

Exact Correlation between Actual and Estimated Errors in Discrete Classification

Ulisses Braga-Neto ^{a,*}, Edward Dougherty ^{a,b,c}

^a*Department of Electrical and Computer Engineering, Texas A&M University,
College Station, TX*

^b*Computational Biology Division, Translational Genomics Research Institute,
Phoenix, AZ*

^c*Department of Pathology, University of Texas MD Anderson Cancer Center,
Houston, TX*

Abstract

Discrete Classification problems are important in pattern recognition applications. The most often used discrete classification rule is the discrete histogram rule. In this letter we provide exact expressions for the correlation coefficient between the actual error and the resubstitution and leave-one-out cross-validation error estimators for the discrete histogram rule. We show with an example that correlations between actual and estimated errors are generally poor, and that in fact leave-one-out cross-validation can display negative correlation when sample sizes are small and classifier complexity is large. We observe that correlation decreases with increasing classifier complexity and increasing sample size does not necessarily produce an increase in correlation. The exact expressions given here can be computed reasonably fast for given sample size, dimensionality, and model parameters, which is useful because, as also illustrated in this letter, Monte-Carlo approximations of the correlation coefficient are generally poor, even at a large number of simulated data sets.

Key words: Error Estimation; Discrete Classification; Correlation Coefficient; Histogram Rule; Resubstitution; Leave-one-out; Cross-validation.

* Corresponding author.

Email addresses: ulisses@ece.tamu.edu (Ulisses Braga-Neto), edward@ece.tamu.edu (Edward Dougherty).

1 Introduction

A full probabilistic understanding of the relationship between an error estimator and the actual error of a sample-designed classifier rests with the joint distribution of the actual and estimated errors relative to the sampling distribution for the underlying feature-label distribution. While knowing the distribution of the error estimator is very important, it alone does not give the complete description of the interaction between the error estimator and the actual error. Of particular importance is the correlation between the actual and estimator errors, which we will label by ε_n and $\hat{\varepsilon}_n$, respectively, where n is the size of the sample. Since $\hat{\varepsilon}_n$ is used in place of ε_n in classifier application, ideally we would like ε_n and $\hat{\varepsilon}_n$ to be perfectly correlated. In fact, as investigated via simulations in [Hanczar et al.(2007)], they are often very poorly correlated. The effect of this lack of correlation can be seen by considering the variance of the deviation $\hat{\varepsilon}_n - \varepsilon_n$, which is given by

$$\text{Var}(\hat{\varepsilon}_n - \varepsilon_n) = \text{Var}(\hat{\varepsilon}_n) + \text{Var}(\varepsilon_n) - 2\rho(\hat{\varepsilon}_n, \varepsilon_n)\sqrt{\text{Var}(\hat{\varepsilon}_n)\text{Var}(\varepsilon_n)} \quad (1)$$

where ρ is the correlation coefficient for ε_n and $\hat{\varepsilon}_n$. A smaller correlation between error estimator and actual error leads to a larger variance for the deviation and vice-versa. The larger the deviation variance, the larger the root-mean-square (RMS) error between ε_n and $\hat{\varepsilon}_n$. If the sample is very large, then the variances of ε_n and $\hat{\varepsilon}_n$ tend to be small, so that the deviation variance is small; however, when the sample is small, these variances tend to be large, so that strong correlation is needed to offset these variances. Thus, the correlation between the actual and estimated errors plays a vital role in assessing the goodness of the error estimator.

In this letter we provide an exact representation for the correlation coefficient between the actual error and the resubstitution and leave-one-out cross-validation error estimators for the *discrete histogram rule*, also called *multinomial discrimination* [Devroye et al.(1996)], which is important in many practical applications, particularly in medicine, economics, psychology and social science [Goldstein and Dillon(1978)]. While other discrete classification rules of practical significance exist (e.g., see [Asparoukhov and Krzanowski(2001), Celeux and Mkhadri(1992)]), the discrete histogram rule is simple enough to allow the exact analytical study of its properties, while at the same time being able to illuminate issues related to classification in general. The classical references on classification error for the discrete histogram rule concern only the actual classification error, or the bias of the apparent, or *resubstitution*, error [Hills(1966), Hills(1967), Hughes(1968), Hughes(1969), Glick(1973)].

In [Braga-Neto and Dougherty(2005)], the authors found analytical expressions for exact calculation of the bias, variance and RMS of not only resubstitution, but also of the *leave-one-out cross-validation* error estimator, for the discrete histogram rule. The authors also described in [Braga-Neto and Dougherty(2005)] a *complete enumeration* method to compute the marginal and joint sampling distributions of resubstitution and leave-one-out cross-validation, with respect to the actual classification error; complete enumeration methods, which have been used extensively for discrete data analysis in statistics

[Agresti(1992), Verbeek(1985), Klotz(1966), Hirji et al.(1987)], rely on intensive computational power to list all possible configurations of data and their probabilities, and from this to derive exact statistical properties of the methods of interest. Efficient computer algorithms were discussed in [Braga-Neto and Dougherty(2005)] in order to implement the proposed complete enumeration methods. In [Xu et al.(2006)], these results were extended to the exact computation of confidence intervals and conditional bias.

