116	.133	.416	.086	.121	.357	.092	.096	.443	.073	.094
cv1	.269	.144	.156	.501	.176	.175	.362	.130	.123	.505	.135	.135
cv5f	.303	.158	.173	.485	.158	.158	.379	.129	.123	.515	.137	.139
cf5fr	.299	.125	.141	.497	.135	.134	.371	.114	.109	.506	.117	.117
bootop	.182	.105	.147	.375	.135	.183	.345	.107	.106	.459	.102	.110
bc1	.275	.153	.164	.499	.163	.163	.376	.115	.113	.499	.109	.109
bc2	.180	.231	.250	.498	.259	.258	.355	.151	.147	.493	.161	.161
Errhat2	.256	.118	.136	.458	.142	.147	.358	.109	.107	.472	.103	.107
errbar	.070	.075	.214	.209	.104	.310	.267	.090	.128	.355	.084	.168
R	.671	.234		.931	.136		.657	.300		.936	.165	

- Experiments #20–24 refer to real datasets taken from the machine learning archive at UC Irvine. The dimensions of these datasets are (846, 19), (683, 10), and (562, 18) (after removing incomplete observations) with four, two, and 15 classes. We followed Kohavi (1995) and chose a random subset of the data to act as the training set, choosing the training set size n so that μ_n still sloped downward. The idea is that if n is so large that the error curve is flat, then the error rate estimation problem is too easy, because the potential biases arising from changing the training set size will be small. We chose training set sizes 100, 36, and 80. The soybean data actually have 35 categorical predictors, many with more than two possible values. To keep the computations manageable, we used only the 15 binary predictors.
- The prediction rules are LDF, Fisher's linear discriminant analysis; 1-NN and 3-NN, one-nearest-neighbor and three-nearest-neighbor classifiers; TREES, a classification tree using the tree function in S-PLUS; and QDF, quadratic discriminant function (i.e., estimating a separate mean and covariance in each class and using the Gaussian log-likelihood for classification).

Tables 3–8 report the performance of several error rate estimators in the 24 sampling experiments. In each table the Exp and SD columns give means and standard deviations, and RMS is the square root of the average squared error for estimating $\mathrm{Err}(\mathbf{x}, F)$, (5). The bootstrap estimators all used $B = 50$ bootstrap replications per simulation.

The error rate estimators include $\widehat{\mathrm{Err}}^{(.632)}$, $\widehat{\mathrm{Err}}^{(.632+)}$, $\widehat{\mathrm{Err}}^{(1)}$, and four different cross-validation rules: cv1, cv5f, and cv10f are leave-one-out and fivefold and tenfold cross-validations, whereas cv5fr is fivefold cross-validation averaged over 10 random choices of the partition (making the total number of recomputations 50, the same as for the bootstrap rules). Also shown are other bias-corrected versions of $\widehat{\mathrm{Err}}^{(1)}$ called bootop, bc1, bc2, and $\widehat{\mathrm{Err}}^{(2)}$; see Section 6. The tables also give statistical summaries for Err, $\overline{\mathrm{err}}$, and $\hat{R}'$; see (31).

The results vary considerably from experiment to experiment, but in terms of RMS error the .632+ rule is an overall winner. In Figure 4 the solid line graphs $\mathrm{RMS}\{\widehat{\mathrm{Err}}^{(.632+)}\}/\mathrm{RMS}\{\widehat{\mathrm{Err}}^{(\mathrm{cv1})}\}$ versus the true expected error μ, (6). The median value of the ratio for the 24 experiments was .78. The dotted line is the rms ratio for estimating, μ rather than Err, a measure that is slightly more

Table 4. Results for the 1-NN Classifier

	5: (14, 5)			6: (14, 5, ind)			7: (20, 2)			8: (20, 2, ind)		
	Exp	*SD*	*RMS*	*Exp*	*SD*	*RMS*	*Exp*	*SD*	*RMS*	*Exp*	*SD*	*RMS*
Err	.293	.056	0	.500	.011	0	.418	.047	0	.500	.011	0
Err^{1}	.303	.134	.122	.491	.132	.132	.424	.105	.095	.507	.097	.097
$\mathrm{Err}^{.632}$	.192	.085	.129	.310	.083	.207	.268	.067	.162	.320	.062	.190
$\mathrm{Err}^{.632+}$	.257	.127	.120	.413	.094	.128	.380	.101	.099	.439	.068	.092
cv1	.287	.161	.151	.496	.169	.168	.419	.133	.123	.513	.136	.136
cv5f	.297	.167	.155	.490	.162	.163	.423	.144	.134	.508	.139	.139
cv5fr	.297	.144	.133	.496	.138	.138	.420	.122	.110	.509	.117	.117
bootop	.107	.047	.194	.172	.046	.331	.150	.037	.271	.180	.035	.322
bc1	.307	.139	.128	.490	.143	.143	.423	.109	.100	.506	.102	.101
bc2	.313	.158	.149	.486	.179	.179	.421	.131	.124	.503	.126	.126
Err^{2}	.197	.088	.126	.319	.087	.201	.274	.069	.157	.327	.063	.185
$\overline{\mathrm{Err}}$	0	0	.298	0	0	.500	0	0	.420	0	0	.500
R	.641	.252		.922	.134		.853	.169		.949	.099	

