Machine Learning, 52, 11–30, 2003

Relation Between Permutation-Test *P* Values and Classifier Error Estimates

TAILEN HSING
Department of Statistics, Texas A&M University, USA

SANJU ATTOOR
Department of Electrical Engineering, Texas A&M University, USA

EDWARD DOUGHERTY edward@ee.tamu.edu
Department of Electrical Engineering, Texas A&M University, USA; Department of Pathology, University of Texas M.D. Anderson Cancer Center, USA

Editors: Paola Sebastiani, Isaac S. Kohane and Marco F. Ramoni

Abstract. Gene-expression-based classifiers suffer from the small number of microarrays usually available for classifier design. Hence, one is confronted with the dual problem of designing a classifier and estimating its error with only a small sample. Permutation testing has been recommended to assess the dependency of a designed classifier on the specific data set. This involves randomly permuting the labels of the data points, estimating the error of the designed classifiers for each permutation, and then finding the p value of the error for the actual labeling relative to the population of errors for the random labelings. This paper addresses the issue of whether or not this p value is informative. It provides both analytic and simulation results to show that the permutation p value is, up to very small deviation, a function of the error estimate. Moreover, even though the p value is a monotonically increasing function of the error estimate, in the range of the error where the majority of the p values lie, the function is very slowly increasing, so that inversion is problematic. Hence, the conclusion is that the p value is less informative than the error estimate. This result demonstrates that random labeling does not provide any further insight into the accuracy of the classifier or the precision of the error estimate. We have no knowledge beyond the error estimate itself and the various distribution-free, classifier-specific bounds developed for this estimate.

Keywords: classification, error estimation, genomics, microarrays, p value, pattern recognition

1. Introduction

Gene-expression microarrays provide expression measurements for thousands of genes at once (Schena et al., 1995; De Risi, Iyer, & Brown, 1997; Duggan et al., 1999). A key goal is to perform classification via different expression patterns—for instance, cancer classification (Golub et al., 1999; Khan et al., 2001; Hedenfalk et al., 2001). This requires designing a classifier that takes a vector of gene expression levels as input, and outputs a class label predicting the class containing the input vector. Classification can be between different cancers, different stages of tumor development, or many other such differences. Owing to the large number of variables (expression levels) and the small samples (number of microarrays),

many thorny issues for classifier design, error estimation, and feature selection are inherent in microarray-based classification (Dougherty, 2001). In this vein, a number of recent papers have suggested the use of permutation-based p values for obtaining information regarding the selection of relevant genes or for assessing the quality of classification when using expression data. Essentially, a statistic relating to class discrimination is computed from the data, the class labels are randomized some large number of times, the statistic is computed for each re-labeling, a histogram is formed from these re-label statistics, and the p value of the statistic corresponding to the actual labeling is computed. A salient issue is whether this p value is informative.

In this paper we will focus on the error estimate of a classifier designed from the data and its p value relative to errors for classifiers derived from random re-labeling of the data (Pomeroy et al., 2002; Lesnick, Dacwag, & Golub, 2002). In some sense, this setting is primordial because we are ultimately interested in classification. Given the error estimate, there are two issues to consider:

1. Is the p value useful in comparing different classifiers?

   The key performance measure of a classifier is its error relative to the true distribution. In practice, this error can only be estimated. If the p value gives insight into the distribution of the error or the reliability of the estimated error, then a point can be made by including the p value in assessing and/or comparing classifiers. The critical point is that the randomly relabeled data contain little or no information on the true joint distribution of the label and gene expression levels. Hence any insight based on the issue mentioned above based on the p value has to come solely from the estimated error. As such, the p value has nothing to add to what is already known in the literature with regard to the issue (Devroye, Gyorfi, & Lugosi, 1996). The reduction of information caused by re-labeling in a bivariate sample is by no means unique to the classification problem focused on here (cf. Section 4).
2. Is the p value useful in ranking the genes in a common study using a common classifier?