In [Braga-Neto and Dougherty(2005)], we did not consider the problem of computing the correlation between the resubstitution or leave-one-out cross-validation errors and the actual classification error. We do this in the present letter, by providing exact expressions for the correlation coefficient, which are faster to compute than by complete enumeration. They are also exact, providing an advantage over Monte-Carlo approximations, which are quite inaccurate for the computation of the correlation coefficient. Not only will we see that the resubstitution and leave-one-out cross-validation error estimators are generally poorly correlated with the actual error, but that it is even possible for leave-one-out cross-validation to display negative correlation when sample sizes are small and classifier complexity is large, exactly the situation in which strong correlation is needed to obtain useful estimates. In general, we will see that the correlation decreases with increasing classifier complexity and that increasing sample size does not produce a corresponding increase in correlation between the actual and estimated errors.

2 Discrete Classification

Let X_1, X_2, \dots, X_d be a set of quantized predictor random variables such that each X_i is quantized into a finite number b_i of values, and let Y be a target random variable taking values in $\{0, 1, \dots, c-1\}$ (for simplicity, we assume $c = 2$). Since the predictors as a group take on values in a finite space of $b = \prod_{i=1}^d b_i$ possible states and a bijection can be established between this finite state-space and the sequence of integers $1, \dots, b$, one can alternatively and equivalently assume, without loss of generality, a single predictor variable X taking on values in the set $\{1, \dots, b\}$. The value b is the total number of quantization levels, or the number of “bins,” into which the data are categorized — this parameter provides a direct measure of the complexity of discrete classification.

The discrete classification model is completely specified by the class prior probabilities $c_0 = P(Y = 0)$, $c_1 = P(Y = 1)$, and the class-conditional probabilities: $p_i = P(X = i | Y = 0)$, $q_i = P(X = i | Y = 1)$, for $i = 1, \dots, b$, where $c_0 + c_1 = 1$, $\sum_{i=1}^b p_i = 1$, and $\sum_{i=1}^b q_i = 1$. Let $S_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ be an i.i.d. sample taken from this probability model, and define the *bin counts*:

$$\begin{aligned} U_i &= \sum_{j=1}^n I_{X_j=i} I_{Y_j=0}, \quad i = 1, \dots, b, \\ V_i &= \sum_{j=1}^n I_{X_j=i} I_{Y_j=1}, \quad i = 1, \dots, b, \end{aligned} \tag{2}$$

where I_A is the usual indicator function for event A . Note that $N_0 = \sum_{i=1}^b U_i$ and $N_1 = \sum_{i=1}^b V_i$ are the total number of samples in classes 0 and 1, respectively, with $N_0 + N_1 = n$.

Given observed values $U_i = u_i$ and $V_i = v_i$, for $i = 1, \dots, b$, the *discrete histogram classification rule* produces the discrete classifier given by

$$\psi_n(i) = I_{u_i < v_i} = \begin{cases} 1, & u_i < v_i \\ 0, & \text{otw} \end{cases}, \quad i = 1, \dots, b. \quad (3)$$

The classification error is given by

$$\begin{aligned} \varepsilon_n &= P(Y \neq \psi_n(X)) \\ &= \sum_{i=1}^b \left[P(Y = 0, X = i) I_{\psi_n(i)=1} + P(Y = 1, X = i) I_{\psi_n(i)=0} \right] \\ &= \sum_{i=1}^b [c_0 p_i I_{U_i < V_i} + c_1 q_i I_{U_i \geq V_i}]. \end{aligned} \quad (4)$$

The *resubstitution* or *apparent* error estimator [Smith(1947)], in the case of the discrete histogram rule, is given by

$$\begin{aligned} \hat{\varepsilon}_n^r &= \frac{1}{n} \sum_{i=1}^n |Y_i - \psi_n(i)| \\ &= \frac{1}{n} \sum_{i=1}^b \min\{U_i, V_i\} \\ &= \frac{1}{n} \sum_{i=1}^b [U_i I_{U_i < V_i} + V_i I_{U_i \geq V_i}] \end{aligned} \quad (5)$$

while the *leave-one-out cross-validation* error estimator, usually attributed to [Lachenbruch and Mickey(1968)], is given in this case by

$$\begin{aligned} \hat{\varepsilon}_n^l &= \frac{1}{n} \sum_{i=1}^n |Y_i - \psi_n^i(i)| \\ &= \frac{1}{n} \sum_{i=1}^b [U_i I_{U_i \leq V_i} + V_i I_{U_i \geq V_{i-1}}] \end{aligned} \quad (6)$$

where ψ_n^i is the classifier obtained when sample point (x_i, y_i) is deleted from the data. Note from the above equations that the random variables U_i and V_i , for $i = 1, \dots, b$, are sufficient statistics for determination of the actual, resubstitution, and leave-one-out cross-validation errors.

3 Correlation Between Actual and Estimated Errors

We provide in this section an exact representation for the correlation coefficients $\rho(\varepsilon_n, \hat{\varepsilon}_n^r)$ and $\rho(\varepsilon_n, \hat{\varepsilon}_n^l)$ in terms of the random variables U_i and V_i , for $i = 1, \dots, b$.