Table 5. Results for the 3-NN Classifier

	9: (14, 5)			10: (14, 5, ind)			11: (20, 2)			12: (20, 2, ind)		
	Exp	*SD*	*RMS*	*Exp*	*SD*	*RMS*	*Exp*	*SD*	*RMS*	*Exp*	*SD*	*RMS*
Err	.273	.065	0	.500	.011	0	.399	.062	0	.501	.011	0
Err	.314	.116	.131	.494	.112	.113	.427	.097	.091	.507	.083	.083
$\text{Err}^{.632}$	.245	.099	.110	.400	.100	.142	.346	.084	.093	.412	.074	.115
$\text{Err}^{.632+}$	.277	.113	.122	.421	.087	.119	.388	.091	.090	.437	.066	.091
cv1	.263	.154	.154	.496	.173	.173	.401	.139	.126	.509	.138	.138
cv5f	.273	.154	.155	.491	.161	.161	.405	.133	.124	.511	.143	.143
cv5fr	.290	.133	.139	.495	.144	.145	.411	.123	.110	.509	.117	.117
bootop	.237	.122	.127	.412	.135	.162	.359	.106	.106	.431	.101	.123
bc1	.329	.125	.146	.495	.128	.128	.431	.114	.106	.505	.098	.098
bc2	.355	.187	.213	.499	.210	.209	.438	.181	.175	.502	.177	.176
Err^{2}	.250	.120	.124	.425	.131	.152	.369	.103	.099	.441	.099	.115
$\overline{\text{Err}}$	.126	.091	.175	.239	.106	.283	.207	.078	.208	.250	.080	.263
R	.604	.265		.943	.121		.814	.218		.946	.119	

favorable to the .632+ rule, the median ratio now being .72.

Simulation results must be viewed with caution, especially in an area as broadly defined as the prediction problem. The smoothing argument of Section 2 strongly suggests that it should be possible to improve on cross-validation. With this in mind, Figure 4 demonstrates that $\widehat{\text{Err}}^{(.632+)}$ has made full use of the decreased standard deviation seen in Figure 2. However, the decrease in RMS is less dependable than the decrease in SD, and part of the RMS decrease is due to the truncation at $\hat{\gamma}$ in definitions (31) and (32). The truncation effect is particularly noticeable in the seven no-information experiments.

5. ESTIMATING THE STANDARD ERROR OF $\widehat{\text{Err}}^{(1)}$

The same set of bootstrap replications that gives a point estimate of prediction error can also be used to assess the variability of that estimate. This can be useful for inference purposes, model selection, or comparing two models. The method presented here, called "delta-method-after-bootstrap" by Efron (1992), works well for estimators like $\widehat{\text{Err}}^{(1)}$ that are smooth functions of $\mathbf{x}$. It is more difficult to obtain standard error estimates for cross-validation or the .632+ estimator, and we do not study those estimators in this section.

We discuss estimating the usual *external* standard deviation of $\widehat{\text{Err}}^{(1)}$; that is, the variability in $\widehat{\text{Err}}^{(1)}$ caused by the random choice of $\mathbf{x}$. We also discuss the *internal* variability, due to the random choice of the B bootstrap samples (as at the end of Sec. 2), because it affects the assessment of external variability. Finally, we discuss estimating the standard deviation of the *difference* $\widehat{\text{Err}}^{(1)}(\text{rule 1}) - \widehat{\text{Err}}^{(2)}(\text{rule 2})$, where rule 1 and rule 2 are two different prediction rules applied to the same set of data.

The nonparametric delta method estimate of standard error applies to symmetrically defined statistics, those that are invariant under permutation of the points x_i in $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. In this case we can write the statistic as a function $S(\hat{F})$ of the empirical distribution. The form of S can depend on n, but $S(\hat{F})$ must be smoothly defined in the following sense: Let $\hat{F}_{\varepsilon,i}$ be a version of $\hat{F}$ that puts extra probability on x_i,

$$\hat{F}_{\varepsilon,i} : \Pr \begin{cases} \frac{1-\varepsilon}{n} + \varepsilon & \text{on } x_i \\ \frac{1-\varepsilon}{n} & \text{on } x_j \text{ for } j \neq i. \end{cases} \tag{33}$$

Then we need the derivatives $\partial S(\hat{F}_{\varepsilon,i})/\partial\varepsilon$ to exist at $\varepsilon = 0$. Defining

Table 6. Results for the (100, 10) Problem

	13: LDF			14: 1 nearest neighbor			15: trees			16: QDA		
	Exp	*SD*	*RMS*	*Exp*	*SD*	*RMS*	*Exp*	*SD*	*RMS*	*Exp*	*SD*	*RMS*
Err	.155	.013	0	.188	.017	0	.161	.033	0	.053	.012	0
Err^{1}	.180	.035	.045	.203	.028	.032	.203	.032	.061	.111	.022	.063
$\text{Err}^{.632}$	.157	.033	.036	.128	.018	.063	.145	.024	.042	.074	.016	.029
$\text{Err}^{.632+}$	.160	.033	.037	.151	.025	.044	.161	.028	.041	.079	.018	.034
cv-1	.163	.039	.042	.181	.034	.034	.169	.057	.057	.054	.026	.028
cv5f	.169	.040	.044	.193	.036	.036	.180	.046	.056	.066	.028	.032
cv10	.164	.035	.038	.185	.034	.034	.171	.042	.046	.060	.026	.029
bootop	.157	.037	.039	.073	.010	.116	.116	.025	.058	.049	.015	.019
bc1	.167	.038	.042	.202	.030	.033	.188	.042	.059	.063	.026	.029
bc2	.142	.046	.049	.201	.039	.041	.160	.066	.075	−.024	.044	.088
$\overline{\text{Err}}$	.118	.032	.050	0	0	.188	.047	.019	.119	.011	.010	.045
R	.162	.047		.406	.056		.346	.065		.204	.037	