   In variable-selection problems in regression, it is common to use a p value as a basis for identifying the predictors that have superior predictive powers in predicting the response. A small p value corresponds to high predictive power. This is equivalent to the so-called R^2 criterion. In our setting, genes in a common study can be ranked according to p values based on a common classifier; however, the issue is whether this ranking is significantly different from that based on error rates. Conceivably these rankings will be somewhat different in some problems, in which case one has to provide a careful judgment as to whether it makes sense theoretically or practically to include p values as part of the selection process. If this is not the case, then there is little point of going to great length to compute p values, as the computations are typically tedious.

In this paper we will illustrate the latter situation in the context of classification error by showing that the p value is less informative than the original error estimate itself for two commonly employed classification rules. Both simulation and theoretical results are provided. This does not mean that permutation-test p values are always non-informative relative to the statistic from which they are derived and the purpose to which they are being

used; however, it does indicate that the degree to which proposed tests are informative should be studied.

2. Background

We begin by providing some background material on pattern classification. We refer to a comprehensive text for a more complete discussion (Devroye, Gyorfi, & Lugosi, 1996). The ultimate issue of pattern recognition can be stated in the following manner: Given a *feature vector* $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ composed of random variables, and a binary random variable Y, find a classifier ψ so that $\psi(\mathbf{X})$ is the optimal predictor of Y relative to the error, $\epsilon[\psi] = P(\psi(\mathbf{X}) \neq Y)$, giving the probability that classification is erroneous. Sometimes $\epsilon[\psi]$ is referred to as the "error rate" of the classifier. The values, 0 or 1, of Y are called *class labels*. An optimal classifier, $\psi_\bullet$, is one having minimal error, $\epsilon_\bullet$, among all binary functions of $\mathbf{X}$. $\psi_\bullet$ and $\epsilon_\bullet$ are called the *Bayes classifier* and *Bayes error*, respectively. More general classification problems can be considered in which Y is a discrete random variable corresponding to more than two labels; however, from a theoretical perspective, it is sufficient to consider binary classification (Devroye, Gyorfi, & Lugosi, 1996). Moreover, in genomics there are many inherently binary classification problems, such as separating tumor and non-tumor classes.

The difficulty is that the joint distribution of $(\mathbf{X}, Y)$ is typically unknown and an estimate of the optimal classifier must be obtained from sample data. Two issues arise. First, how does one design (estimate) a classifier from the sample data that provides good classification relative to the full distribution? Second, how does one estimate the error of a designed classifier? If data are unlimited, then it is possible to obtain a designed classifier that estimates the optimal classifier to any desired degree of precision using *test data*, and the error of the designed classifier can be estimated to any desired degree of precision using independent *training data*. The matter becomes problematic when data are limited, and highly problematic when samples are small. The second issue, error estimation, concerns us here.

A classifier ψ may perform well on the sample from which it has been designed but not perform well relative to the distribution as a whole. An estimator must be used to estimate the error of the classifier relative to the distribution. This estimator is a random variable relative to the sample used to compute it. Our full understanding of the quality of the classifier rests with the distribution of the estimator—more specifically, what we know about the distribution of the estimator. One would like to use a classification rule to design a close-to-optimal classifier, and various considerations come into play in the choice of a classification rule, such as sample size, prior distributional knowledge, and the Vapnik-Chervonenkis dimension, which can be used to bound the expected increase in error of the designed classifier relative to the error of the truly optimal classifier (Vapnik & Chervonenkis, 1971); but once the classifier is designed, all that one can hope for is to know as much as possible about the distribution of the error estimator for the designed classifier. In the case of small samples, even if the error estimator is unbiased, its variance may be so large that we can have little confidence in any individual empirical estimate of the error. This is why we are concerned with bounds on error-estimator variances. The error estimate is a statistic pertaining to the performance of the classifier, and the variance of the error

estimator tells us something about this statistic. If other statistics can be obtained to give us more information regarding classifier performance, then they are beneficial. But to be valuable, they must be more informative than the error estimate itself.