It follows from (4), after some algebraic manipulation, that the variance of the actual error can be written as

$$\begin{aligned} \text{Var}(\varepsilon_n) &= \sum_{i=1}^b (c_1 q_i - c_0 p_i)^2 \text{Var}(I_{U_i < V_i}) \\ &+ 2 \sum_{i < j} (c_1 q_i - c_0 p_i)(c_1 q_j - c_0 p_j) \text{Cov}(I_{U_i < V_i}, I_{U_j < V_j}), \end{aligned} \quad (7)$$

where

$$\begin{aligned} \text{Var}(I_{U_i < V_i}) &= P(U_i < V_i)[1 - P(U_i < V_i)] \\ \text{Cov}(I_{U_i < V_i}, I_{U_j < V_j}) &= P(U_i < V_i, U_j < V_j) - P(U_i < V_i)P(U_j < V_j), \end{aligned} \quad (8)$$

in which

$$\begin{aligned} P(U_i < V_i) &= \sum_{k < l} P(U_i = k, V_i = l) \\ P(U_i < V_i, U_j < V_j) &= \sum_{\substack{k < l \\ r < s}} P(U_i = k, V_i = l, U_j = r, V_j = s). \end{aligned} \quad (9)$$

The required probabilities can be computed by using the fact that the random vector $(U_1, \dots, U_b, V_1, \dots, V_b)$ is jointly multinomially distributed with parameters $(n, c_0 p_1, \dots, c_0 p_b, c_1 q_1, \dots, c_1 q_b)$. This implies that (U_i, V_i) and (U_i, V_i, U_j, V_j) are also jointly multinomially distributed, with parameters $(n, c_0 p_i, c_1 q_i)$ and $(n, c_0 p_i, c_1 q_i, c_0 p_j, c_1 q_j)$, respectively, so that

$$\begin{aligned} P(U_i = k, V_i = l) &= \binom{n}{k, l, n-k-l} (c_0 p_i)^k (c_1 q_i)^l (1 - c_0 p_i - c_1 q_i)^{n-k-l} \\ P(U_i = k, V_i = l, U_j = r, V_j = s) &= \binom{n}{k, l, r, s, n-k-l-r-s} (c_0 p_i)^k (c_1 q_i)^l (c_0 p_j)^r (c_1 q_j)^s (1 - c_0 p_i - c_1 q_i - c_0 p_j - c_1 q_j)^{n-k-l-r-s}. \end{aligned} \quad (10)$$

Now, it follows from (5) that the variance of the resubstitution error estimator is given by:

$$\begin{aligned} \text{Var}(\hat{\varepsilon}_n^r) &= \frac{1}{n^2} \sum_{i=1}^b \text{Var}(U_i I_{U_i < V_i}) + \text{Var}(V_i I_{U_i \geq V_i}) + 2 \text{Cov}(U_i I_{U_i < V_i}, V_i I_{U_i \geq V_i}) \\ &+ \frac{2}{n^2} \sum_{i < j} \left[\text{Cov}(U_i I_{U_i < V_i}, U_j I_{U_j < V_j}) + \text{Cov}(U_i I_{U_i < V_i}, V_j I_{U_j \geq V_j}) + \right. \\ &\quad \left. \text{Cov}(U_j I_{U_j < V_j}, V_i I_{U_i \geq V_i}) + \text{Cov}(V_i I_{U_i \geq V_i}, V_j I_{U_j \geq V_j}) \right] \end{aligned} \quad (11)$$

while the variance of the leave-one-out cross-validation error estimator follows from (6):

$$\begin{aligned} \text{Var}(\hat{\varepsilon}_n^l) &= \frac{1}{n^2} \sum_{i=1}^b \text{Var}(U_i I_{U_i \leq v_i}) + \text{Var}(V_i I_{U_i \geq v_{i-1}}) + 2\text{Cov}(U_i I_{U_i \leq v_i}, V_i I_{U_i \geq v_{i-1}}) \\ &+ \frac{2}{n^2} \sum_{i < j} \left[\text{Cov}(U_i I_{U_i \leq v_i}, U_j I_{U_j \leq v_j}) + \text{Cov}(U_i I_{U_i \leq v_i}, V_j I_{U_j \geq v_{j-1}}) + \right. \\ &\quad \left. \text{Cov}(U_j I_{U_j \leq v_j}, V_i I_{U_i \geq v_{i-1}}) + \text{Cov}(V_i I_{U_i \geq v_{i-1}}, V_j I_{U_j \geq v_{j-1}}) \right]. \end{aligned} \quad (12)$$

Exact expressions for the variances and covariances appearing in (11) and (12) can be readily found using the probabilities in (10) — see the Appendix.

The covariance between actual and estimated errors is a quantity of fundamental interest here. For resubstitution, it is given by

$$\begin{aligned} \text{Cov}(\varepsilon_n, \hat{\varepsilon}_n^r) &= \frac{1}{n} \sum_{i,j} c_0 p_i \left(\text{Cov}(I_{U_i < v_i}, U_j I_{U_j < v_j}) + \text{Cov}(I_{U_i < v_i}, V_j I_{U_j \geq v_j}) \right) + \\ &\quad c_1 q_i \left(\text{Cov}(I_{U_i \geq v_i}, U_j I_{U_j < v_j}) + \text{Cov}(I_{U_i \geq v_i}, V_j I_{U_j \geq v_j}) \right) \end{aligned} \quad (13)$$

while for leave-one-out cross-validation, it is given by:

$$\begin{aligned} \text{Cov}(\varepsilon_n, \hat{\varepsilon}_n^l) &= \frac{1}{n} \sum_{i,j} c_0 p_i \left(\text{Cov}(I_{U_i < v_i}, U_j I_{U_j \leq v_j}) + \text{Cov}(I_{U_i < v_i}, V_j I_{U_j \geq v_{j-1}}) \right) + \\ &\quad c_1 q_i \left(\text{Cov}(I_{U_i \geq v_i}, U_j I_{U_j \leq v_j}) + \text{Cov}(I_{U_i \geq v_i}, V_j I_{U_j \geq v_{j-1}}) \right) \end{aligned} \quad (14)$$

(see the Appendix for the expressions for the covariances involved in these equations).