Table 7. LDF for the (20, 2, +), (14, 12, ind), and (20, 2, 4) Problems

	17: (20, 2, +)			18: (14, 12, ind)			19: (20, 2, 4)		
	Exp	SD	RMS	Exp	SD	RMS	Exp	SD	RMS
Err	.187	.028	0	.502	.012	0	.602	.046	0
Errhat1	.221	.088	.094	.496	.067	.069	.627	.083	.096
632	.191	.082	.082	.315	.044	.193	.546	.083	.109
632+	.199	.087	.088	.438	.065	.093	.602	.098	.106
cv-1	.196	.093	.095	.507	.203	.204	.748	.100	.186
cv5f	.207	.104	.106	.492	.147	.148	.763	.101	.195
cv5fr	.204	.092	.094	.504	.093	.094	.730	.085	.162
bootop	.188	.091	.092	.179	.035	.326	.554	.113	.129
bc1	.203	.093	.094	.492	.105	.107	.609	.105	.112
bc2	.169	.115	.115	.485	.193	.195	.577	.157	.162
errbar	.141	.078	.092	.005	.021	.498	.406	.101	.224
R	.271	.196		.967	.074		.743	.189	

$$\hat{D}_i = \frac{1}{n}\left.\frac{\partial S(\hat{F}_{\varepsilon,i})}{\partial \varepsilon}\right|_0, \tag{34}$$

the nonparametric delta method standard error estimate for $S(\hat{F})$ is

$$\widehat{\mathrm{SE}}_{\mathrm{del}}(S) = \left[\sum_1^n \hat{D}_i^2\right]^{1/2} \tag{35}$$

(see Efron 1992, Sec. 5). The vector $\hat{\mathbf{D}} = (\hat{D}_1, \hat{D}_2, \ldots, \hat{D}_n)$ is $1/n$ times the *empirical influence function* of S.

If the prediction rule $r_{\mathbf{x}}(t)$ is a symmetric function of the points x_i in $\mathbf{x}$, as it usually is, then $\widehat{\mathrm{Err}}^{(1)}$ is also symmetrically defined. The expectation in (13) guarantees that $\widehat{\mathrm{Err}}^{(1)}$ will be smoothly defined in $\mathbf{x}$, so we can apply formulas (34)–(35).

We first consider the *ideal case* where $\widehat{\mathrm{Err}}^{(1)}$ is based on all $B = n^n$ possible bootstrap samples $\mathbf{x}^* = (x_{i_1}, x_{i_2}, \ldots, x_{i_n})$, each $i_j \in \{1, 2, \ldots, n\}$. Following the notation in (14)–(16), let

$$q_i^b = I_i^b \cdot Q_i^b \quad \text{and} \quad q_\cdot^b = \frac{1}{n}\sum_{i=1}^n q_i^b. \tag{36}$$

In this notation,

$$\widehat{\mathrm{Err}}^{(1)} = \frac{1}{n}\sum_{i=1}^n \hat{E}_i, \quad \text{where} \quad \hat{E}_i = \sum_{b=1}^B q_i^b \Big/ \sum_{b=1}^B I_i^b. \tag{37}$$

We also define $\hat{C}_i$ to be the bootstrap covariance between N_i^b and q_i^b,

$$\hat{C}_i = \frac{1}{B}\sum_{b=1}^B (N_i^b - 1)q_\cdot^b. \tag{38}$$

The following lemma was proven by Efron and Tibshirani (1995).

Lemma. The derivative (34) for $S(\hat{F}) = \widehat{\mathrm{Err}}^{(1)}$ is

$$\hat{D}_i = \left(2 + \frac{1}{n-1}\right)\frac{\hat{E}_i - \widehat{\mathrm{Err}}^{(1)}}{n} + e_n\hat{C}_i, \tag{39}$$

$e_n \equiv (1 - 1/n)^{-n}$.

A naive estimate of standard error for $\widehat{\mathrm{Err}}^{(1)}$ would be $[\sum_i(\hat{E}_i - \widehat{\mathrm{Err}}^{(1)})^2/n^2]^{1/2}$, based on the false assumption that the $\hat{E}_i$ are independent. This amounts to taking $\hat{D}_i = (\hat{E}_i - \widehat{\mathrm{Err}}^{(1)})/n$ in (35). The actual influences (39) usually result in a larger standard error than the naive estimates.

Table 8. Results for Real Data Examples

	20: veh/LDF			21: veh/1-NN			22: breast/lda			23: breast/1-NN		
	Exp	SD	RMS	Exp	SD	RMS	Exp	SD	RMS	Exp	SD	RMS
Err	.262	.022	0	.442	.023	0	.067	.025	0	.050	.018	0
Errhat1	.300	.035	.057	.476	.050	.067	.098	.040	.055	.058	.046	.041
632	.236	.033	.049	.301	.032	.147	.067	.030	.038	.037	.029	.029
632+	.249	.034	.044	.395	.054	.077	.072	.033	.040	.040	.034	.032
cv-1	.262	.041	.047	.446	.058	.065	.066	.050	.051	.054	.048	.042
cv5f	.275	.052	.058	.458	.062	.068	.084	.053	.056	.056	.053	.046
cv10	.269	.044	.047	.452	.059	.067	.073	.057	.058	.060	.054	.047
bootop	.222	.043	.065	.172	.018	.271	.048	.029	.042	.021	.016	.034
errbar	.128	.035	.142	0	0	.442	.013	.019	.062	0	0	.054
R	.280	.036		.640	.067		.202	.081		.134	.106	