The Bayes classifier is defined in a natural way: for any vector $\mathbf{x}$, $\psi_\bullet(\mathbf{x}) = 1$ if $P(Y = 1 \mid \mathbf{x}) > \frac{1}{2}$, and $\psi_\bullet(\mathbf{x}) = 0$ otherwise. Typically, the conditional probabilities are unknown. A sample of feature-label pairs and a classification rule are used to construct a classifier ψ_n. Bayes error is estimated by the error ϵ_n of ψ_n. Because $\epsilon_\bullet$ is minimal, $\epsilon_n \geq \epsilon_\bullet$, and there is a design cost $\Delta_n = \epsilon_n - \epsilon_\bullet$. Hopefully, $\Delta_n \to 0$ as the sample size grows. This will depend on the classification rule and the distribution of the feature-label pair $(\mathbf{X}, Y)$. A classification rule is said to be *consistent* for the distribution of $(\mathbf{X}, Y)$ if $E[\Delta_n] \to 0$ as $n \to \infty$, where the expectation is relative to the distribution of the sample. A rule is *universally consistent* if it is consistent for all feature-label distributions. For small samples, consistency is often of little consequence.

To reduce design error, one can constrain optimization to a pre-specified function class. Constraint can reduce design error, but at the cost of increasing the error of the best classifier. The efficacy of using a particular constraint depends on whether the reduction in design error exceeds the cost of constraint.

In this paper we consider two classification rules. For an odd positive integer k, the *k-nearest-neighbor* (kNN) *rule* is defined in the following manner: the k points closest to $\mathbf{x}$ are selected and $\psi_n(\mathbf{x})$ is defined to be 0 or 1 according to which is the majority among the labels of the chosen points. The kNN rule is universally consistent if $k \to \infty$ in such a way that $\frac{k}{n} \to 0$ as $n \to 0$.

We consider constrained optimization in the form of linear discrimination, in which case the classifier is a *perceptron*, defined by $\psi(\mathbf{X}) = T(a_1 X_1 + \cdots + a_m X_m + b)$, where T is a threshold function, $T(z) = 0$ if $z \leq 0$, and $T(z) = 1$ if $z > 0$. A perceptron splits the feature space into two by the hyperplane defined by setting the sum defining ψ equal to 0. Design of a perceptron requires estimating the coefficients $a_1, a_2, \ldots, a_m$, and b. Here we use the method of linear support vector machines to design the perceptron.

We consider two types of error estimation. One approach is to use all sample data to design a classifier ψ_n, and estimate ϵ_n by applying ψ_n to the same data. The *resubstitution estimate*, $\bar{\epsilon}_n$, is the fraction of errors made by ψ_n. Typically, $\bar{\epsilon}_n$ is biased low, meaning $E[\bar{\epsilon}_n] \leq E[\epsilon_n]$. For small samples, the bias can be severe. It improves for large samples.

With *cross-validation*, classifiers are designed from parts of the sample, each is tested on the remaining data, and ϵ_n is estimated by averaging the errors. For *leave-one-out estimation*, n classifiers are designed from sample subsets formed by leaving out one sample pair. Each is applied to the left-out pair, and the estimator $\hat{\epsilon}_n$ is $\frac{1}{n}$ times the number of errors made by the n classifiers. Since the classifiers are designed on sample sizes of $n - 1$, $\hat{\epsilon}_n$ actually estimates the error ϵ_{n-1}. It is an unbiased estimator of ϵ_{n-1}, meaning that $E[\hat{\epsilon}_n] = E[\epsilon_{n-1}]$. Unbiasedness is important, but the variance of the estimator is also a key concern for small n. For the kNN rule, $E[|\hat{\epsilon}_n - \epsilon_n|^2] \leq \frac{(6k+1)}{n}$ (Rogers & Wagner, 1978). Given that $\hat{\epsilon}_n$ is approximately an unbiased estimator of ϵ_n, this inequality bounds the variance of $\hat{\epsilon}_n - \epsilon_n$. Taken together, the approximate unbiasedness and this bound provide information concerning the performance of the leave-one-out estimator. Leave-one-estimation is usually preferred to resubstitution for small samples owing to the bias of the latter; however, the

leave-one-out estimator generally possesses significantly greater variance than the resubstitution estimator, so that it too can produce a dangerously high rate of overly optimistic estimates.