Having computed the quantities in eqs. (7), (11)–(14), for given n and b , and model parameters c_0 , $c_1 = 1 - c_0$, and p_i, q_i , for $i = 1, \dots, b$, one can find the exact correlation coefficients:

$$\rho(\varepsilon_n, \hat{\varepsilon}_n^r) = \frac{\text{Cov}(\varepsilon_n, \hat{\varepsilon}_n^r)}{\sqrt{\text{Var}(\varepsilon_n) \text{Var}(\hat{\varepsilon}_n^r)}} \quad (15)$$

and

$$\rho(\varepsilon_n, \hat{\varepsilon}_n^l) = \frac{\text{Cov}(\varepsilon_n, \hat{\varepsilon}_n^l)}{\sqrt{\text{Var}(\varepsilon_n) \text{Var}(\hat{\varepsilon}_n^l)}}. \quad (16)$$

4 Examples

Figure 1 displays plots of the exact correlation between the actual error and the resubstitution and leave-one-out cross-validation error estimators, obtained with the previous expressions. Correlation is plotted versus sample size, for different bin sizes and probability models of distinct difficulty, as determined by the optimal (Bayes) classification error, from easy (Bayes

error = 10%) to difficult (Bayes error = 40%). The bin sizes are selected to correspond to the cases of 2,3,4, and 5 binary predictors. In this example, $c_0 = c_1 = 0.5$, and the class-conditional distributions are parametric Zipf (power-law) distributions: $p_i = Ki^{-\alpha}$ and $q_i = p_{b-i+1}$, for $i = 1, \dots, b$, where the normalizing constant is given by $K = [\sum_{i=1}^b i^{-\alpha}]^{-1}$; the parameter $\alpha > 0$ is adjusted to give the desired target Bayes error. This is also the model used in [Braga-Neto and Dougherty(2005)].

[Fig. 1 about here.]

We can observe that the correlation is generally low (below 0.3) in most cases. At small and large classification difficulty, the behavior of the correlation is much more regular than at intermediate difficulty models, both for resubstitution and leave-one-out cross-validation. We can also observe that at small sample sizes, correlation for resubstitution is larger than for leave-one-out cross-validation, with few exceptions. Correlation for leave-one-out cross-validation is more sensitive to complexity of the classification rule, as measured by bin size; it tends to decrease for larger bin size, and in one striking case, it becomes negative, at the very small-sample situation of $n = 20$ and $b = 32$. This behavior of the correlation precisely mirrors the behavior of the deviation variance $\text{Var}(\hat{\varepsilon}_n^l - \varepsilon_n)$, which is known to be large for the cross-validation error estimator under complex models and small sample sizes [Devroye et al.(1996), Braga-Neto and Dougherty(2004)], and is in complete accord with (1).

To try to understand further the correlation obtained for resubstitution and leave-one-out cross-validation, it is useful to examine the joint sampling distribution of these error estimators and the actual error. The exact joint distributions can be computed by the complete enumeration method described in [Braga-Neto and Dougherty(2005)]. Figure 2 displays plots of the exact joint distribution between the actual error and the resubstitution and leave-one-out cross-validation error estimators, for a small-sample case, $n = 20$ and $b = 8$, and probability models of intermediate difficulty (Bayes error = 20% and 30%) — these are the same models used in connection with Figure 1. We can observe that the joint distribution for resubstitution is much more compact than for leave-one-out cross-validation, which explains in part its larger correlation in small-sample cases.

[Fig. 2 about here.]

The exact expressions for the correlation coefficient presented in this paper are also important due to the poor accuracy displayed by the corresponding Monte-Carlo estimates. Figure 3 displays plots of the Monte-Carlo approximations of the correlation coefficient versus sample size, for the model with Bayes error = 0.40 in Figure 1. Comparing the results in Figure 3 with the corresponding exact results in Figure 1, one observes that there is considerable variance in the MC estimates, even at $M=50000$ simulated training data sets.

[Fig. 3 about here.]

Appendix

We give below the expressions necessary for the computation of equations (11)–(14).

For eq. (11),

$$\begin{aligned}\text{Var}(U_i I_{U_i < V_i}) &= E[U_i^2 I_{U_i < V_i}] - (E[U_i I_{U_i < V_i}])^2 \\ &= \sum_{k < l} k^2 P(U_i = k, V_i = l) - \left(\sum_{k < l} k P(U_i = k, V_i = l) \right)^2\end{aligned}\quad (17)$$

$$\begin{aligned}\text{Var}(V_i I_{U_i \geq V_i}) &= E[V_i^2 I_{U_i \geq V_i}] - (E[V_i I_{U_i \geq V_i}])^2 \\ &= \sum_{k \geq l} l^2 P(U_i = k, V_i = l) - \left(\sum_{k \geq l} l P(U_i = k, V_i = l) \right)^2\end{aligned}\quad (18)$$