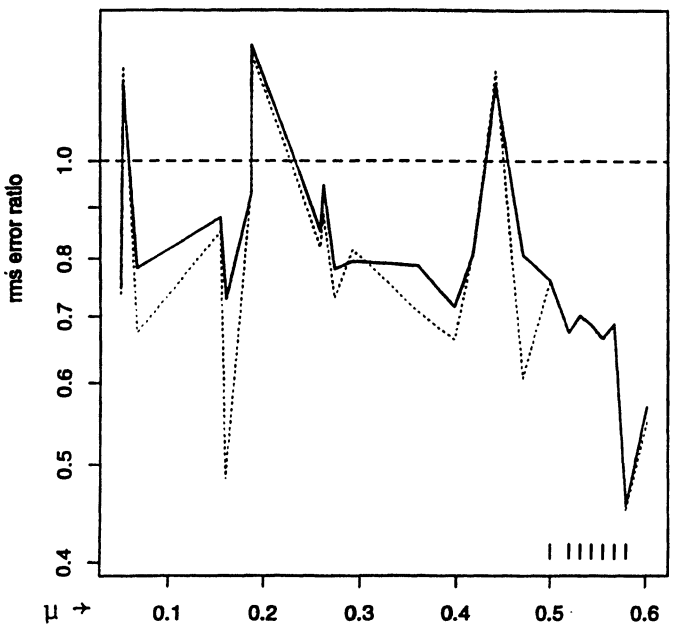

Figure 4. The RMS$\{\widehat{Err}^{(.632+)}\}$/RMS$\{\widehat{Err}^{(cv1)}\}$/RMS Ratio From Tables 3–8, Plotted Versus Expected True Error μ (Solid line) and the RMS Ratio for Estimating μ Instead of Err (Dotted line). Dashes indicate the seven no-information experiments. The vertical axis is plotted logarithmically.

In practice, we have only B bootstrap samples, with B being much smaller than n^n. In that case, we can set

$$\hat{D}_i = \left(2 + \frac{1}{n-1}\right)\frac{\hat{E}_i - \widehat{\text{Err}}^{(1)}}{n} + \frac{\sum_{b=1}^{B}(N_i^b - \bar{N}_i)q_\cdot^b}{\sum_{b=1}^{B} I_i^b}, \tag{40}$$

where $\hat{E}_i$ and $\widehat{\text{Err}}^{(1)}$ are as given in (37), and $\bar{N}_i = \sum_{b=1}^{B} N_i^b/B$. ($\bar{N}_i = 1$ for a balanced set of bootstrap samples.) The bootstrap expectation of I_i^b equals e_n^{-1}, so (40) goes to (39) as we approach the ideal case $B = n^n$.

Finally, we can use the jackknife to assess the internal error of the $\hat{D}_i$ estimates, namely the Monte Carlo error that comes from using only a limited number of bootstrap replications. Let $\hat{D}_{i(b)}$ indicate the value of $\hat{D}_i$ calculated from the $B-1$ bootstrap samples not including the bth sample. Simple computational formulas for $\hat{D}_{i(b)}$ are available along the lines of (17). The internal standard error of $\hat{E}_i$ is given by the jackknife formula

$$\hat{\Delta}_i = \left[\frac{B-1}{B}\sum_b (\hat{D}_{i(b)} - \hat{D}_{i(\cdot)})^2\right]^{1/2}, \tag{41}$$

$\hat{D}_{i(\cdot)} \equiv \sum_b \hat{D}_{i(b)}/B$, with the total internal error

$$\widehat{\text{SE}}_{\text{int}} = \left[\sum_{i=1}^{n} \hat{\Delta}_i^2\right]^{1/2}. \tag{42}$$

This leads to an adjusted standard error estimate for $\widehat{\text{Err}}^{(1)}$,

$$\widehat{\text{SE}}_{\text{adj}} = \left[\sum_{i=1}^{n} \hat{D}_i^2 - \sum_{i=1}^{n} \hat{\Delta}_i^2\right]^{1/2}. \tag{43}$$

These formulas were applied to a single realization of experiment #3, having $n = 20$ points as in Table 2. $B = 1{,}000$ bootstrap replications were generated, and the standard error formulas (25), (32), and (33) were calculated from the first 50, the first 100, and so on. The results appear in Table 9. Using all $B = 1{,}000$ replications gave $\widehat{\text{SE}}_{\text{del}} = .100$, nearly the same as $\widehat{\text{SE}}_{\text{adj}} = .097$. This might be compared to the actual standard deviation .110 for $\widehat{\text{Err}}^{(1)}$ in experiment #3, though of course we expect any data-based standard error estimate to vary from sample to sample.

The right side of Table 9 shows $\widehat{\text{SE}}_{\text{del}}$ and $\widehat{\text{SE}}_{\text{adj}}$ for successive groups of 100 bootstrap replications. The values of $\widehat{SE}_{\text{del}}$ are remarkably stable but biased upward from the answer based on all 1,000 replications; the bias-adjusted values $\widehat{\text{SE}}_{\text{adj}}$ are less biased but about twice as variable from group to group. In this example both $\widehat{\text{SE}}_{\text{del}}$ and $\widehat{\text{SE}}_{\text{adj}}$ gave useful results even for B as small as 100.

The delta method cannot be directly applied to find the standard error of $\widehat{\text{Err}}^{(.632+)}$, because the .632+ rule involves $\overline{\text{err}}$, an unsmooth function of $\mathbf{x}$. A reasonable estimate of standard error for $\widehat{\text{Err}}^{(.632+)}$ is obtained by multiplying (35) or (43) by $\widehat{\text{Err}}^{(.632+)}/\widehat{\text{Err}}^{(1)}$. This is reasonable, because the coefficient of variation for the two estimators was nearly the same in our experiments.