Our concern is with the relationship between the estimated classifier error and the p value for this error based on errors pertaining to randomizations of the data $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$. The intention behind the use of such p values is that they provide useful information in the selection of relevant genes or assessing the quality of the classification. The basic justification comes from the context of hypothesis testing in the following manner. Let the null hypothesis H_0 be that there is no relationship between $\mathbf{X}$ and Y and the alternative hypothesis H_1 be that the contrary is true. The test or decision rule is to reject H_0 in favor of H_1 when the estimated error is small. The following assumption is made: under the null hypothesis, all permutations of the labels are equally likely (see the Appendix for a more rigorous statement of this condition and the manner in which it relates to type-1 error). Let $\tilde{Y}_i$, $i = 1, 2, \ldots, n$, be a sample drawn at random from the set of all permutations of $(y_1, y_2, \ldots, y_n)$ and let $\tilde{\epsilon}_n$ denote the resubstitution error relative to the randomization. $\tilde{\epsilon}_n$ is a random variable relative to the set of all randomizations. The p value of the test is then defined by $p = P(\tilde{\epsilon}_n \leq \bar{\epsilon}_n)$ (Good, 1994). An analogous definition applies to the leave-one-out (or other error) estimate.

3. Simulation studies

To study the relationship between the permutation-test p value and the error estimates, we begin with a simulation study involving the kNN rule, which has been used in a number of expression-based classification studies (Pomeroy et al., 2002; Lesnick, Dacwag, & Golub, 2002; Armstrong et al., 2002; Yeang et al., 2001). We use $k = 3$ and a two-dimensional classifier. We consider two basic models, $M_1(\epsilon)$ and $M_2(\epsilon)$, where ϵ denotes the Bayes error. For $M_1(\epsilon)$, the conditional densities for labels 0 and 1 are normally distributed with $f(\mathbf{x} \mid 0) \sim N(\mu_0, \mathbf{1})$ and $f(x \mid 1) \sim N(\mu_1, \mathbf{1})$, where μ_k is the mean vector, $\mathbf{1}$ is the covariance matrix for uncorrelated marginal densities, each having unit variance, and ϵ is determined by the distance between the means. The Bayes classifier for $M_1(\epsilon)$ is a perceptron. For $M_2(\epsilon)$, the conditional densities for labels 0 and 1 are given by $f(\mathbf{x} \mid 0) = \frac{1}{2}(f_{00}(\mathbf{x}) + f_{01}(\mathbf{x}))$ and $f(\mathbf{x} \mid 1) = \frac{1}{2}(f_{10}(\mathbf{x}) + f_{11}(\mathbf{x}))$, where $f_{00}(\mathbf{x}) \sim N(\mu_{00}, \mathbf{1})$, $f_{01}(\mathbf{x}) \sim N(\mu_{01}, \mathbf{1})$, $f_{10}(\mathbf{x}) \sim N(\mu_{10}, \mathbf{1})$, and $f_{11}(\mathbf{x}) \sim N(\mu_{11}, \mathbf{1})$, and where the means are situated at the corners of a square with μ_{00} diagonally opposite from μ_{01}, μ_{10} diagonally opposite from μ_{11}, and the spacing of the square is such that the Bayes error is ϵ. Experiments are run with errors of $\epsilon = 0.05$, 0.10, and 0.15, and sample sizes of $n = 20$, 40, and 60. This range of sample sizes reflects studies reported in the literature: 38 (Golub et al., 1999); 22 (Hedenfalk et al., 2001); 34, 42, and 60 (Pomeroy et al., 2002); 32, 62, and 72, (Ben-Dor et al., 2000); 17 (Bhattacharjee et al., 2001). The labels are evenly split. For each error and sample size, 500 samples are generated, the kNN classifiers designed, and both the resubstitution and leave-one-out error estimates computed. For each sample, the permutation p value is computed using 4000 re-labelings. For a fixed model, error, and sample size, the 500 samples yield a histogram of error estimates. Relative-frequency histograms for error estimates and plots for the mean p value as a function of the error estimate have been constructed for all cases.