$$\begin{aligned}\text{Cov}(U_i I_{U_i < V_i}, V_i I_{U_i \geq V_i}) &= E[U_i V_i I_{U_i < V_i} I_{U_i \geq V_i}] - E[U_i I_{U_i < V_i}] E[V_i I_{U_i \geq V_i}] \\ &= - \left(\sum_{k < l} k P(U_i = k, V_i = l) \right) \left(\sum_{k \geq l} l P(U_i = k, V_i = l) \right)\end{aligned}\quad (19)$$

$$\begin{aligned}\text{Cov}(U_i I_{U_i < V_i}, U_j I_{U_j < V_j}) &= E[U_i U_j I_{U_i < V_i} I_{U_j < V_j}] - E[U_i I_{U_i < V_i}] E[U_j I_{U_j < V_j}] \\ &= \sum_{\substack{k < l \\ r < s}} kr P(U_i = k, V_i = l, U_j = r, V_j = s) - \left(\sum_{k < l} k P(U_i = k, V_i = l) \right) \left(\sum_{k < l} k P(U_j = k, V_j = l) \right)\end{aligned}\quad (20)$$

while $\text{Cov}(U_i I_{U_i < V_i}, V_j I_{U_j \geq V_j})$, $\text{Cov}(U_j I_{U_j < V_j}, V_i I_{U_i \geq V_i})$, and $\text{Cov}(V_i I_{U_i \geq V_i}, V_j I_{U_j \geq V_j})$ are found analogously as in (20).

Similarly, for eq. (12),

$$\text{Var}(U_i I_{U_i \leq V_i}) = \sum_{k \leq l} k^2 P(U_i = k, V_i = l) - \left(\sum_{k \leq l} k P(U_i = k, V_i = l) \right)^2\quad (21)$$

$$\text{Var}(V_i I_{U_i \geq V_i - 1}) = \sum_{k \geq l - 1} l^2 P(U_i = k, V_i = l) - \left(\sum_{k \geq l - 1} l P(U_i = k, V_i = l) \right)^2\quad (22)$$

$$\begin{aligned}\text{Cov}(U_i I_{U_i \leq V_i}, V_i I_{U_i \geq V_i - 1}) &= \\ &= \sum_{l - 1 \leq k \leq l} kl P(U_i = k, V_i = l) - \left(\sum_{k \leq l} k P(U_i = k, V_i = l) \right) \left(\sum_{k \geq l - 1} l P(U_i = k, V_i = l) \right)\end{aligned}\quad (23)$$

$$\begin{aligned}
& \text{Cov}(U_i I_{U_i \leq V_i}, U_j I_{U_j \leq V_j}) = \\
& = \sum_{\substack{k \leq l \\ r \leq s}} kr P(U_i = k, V_i = l, U_j = r, V_j = s) - \left(\sum_{k \leq l} k P(U_i = k, V_i = l) \right) \left(\sum_{k \leq l} k P(U_j = k, V_j = l) \right)
\end{aligned} \tag{24}$$

while $\text{Cov}(U_i I_{U_i \leq V_i}, V_j I_{U_j \geq V_j - 1})$, $\text{Cov}(U_j I_{U_j \leq V_j}, V_i I_{U_i \geq V_i - 1})$, and $\text{Cov}(V_i I_{U_i \geq V_i - 1}, V_j I_{U_j \geq V_j - 1})$ are found analogously as in (24).

For eq. (13),

$$\begin{aligned}
& \text{Cov}(I_{U_i < V_i}, U_j I_{U_j < V_j}) = E[U_j I_{U_i < V_i} I_{U_j < V_j}] - E[I_{U_i < V_i}] E[U_j I_{U_j < V_j}] \\
& = \sum_{\substack{k < l \\ r < s}} r P(U_i = k, V_i = l, U_j = r, V_j = s) - P(U_i < V_i) \left(\sum_{k < l} k P(U_j = k, V_j = l) \right)
\end{aligned} \tag{25}$$

When $i = j$, this reduces to:

$$\begin{aligned}
& \text{Cov}(I_{U_i < V_i}, U_i I_{U_i < V_i}) = E[U_i I_{U_i < V_i}] - E[I_{U_i < V_i}] E[U_i I_{U_i < V_i}] \\
& = [1 - P(U_i < V_i)] E[U_i I_{U_i < V_i}] \\
& = P(U_i \geq V_i) \left(\sum_{k < l} k P(U_i = k, V_i = l) \right)
\end{aligned} \tag{26}$$

The covariances $\text{Cov}(I_{U_i < V_i}, V_j I_{U_j \geq V_j})$, $\text{Cov}(I_{U_i \geq V_i}, U_j I_{U_j < V_j})$, and $\text{Cov}(I_{U_i \geq V_i}, V_j I_{U_j \geq V_j})$ are found analogously as in (25) and (26).