Finally, suppose that we apply two different prediction rules, $r'_{\mathbf{x}}$ and $r''_{\mathbf{x}}$, to the same training set and wish to assess the significance of the difference $\widehat{\text{Diff}} = \widehat{\text{Err}}^{(1)'} - \widehat{\text{Err}}^{(1)''}$ between the error rate estimates. For example, $r'_{\mathbf{x}}$ and $r''_{\mathbf{x}}$ could be LDF and NN, as in Figure 1. The previous theory goes through if we change the definition of Q_i^b in (36) to

$$Q_i^b = Q[y_i, r_{\mathbf{x}}(\mathbf{t}_i)'] - Q[y_i, r_{\mathbf{x}}(\mathbf{t}_i)'']. \tag{44}$$

Then the delta-method estimate of standard error for $\widehat{\text{Diff}}$ is $(\sum_i \hat{D}_i^2)^{1/2}$, where

$$\hat{D}_i = \left(2 + \frac{1}{n-1}\right)\frac{\widehat{\text{Diff}}_i - \widehat{\text{Diff}}}{n} + e_i\hat{C}_i. \tag{45}$$

Here $\widehat{\text{Diff}}_i = \sum_b q_i^b / \sum_b I_i^b$, $q_i^b \equiv I_i^b Q_i^b$, and everything else is as defined in (38) and (39).

Table 9. B = 1,000 Bootstrap Replications From a Single Realization of Experiment #3, Table 2

B	$\widehat{SE}_{del}$	$\widehat{SE}_{int}$	$\widehat{SE}_{adj}$	B	$\widehat{SE}_{del}$	$\widehat{SE}_{int}$	$\widehat{SE}_{adj}$
				1:00	.131	.063	.115
1:50	.187	.095	.161	101:200	.122	.077	.094
				201:300	.128	.068	.108
1:100	.131	.063	.115	301:400	.118	.068	.097
				401:500	.134	.076	.110
1:200	.119	.049	.109	501:600	.116	.085	.079
				601:700	.126	.060	.111
1:400	.110	.034	.105	701:800	.108	.076	.077
				801:900	.109	.084	.068
1:1000	.100	.023	.097	901:1000	.116	.082	.082
				Mean	.121	.074	.0941
				(Std.dev.)	(.009)	(.009)	(.0168)

NOTE: Standard error estimates (25), (32), and (33) were based on portions of the 1,000 replications.

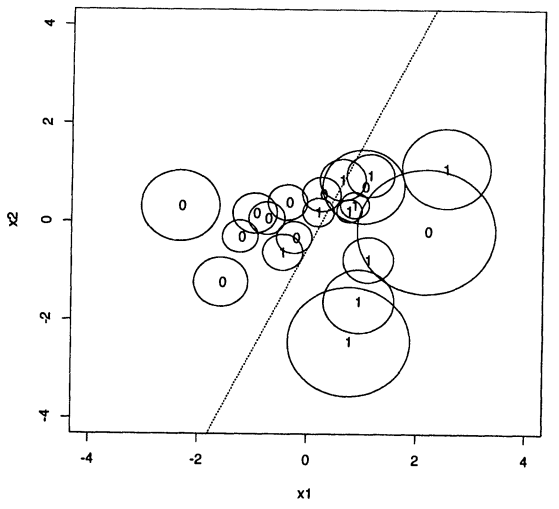

Figure 5. Two Gaussian Classes in Two Dimensions, From the (20, 2) Problem. The circles represent neighborhoods of probability content Δ = .05 around each training point. The dotted line represents the LDF decision boundary.

6. DISTANCE AND BIAS CORRECTIONS

One way to understand the biases of the various error rate estimators is in terms of the distance of test points from the training set: $\overline{\text{err}}$, (7), is the error rate for test points that are zero distance away from the training set $\mathbf{x}$, whereas the true value Err, (5), is the error rate for a new test point x_0 that may lie some distance away from $\mathbf{x}$. Because we expect the error rate of a rule $r_{\mathbf{x}}(\mathbf{t}_0)$ to increase with distance from $\mathbf{x}$, $\overline{\text{err}}$ underestimates Err. Cross-validation has the test points nearly the right distance away from the training set and so is nearly unbiased. The leave-one-out bootstrap $\widehat{\text{Err}}^{(1)}$, (14), has x_0 too far away from the bootstrap training sets $\mathbf{x}^*$, because these are supported on only about $.632n$ of the points in $\mathbf{x}$, producing an upward bias.

A quantitative version of the distance argument leads to the .632+ rule (29). This section presents the argument, which is really quite rough, and then goes on to discuss other more careful bias-correction methods. However, these "better" methods did not produce better estimators in our experiments, with the reduced bias paid for by too great an increase in variance.

Efron (1983) used distance methods to motivate $\widehat{\text{Err}}^{(.632)}$, (24). Here is a brief review of the arguments in that article, which lead up to the motivation for $\widehat{\text{Err}}^{(.632+)}$. Given a system of neighborhoods around points $x = (\mathbf{t}, y)$, let $S(x, \Delta)$ indicate the neighborhood of x having probability content Δ,

$$\Pr\{X_0 \in S(x, \Delta)\} = \Delta. \tag{46}$$

(In this section capital letters indicate random quantities distributed according to F [e.g., X_0 in (46)], and lower-case x values are considered fixed.) As $\Delta \to 0$, we assume that the neighborhood $S(x, \Delta)$ shrinks toward the point x. The distance of test point x_0 from a training set $\mathbf{x}$ is defined by its distance from the nearest point in $\mathbf{x}$,

$$\delta(x_0, \mathbf{x}) = \inf\left\{\Delta : x_0 \in \bigcup_{i=1}^{n} S(x_i, \Delta)\right\}. \tag{47}$$

Figure 5 shows neighborhoods of probability content $\Delta = .05$ for a realization from the (20, 2) problem. Here we have chosen $S(x, \Delta)$ to be circles in the planes $y = 0$ and $y = 1$. That is, if $x_0 = (\mathbf{t}_0, y_0)$, then $S(x_0, \Delta) = \{(\mathbf{t}, y) : y = y_0 \text{ and } \|\mathbf{t} - \mathbf{t}_0\| \le r\}$, where r is chosen so that (46) holds. Notice how the neighborhoods grow smaller near the decision boundary; this occurs because the probability in (46) refers not to the distribution of $\mathbf{t}$ but rather to the joint distribution of $\mathbf{t}$ and y.