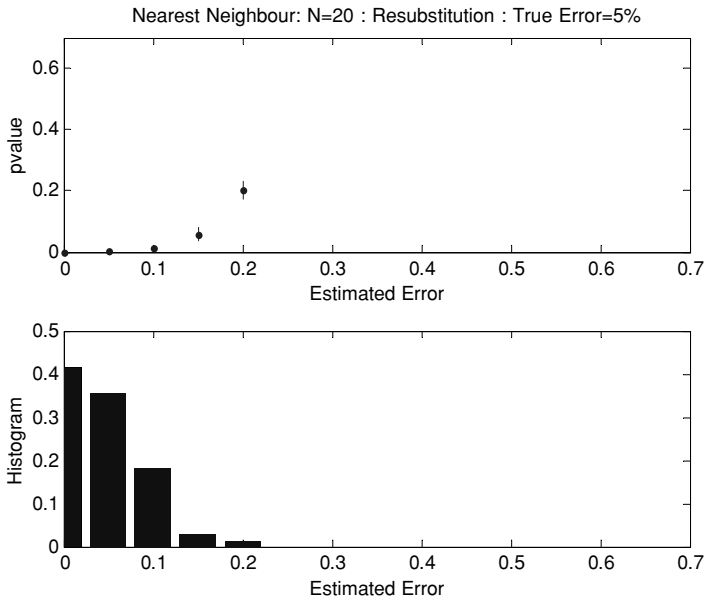


Figure 1. Error histogram and mean p value for resubstitution, 3NN, $n = 20$, model $M_1(0.05)$.

These are provided on our website (http://stat.tamu.edu/∼thsing/pvalue/pvalue.html). We consider some of these in detail here.

The relative-frequency histogram in figure 1 is for $M_1(\epsilon)$, $\epsilon = 0.05$, $n = 20$, and the resubstitution estimate. Above the histogram we see a plot of the mean p value for each observed value of $\bar{\epsilon}_n$, where the vertical bar through each dot denotes the one-standard-deviation range (which in many cases is so small that the bar is not visible outside the dot). Figure 2 corresponds to the same scenario as figure 1, except that the leave-one-out estimator, $\hat{\epsilon}_n$, is used. Before considering the p values, note that the average value of the resubstitution error is less than 0.05, which is low-biased since the error of the designed classifier must exceed 0.05. The average value of the leave-one-out error is somewhat above 0.05, which is indicative of the fact that it is an unbiased estimator of the error of the designed classifier. Note also the greater variance for the leave-one-out estimator.

Turning to the p values, the very low standard deviations indicate that they are tightly regressed on the error estimates for both resubstitution and leave-one-out. In the latter case $E[p] \approx 0$ for $\hat{\epsilon}_n \leq 0.15$. From the histogram we see that the range $\hat{\epsilon}_n \leq 0.15$ represents a large majority of the 500 trials. The p value only increases for a small minority of the cases. Similar phenomena are observed in figure 1 for resubstitution, where a large majority of errors are either 0 or 0.05, and $E[p] \approx 0$ for those cases. The regression of the p value on the error estimates becomes even tighter as the sample size increases, as evidenced by figures 3 and 4, which correspond to figures 1 and 2, except that $n = 40$, and by figures 5 and 6, for which $n = 60$. In the latter case, for both error estimators, $E[p] \approx 0$ for all but a few extreme error values. Figures 7 through 12 correspond to figures 1 through 6, except that the Bayes error is 0.10. Similar phenomena can be observed.