Similarly, for eq. (14),

$$\begin{aligned}
& \text{Cov}(I_{U_i < V_i}, U_j I_{U_j \leq V_j}) = E[U_j I_{U_i < V_i} I_{U_j \leq V_j}] - E[I_{U_i < V_i}] E[U_j I_{U_j \leq V_j}] \\
& = \sum_{\substack{k < l \\ r \leq s}} r P(U_i = k, V_i = l, U_j = r, V_j = s) - P(U_i < V_i) \left(\sum_{k \leq l} k P(U_j = k, V_j = l) \right)
\end{aligned} \tag{27}$$

When $i = j$, this reduces to:

$$\begin{aligned}
& \text{Cov}(I_{U_i < V_i}, U_i I_{U_i \leq V_i}) = E[U_i I_{U_i < V_i}] - E[I_{U_i < V_i}] E[U_i I_{U_i \leq V_i}] \\
& = \left(\sum_{k < l} k P(U_i = k, V_i = l) \right) - P(U_i < V_i) \left(\sum_{k \leq l} k P(U_i = k, V_i = k) \right)
\end{aligned} \tag{28}$$

The covariances $\text{Cov}(I_{U_i < V_i}, V_j I_{U_j \geq V_j - 1})$, $\text{Cov}(I_{U_i \geq V_i}, U_j I_{U_j \leq V_j})$, and $\text{Cov}(I_{U_i \geq V_i}, V_j I_{U_j \geq V_j - 1})$ are found analogously as in (27).

Note that, for computational efficiency, one can precompute and store many of the terms in the formulas above, such as $P(U_i < V_i)$, $E[U_i I_{U_i < V_i}]$, $E[V_i I_{U_i < V_i}]$, and so forth, and reuse

them multiple times in the calculations.

References

- [Agresti(1992)] Agresti, A., 1992. A survey of exact inference for contingency tables. *Statistical Science* 7 (1), 131–153.
- [Asparoukhov and Krzanowski(2001)] Asparoukhov, O., Krzanowski, W., 2001. A comparison of discriminant procedures for binary variables. *Computational Statistics and Data Analysis* 38, 139–160.
- [Braga-Neto and Dougherty(2004)] Braga-Neto, U., Dougherty, E., 2004. Is cross-validation valid for microarray classification? *Bioinformatics* 20 (3), 374–380.
- [Braga-Neto and Dougherty(2005)] Braga-Neto, U., Dougherty, E., 2005. Exact performance of error estimators for discrete classifiers. *Pattern Recognition* 38 (11), 1799–1814.
- [Celeux and Mkhadri(1992)] Celeux, G., Mkhadri, A., 1992. Discrete regularized discriminant analysis. *Statistics and Computing* 2, 143–151.
- [Devroye et al.(1996)] Devroye, L., Györfi, L., Lugosi, G., 1996. *A Probabilistic Theory of Pattern Recognition*. Springer, New York.
- [Glick(1973)] Glick, N., 1973. Sample-based multinomial classification. *Biometrics* 29 (2), 241–256.
- [Goldstein and Dillon(1978)] Goldstein, M., Dillon, W., 1978. *Discrete Discriminant Analysis*. Wiley, New York.
- [Hanczar et al.(2007)] Hanczar, B., Hua, J., Dougherty, E., 2007. Decorrelation of the true and estimated classifier errors in high-dimensional settings. Article ID 38473, 12 pages.
- [Hills(1966)] Hills, M., 1966. Allocation rules and their error rates. *Journal of the Royal Statistical Society. Series B (Methodological)* 28 (1), 1–31.
- [Hills(1967)] Hills, M., 1967. Discrimination and allocation with discrete data. *Applied Statistics* 16 (3), 237–250.
- [Hirji et al.(1987)] Hirji, K., Mehta, C., Patel, N., 1987. Computing distributions for exact logistic regression. *Journal of the American Statistical Association* 82 (400), 1110–1117.
- [Hughes(1968)] Hughes, G., 1968. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory* 14 (1), 55–63.
- [Hughes(1969)] Hughes, G., 1969. Number of pattern classifier design samples per class. *IEEE Transactions on Information Theory* 15 (5), 615–618.
- [Klotz(1966)] Klotz, J., 1966. The wilcoxon, ties, and the computer. *Journal of the American Statistical Association* 61 (315), 772–787.
- [Lachenbruch and Mickey(1968)] Lachenbruch, P., Mickey, M., 1968. Estimation of error rates in discriminant analysis. *Technometrics* 10, 1–11.

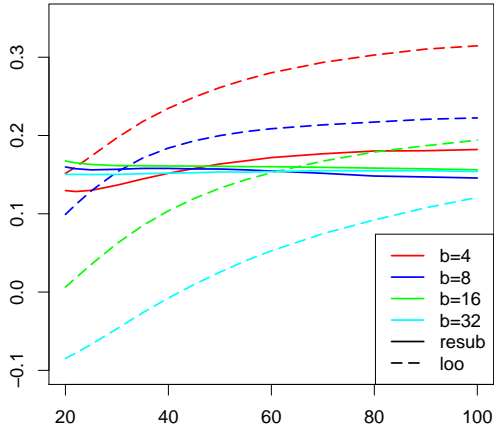
- [Smith(1947)] Smith, C., 1947. Some examples of discrimination. *Annals of Eugenics* 18, 272–282.
- [Verbeek(1985)] Verbeek, A., 1985. A survey of algorithms for exact distributions of test statistics in rxc contingency tables with fixed margins. *Computational Statistics and Data Analysis* 3, 159–185.
- [Xu et al.(2006)] Xu, Q., Hua, J., Braga-Neto, U., Xiong, Z., Suh, E., Dougherty, E., 2006. Confidence intervals for the true classification error conditioned on the estimated error. *Technology in Cancer Research and Treatment* 5 (6), 579–590.