Let

$$\mu(\Delta) = E\{Q(X_0, \mathbf{X}) | \delta(X_0, \mathbf{X}) = \Delta\}, \tag{48}$$

the expected prediction error for test points distance Δ from the training set. The true expected error (6) is given by

$$\mu = \int_0^1 \mu(\Delta) g(\Delta)\, d\Delta, \tag{49}$$

where $g(\Delta)$ is the density of $\delta(X_0, \mathbf{X})$. Under reasonable conditions, $g(\Delta)$ approaches the exponential density

$$g(\Delta) \doteq ne^{-n\Delta} \qquad (\Delta \in (0, \infty)), \tag{50}$$

(see 1983, appendix). Another important fact is that

$$\mu(0) \equiv \nu \equiv E_F\{\overline{\text{err}}\}, \tag{51}$$

which is just another way of saying that $\overline{\text{err}}$ is the error rate for test points zero distance away from $\mathbf{x}$.

We can also define a bootstrap analog of $\mu(\Delta)$:

$$\mu_*(\Delta) = E\{Q(X_0^*, \mathbf{X}^*) | \delta(X_0^*, \mathbf{X}^*) = \Delta\}, \tag{52}$$

with the expectation in (52) being over the choice of $X_1, X_2, \ldots, X_n \sim F$ and then $X_0^*, X_1^*, X_2^*, \ldots, X_n^* \sim \hat{F}$. Notice that if $\delta > 0$, then X_0^* must not equal any of the X_i points in $\mathbf{X}^*$. This and definition (14) give

$$\xi \equiv E\{\widehat{\text{Err}}^{(1)}\} = \int_0^1 \mu_*(\Delta) g_*(\Delta)\, d\Delta, \tag{53}$$

where $g_*(\Delta)$ is the density of $\delta^* = \delta(X_0^*, \mathbf{X}^*)$ given that $\delta^* > 0$.

Because bootstrap samples are supported on about $.632n$ of the points in the training set, the same argument that gives $g(\Delta) \doteq ne^{n\Delta}$ also shows that

$$g_*(\Delta) \doteq .632ne^{-.632n\Delta}. \tag{54}$$

Efron (1983) further supposed that

$$\mu_*(\Delta) \doteq \mu(\Delta) \quad \text{and} \quad \mu(\Delta) \doteq \nu + \beta\Delta \tag{55}$$

for some value β, with the intercept ν coming from (51). Combining these approximations gives

$$\mu \doteq \nu + \frac{\beta}{n} \quad \text{and} \quad \xi \doteq \nu + \frac{\beta}{.632n}, \tag{56}$$

so

$$\mu \doteq \{.318\nu + .632\xi\}. \tag{57}$$

Substituting $\overline{\text{err}}$ for ν and $\widehat{\text{Err}}^{(1)}$ for ξ results in $\widehat{\text{Err}}^{(.632)}$, (24).

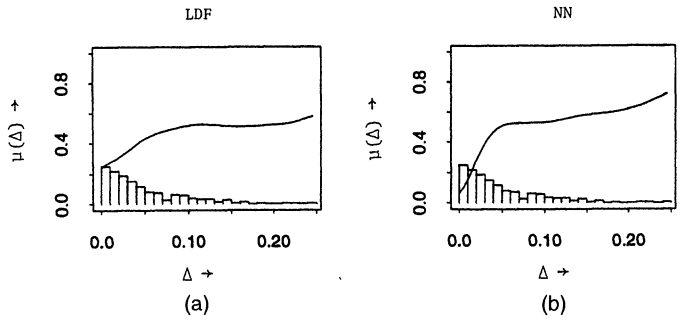

Figure 6. $\mu(\Delta)$ Curves, (38), for Experiments #3 (a) and #7 (b). The linearity assumption $\nu(\Delta) = \nu + \beta\Delta$ is reasonably accurate for the LDF rule but not for NN. The histogram for distance δ (47) supports the exponential approximation (50).

Figure 6 shows $\mu(\Delta)$ for experiments #3 and #7, estimated using all of the data from each set of 200 simulations. The linearity assumption $\mu = \nu + \beta\Delta$ in (55) is reasonably accurate for the LDF rule of experiment #3 but not for the NN rule of #7. The expected apparent error $\nu = \mu(0)$ equals 0 for NN, producing a sharp bend near 0 in the $\mu(\Delta)$ curve. The .632+ rule of Section 3 replaces linearity with an exponential curve, better able to match the form of $\mu(\Delta)$ seen in Figure 5. Now we assume that

$$\mu(\Delta) = \gamma - e^{-(\alpha+\beta\Delta)}. \tag{58}$$

Here α and β are positive parameters and γ is an upper asymptote for $\mu(\Delta)$ as Δ gets large. Formulas (50) and (54), taken literally, produce simple expressions for μ, (49), and for ξ, (53):

$$\mu = \frac{\beta\gamma + n\nu}{\beta + n} \quad \text{and} \quad \xi = \frac{\beta\gamma + .632n\nu}{\beta + .632n}. \tag{59}$$

Combined with $\nu = \gamma - e^{-\alpha}$ from (51), (59) gives

$$\mu = (1 - w)\nu + w\xi, \tag{60}$$

where

$$w = \frac{.632}{1 - .368R} \quad \text{and} \quad R = \frac{\xi - \nu}{\gamma - \nu}. \tag{61}$$

$\widehat{\text{Err}}^{(.632+)}$, (29), is obtained by substituting $\overline{\text{err}}$ for ν, $\widehat{\text{Err}}^{(1)}$ for ξ, and $\hat{\gamma}$, (27), for γ.