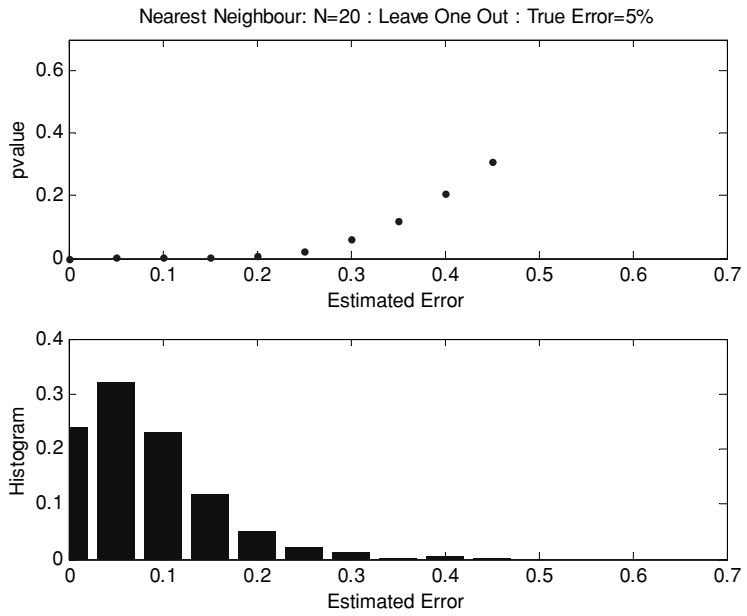


Figure 2. Error histogram and mean p value for leave-one-out, 3NN, $n = 20$, model $M_1(0.05)$.

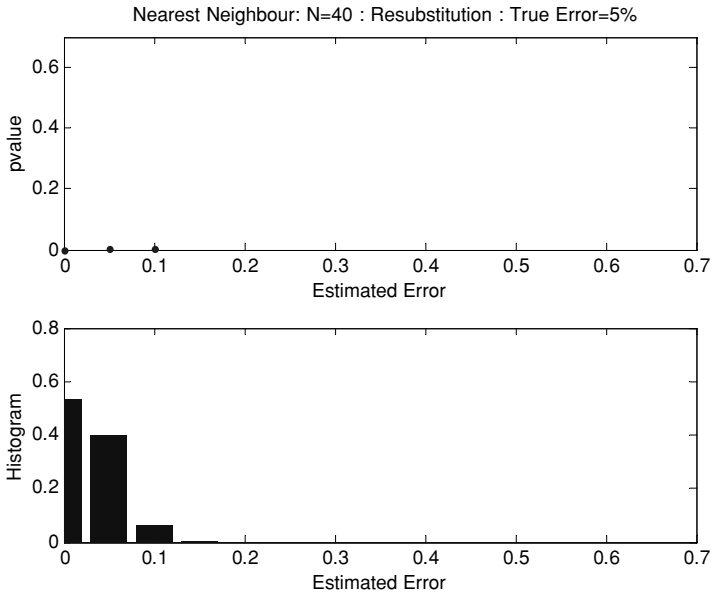


Figure 3. Error histogram and mean p value for resubstitution, 3NN, $n = 40$, model $M_1(0.05)$.

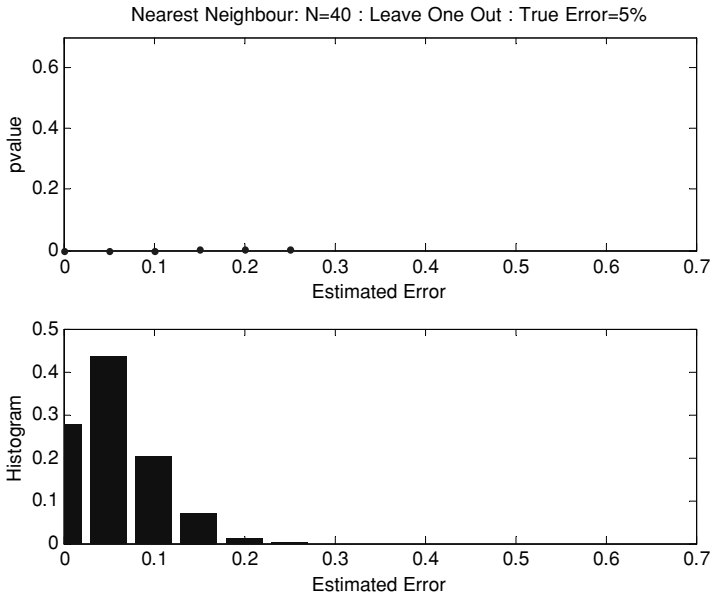


Figure 4. Error histogram and mean p value for leave-one-out, 3NN, $n = 20$, model $M_1(0.05)$.