About the Author—ULISSES BRAGA-NETO received the Ph.D. degree in electrical and computer engineering from The Johns Hopkins University, Baltimore, Maryland, in 2002. He held a post-doctoral fellowship in the section of clinical cancer genetics at the University of Texas M.D. Anderson Cancer Center, Houston, from 2002 to 2004. In 2004, he joined the Virology and Experimental Therapy Laboratory at the Oswaldo Cruz Foundation (FIOCRUZ), in Recife, Brazil. Since January 2007, he has been an assistant professor at the Department of Electrical and Computer Engineering of Texas A&M University, College Station, Texas. His current research interests include genomic and immunomic signal processing, computational biology, and statistical pattern recognition, with applications in the study of cancer and infectious diseases.

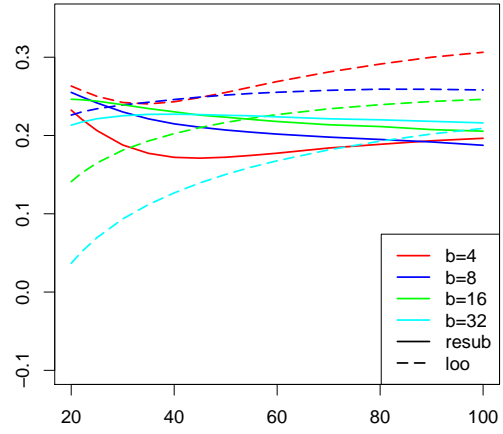
About the Author—EDWARD DOUGHERTY is a Professor in the Department of Electrical and Computer Engineering at Texas A&M University in College Station, TX, where he holds the Robert M. Kennedy Chair and is Director of the Genomic Signal Processing Laboratory. He is also the Director of the Computational Biology Division of the Translational Genomics Research Institute in Phoenix, AZ. He holds a Ph.D. in mathematics from Rutgers University and an M.S. in Computer Science from Stevens Institute of Technology, and has been awarded the Doctor Honoris Causa by the Tampere University of Technology in Finland. He is a fellow of SPIE, has received the SPIE President's Award, and served as the editor of the SPIE/IS&T Journal of Electronic Imaging. At Texas A&M he has received the Association of Former Students Distinguished Achievement Award in Research, been named Fellow of the Texas Engineering Experiment Station, and named Halliburton Professor of the Dwight Look College of Engineering. Prof. Dougherty is author of fourteen books, editor of five others, and author of more than two hundred journal papers. He has contributed extensively to the statistical design of nonlinear operators for image processing and the consequent application of pattern recognition theory to nonlinear image processing. His research in genomic signal processing is aimed at diagnosis and prognosis based on genetic signatures and using gene regulatory networks to develop therapies based on the disruption or mitigation of aberrant gene function contributing to the pathology of a disease.

List of Figures

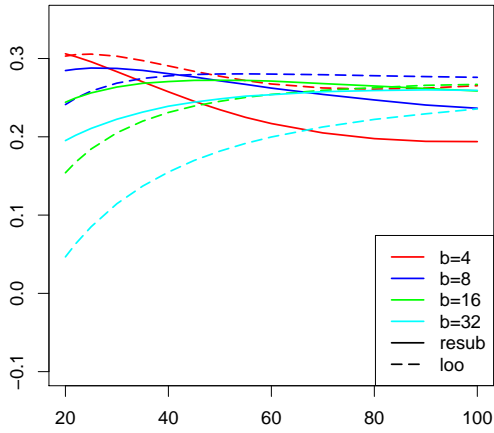
- 1 Exact correlation between the actual error and the resubstitution and leave-one-out cross-validation error estimators versus sample size, for different bin sizes and probability models of distinct difficulty, as determined by the Bayes classification error. 13
- 2 Exact joint distribution between the actual error and the resubstitution and leave-one-out cross-validation error estimators, for $n = 20$ and $b = 8$, and probability models of intermediate difficulty, as determined by the Bayes classification error. 14
- 3 Monte-Carlo approximation of the correlation between the actual error and the resubstitution and leave-one-out cross-validation error estimators versus sample size, for different bin sizes, corresponding to the model with Bayes error = 0.40 in Figure 1, using different numbers M of simulated data sets. 15