All of this is a reasonable plausibility argument for the .632+ rule, but not much more than that. The assumption $\mu(\Delta) \doteq \mu_*(\Delta)$ in (55) is particularly vulnerable to criticism, though in the example shown in figure 2 of Efron (1983) it is quite accurate. A theoretically more defensible approach to the bias correction of $\widehat{\text{Err}}^{(1)}$ can be made using Efron's analysis of variance (ANOVA) decomposition arguments (Efron 1983); see the Appendix.

7. CONCLUSIONS

Our studies show that leave-one-out cross-validation is reasonably unbiased but can suffer from high variability in some problems. Fivefold or tenfold cross-validation exhibits lower variance but higher bias when the error rate curve is still sloping at the given training set size. Similarly, the leave-one-out bootstrap has low variance but sometimes has noticeable bias. The new .632+ estimator is the best overall performer, combining low variance with only moderate bias. All in all, we feel that bias was not a major problem for $\widehat{\text{Err}}^{(.632+)}$ in our simulation studies, and that attempts to correct bias were too expensive in terms of added variability. At the same time, it seems possible that further research might succeed in producing a better compromise between the unbiasedness of cross-validation and the reduced variance of the leave-one-out bootstrap.

APPENDIX: SOME ASYMPTOTIC ANALYSIS OF THE ERROR RATE ESTIMATORS

Define

$$a = -nE\{Q(X_1, \mathbf{X}) - \mu\} \quad \text{and} \quad b = n^2 E\{Q(X_1, \mathbf{X}') - \mu\}, \tag{A.1}$$

where

$$\mathbf{X} = (X_1, X_2, X_3, \ldots, X_n) \quad \text{and} \quad \mathbf{X}' = (X_2, X_2, X_3, \ldots, X_n). \tag{A.2}$$

(Both a and b will usually be positive.) Then the formal ANOVA decompositions of Efron (1983, Sec. 7) give

$$E\{\overline{\text{err}}\} = \mu - a/n, \qquad E\{\widehat{\text{Err}}^{(\text{cv1})}\} = \mu + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right),$$

and

$$E\{\widehat{\text{Err}}^{(1)}\} = \mu + \frac{b}{2n} + \left(\frac{1}{n^2}\right). \tag{A.3}$$

Also, letting $\hat{\mu} \equiv E_{\hat{F}} E_{0\hat{F}} Q(X_0^*, \mathbf{X}^*)$ denote the nonparametric MLE of $\mu = EQ(X_0, \mathbf{X})$,

$$E\{\widehat{\text{Err}}^{(1)} - \hat{\mu}\} = \frac{a}{n} + O\left(\frac{1}{n^2}\right). \tag{A.4}$$

We can combine (A.3) and (A.4) to obtain a bias-corrected version of $\widehat{\text{Err}}^{(1)}$ that, like $\widehat{\text{Err}}^{(\text{cv1})}$, has bias of order $1/n^2$ rather than $1/n$:

$$\widehat{\text{Err}}^{(2)} = \widehat{\text{Err}}^{(1)} - \hat{\mu} + \overline{\text{err}}. \tag{A.5}$$

$\widehat{\text{Err}}^{(2)}$ can be attractively reexpressed in terms of the bootstrap covariances between I_i^b and Q_i^b. Following the notation in (36) and (39), it turns out that

$$\widehat{\text{Err}}^{(2)} = \overline{\text{err}} + \frac{e_n}{n} \sum_{i=1}^{n} \widehat{\text{cov}}_i, \tag{A.6}$$

where

$$\widehat{\text{cov}}_i = \frac{1}{B} \sum_{b=1}^{B} (I_i^b - \bar{I}_i) \cdot Q_i^b, \tag{A.7}$$

$\bar{I}_i = \sum_b I_i^b / B$. Formula (A.6) says that we can bias correct the apparent error rate $\overline{\text{err}}$ by adding e_n times the average covariance between I_I^b (15), the absence or presence of x_i in $\mathbf{x}^*$, and Q_i^b (16), whether or not $r_{\mathbf{x}^*}(\mathbf{t}_i)$ incorrectly predicts y_i. These covariances will usually be positive.

Despite its attractions, $\widehat{\text{Err}}^{(2)}$ was an underachiever in the work of Efron (1983), where it appeared in table 2 as $\hat{\omega}^{(0)}$, and also in the experiments here. Generally speaking, it gains only about half of the RMS advantage over $\widehat{\text{Err}}^{(\text{cv1})}$ enjoyed by $\widehat{\text{Err}}^{(1)}$. Moreover, its bias-correction powers fail for cases like experiment #7, where the formal expansions (A.3) and (A.4) are also misleading.

The estimator called bootop in Tables 3–8 was defined as

$$\widehat{\mathrm{Err}}^{(\mathrm{bootop})} = \overline{\mathrm{err}} - \frac{1}{n}\sum_{i=1}^{n} \mathrm{cov}_*(N_i^b, Q_i^b) \tag{A.8}$$

in (2.10) of Efron (1983), "bootop" standing for "bootstrap optimism." Here $\mathrm{cov}_*(N_i^b, Q_i^b) = \sum_b (N_i^b - 1)Q_i^b/B$. Efron's (1983) section 7 showed that the bootop rule, like $\widehat{\mathrm{Err}}^{(\mathrm{cv1})}$ and $\widehat{\mathrm{Err}}^{(2)}$, has bias of order only $1/n^2$ instead of $1/n$. This does not keep it from being badly biased downward in several of the sampling experiments, particularly for the NN rules.