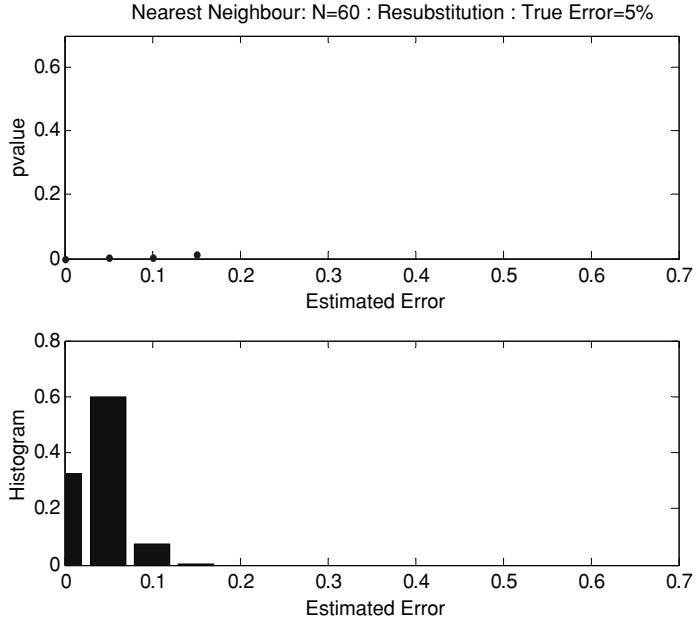


Figure 5. Error histogram and mean p value for resubstitution, 3NN, $n = 60$, model $M_1(0.05)$.

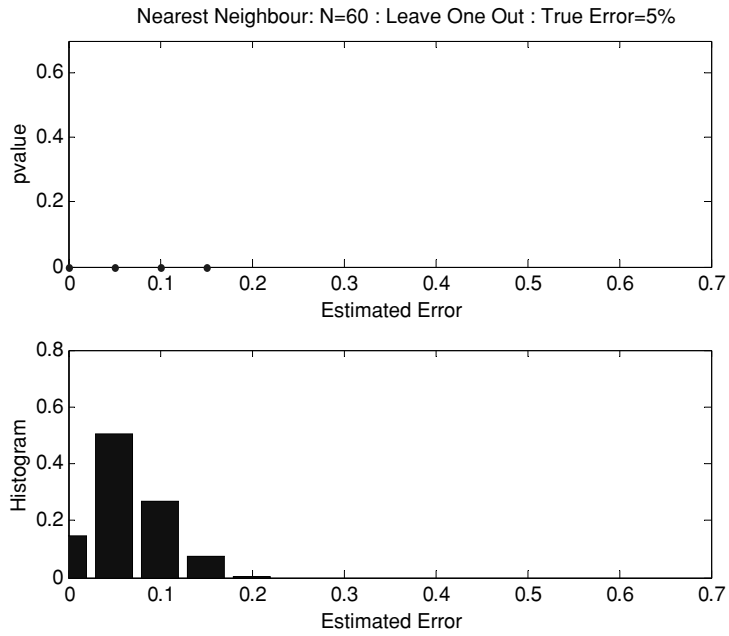


Figure 6. Error histogram and mean p value for leave-one-out, 3NN, $n = 20$, model $M_1(0.05)$.