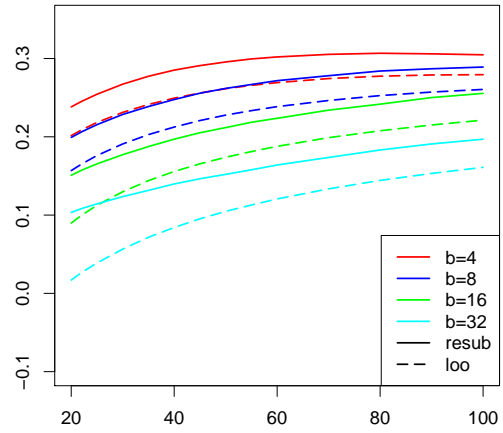
Bayes Error = 0.10



Bayes Error = 0.20

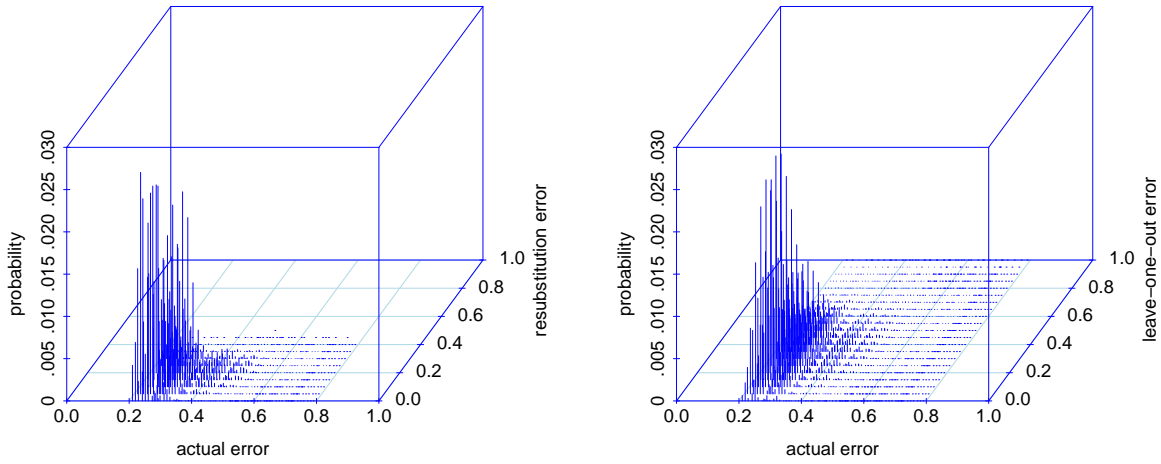


Bayes Error = 0.30

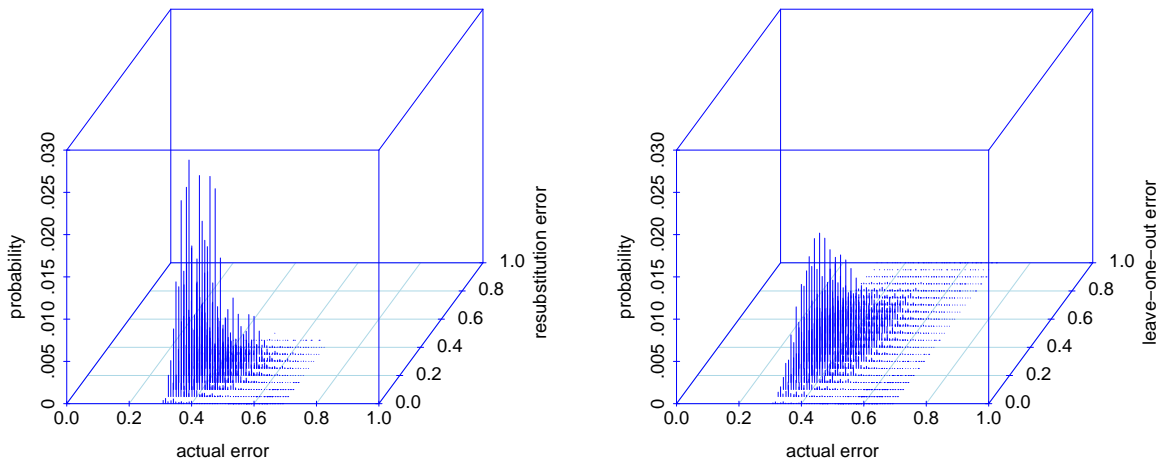


Bayes Error = 0.40

Fig. 1. Exact correlation between the actual error and the resubstitution and leave-one-out cross-validation error estimators versus sample size, for different bin sizes and probability models of distinct difficulty, as determined by the Bayes classification error.

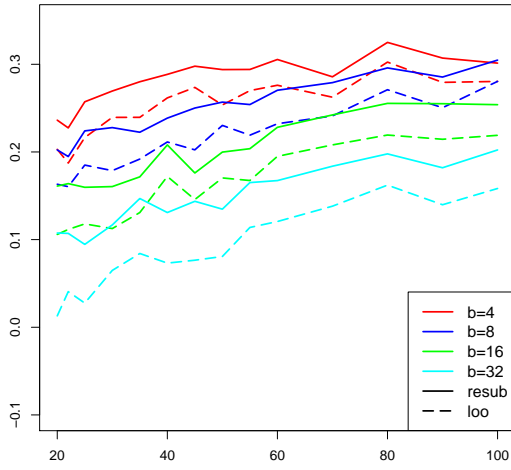


Bayes Error = 0.20

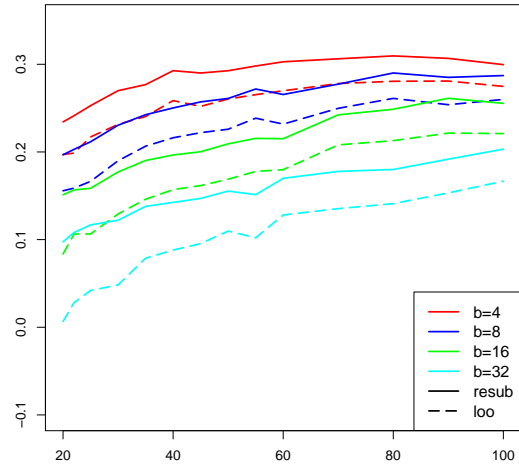


Bayes Error = 0.30

Fig. 2. Exact joint distribution between the actual error and the resubstitution and leave-one-out cross-validation error estimators, for $n = 20$ and $b = 8$, and probability models of intermediate difficulty, as determined by the Bayes classification error.



$M = 10000$



$M = 50000$

Fig. 3. Monte-Carlo approximation of the correlation between the actual error and the resubstitution and leave-one-out cross-validation error estimators versus sample size, for different bin sizes, corresponding to the model with Bayes error = 0.40 in Figure 1, using different numbers M of simulated data sets.