We also tried to bias correct $\widehat{\mathrm{Err}}^{(1)}$ using a second level of bootstrapping. For each training set $\mathbf{x}$, 50 second-level bootstrap samples were drawn by resampling (one time each) from the 50 original bootstrap samples. The number of distinct original points x_i appearing in a second-level bootstrap sample is approximately $.502 \cdot n$, compared to $.632 \cdot n$ for a first-level sample. Let $\widehat{\mathrm{Err}}^{(\mathrm{sec})}$ be the $\widehat{\mathrm{Err}}^{(1)}$ statistic (17) computed using the second-level samples instead of the first-level samples. The rules called bc1 and bc2 in Tables 3–7 are linear combinations of $\widehat{\mathrm{Err}}^{(1)}$ and $\widehat{\mathrm{Err}}^{(\mathrm{sec})}$,

$$\widehat{\mathrm{Err}}^{(\mathrm{bc1})} = 2 \cdot \widehat{\mathrm{Err}}^{(1)} - \widehat{\mathrm{Err}}^{(\mathrm{sec})}$$

and

$$\widehat{\mathrm{Err}}^{(\mathrm{bc2})} = 3.83\widehat{\mathrm{Err}}^{(1)} - 2.83\widehat{\mathrm{Err}}^{(\mathrm{sec})}. \tag{A.9}$$

The first of these is suggested by the usual formulas for bootstrap bias correction. The second is based on linearly extrapolating $\hat{\mu}_{.502n} = \widehat{\mathrm{Err}}^{(\mathrm{sec})}$ and $\hat{\mu}_{.632n} = \mathrm{Err}^{(1)}$ to an estimate for $\hat{\mu}_n$. The bias correction works reasonably in Tables 3–7, but once again at a substantial price in variability.

[Received May 1995. Revised July 1996.]

REFERENCES

Allen, D. M. (1974), "The Relationship Between Variable Selection and Data Augmentation and a Method of Prediction," *Technometrics*, 16, 125–127.

Breiman, L. (1994), "Bagging Predictors," technical report, University of California, Berkeley.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984), *Classification and Regression Trees*, Pacific Grove, CA: Wadsworth.

Breiman, L., and Spector, P. (1992), "Submodel Selection and Evaluation in Regression: The x-Random Case," *International Statistical Review*, 60, 291–319.

Chernick, M., Murthy, V., and Nealy, C. (1985), "Application of Bootstrap and Other Resampling Methods: Evaluation of Classifier Performance," *Pattern Recognition Letters*, 4, 167–178.

——— (1986), Correction to "Application of Bootstrap and Other esampling Methods: Evaluation of Classifier Performance," *Pattern Recognition Letters*, 4, 133–142.

Cosman, P., Perlmutter, K., Perlmutter, S., Olshen, R., and Gray, R. (1991), "Training Sequence Size and Vector Quantizer Performance," in *25th Asilomar Conference on Signals, Systems, and Computers*, November 4–6, 1991, ed. Ray R. Chen, Los Alamitos, CA: IEEE Computer Society Press, pp. 434–448.

Davison, A. C., Hinkley, D. V., and Schechtman, E. (1986), "Efficient Bootstrap Simulations," *Biometrika*, 73, 555–566.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

——— (1983), "Estimating the Error Rate of a Prediction Rule: Some Improvements on Cross-Validation," *Journal of the American Statistical Association*, 78, 316–331.

——— (1986), "How Biased is the Apparent Error Rate of a Prediction Rule?," *Journal of the American Statistical Association*, 81, 461–470.

——— (1992), "Jackknife-After-Bootstrap Standard Errors and Influence Functions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 83–111.

Efron, B., and Gong, G. (1983), "A Leisurely Look at the Bootstrap, The Jackknife and Cross-Validation," *The American Statistician*, 37, 36–48.

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

——— (1995), "Cross-Validation and the Bootstrap: Estimating the Error Rate of a Prediction Rule," Technical Report 176, Stanford University, Department of Statistics.

Friedman, J. (1994), "Flexible Metric Nearest-Neighbor Classification," Technical Report, Stanford University, Department of Statistics.

Geisser, S. (1975), "The Predictive Sample Reuse Method With Applications," *Journal of the American Statistical Association*, 70, 320–328.

Jain, A., Dubes, R. P., and Chen, C. (1987), "Bootstrap Techniques for Error Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9, 628–633.

Kohavi, R. (1995), "A Study of Cross-Validation and Bootstrap for Accuracy Assessment and model selection," technical report, Stanford University, Department of Computer Sciences.

Mallows, C. (1973), "Some Comments on Cp," *Technometrics*, 15, 661–675.

McLachlan, G. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley.

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.

——— (1977), "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion," *Journal of the Royal Statistical Society*, Ser. B, 39, 44–47.

Wahba, G. (1980), "Spline Bases, Regularization, and Generalized Cross-Validation for Solving Approximation Problems With Large Quantities of Noisy Data," in *Proceedings of the International Conference on Approximation Theory in Honour of George Lorenz*, Austin, TX: Academic Press.

Zhang, P. (1993), "Model Selection via Multifold CV," *The Annals of Statistics*, 21, 299–311.

——— (1995), "APE and Models for Categorical Panel Data," *Scandinavian Journal of Statistics*, 22, 83–94.