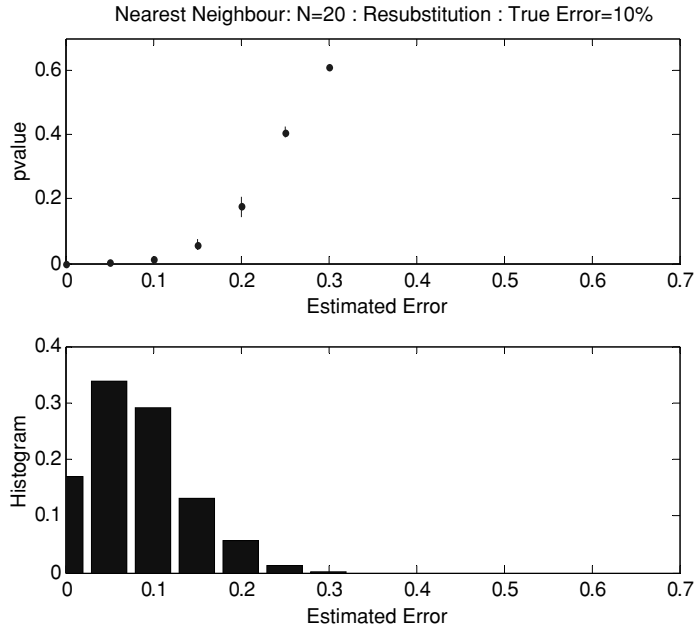


Figure 7. Error histogram and mean p value for resubstitution, 3NN, $n = 20$, model $M_1(0.10)$.

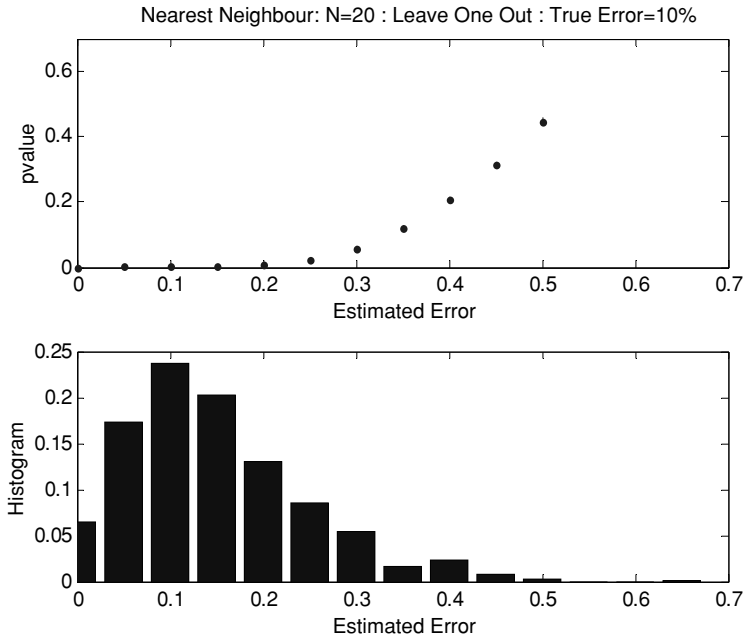


Figure 8. Error histogram and mean p value for leave-one-out, 3NN, $n = 20$, model $M_1(0.10)$.

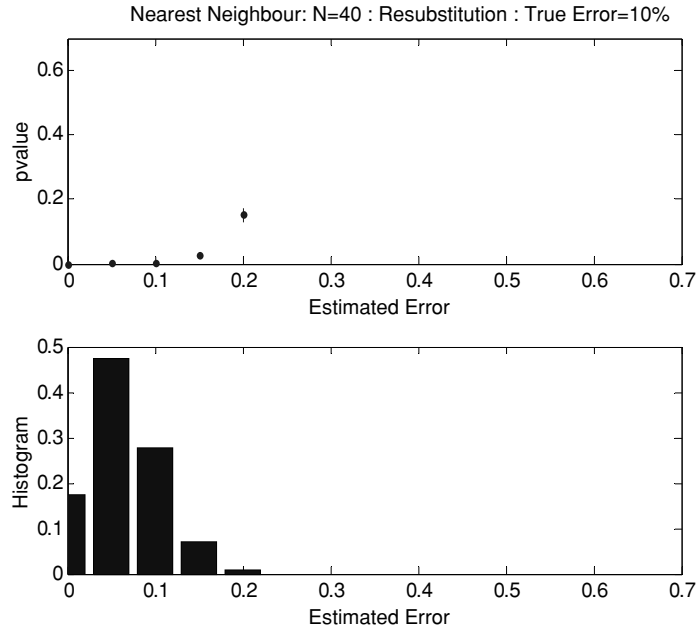


Figure 9. Error histogram and mean p value for resubstitution, 3NN, $n = 40$, model $M_1(0.10)$.

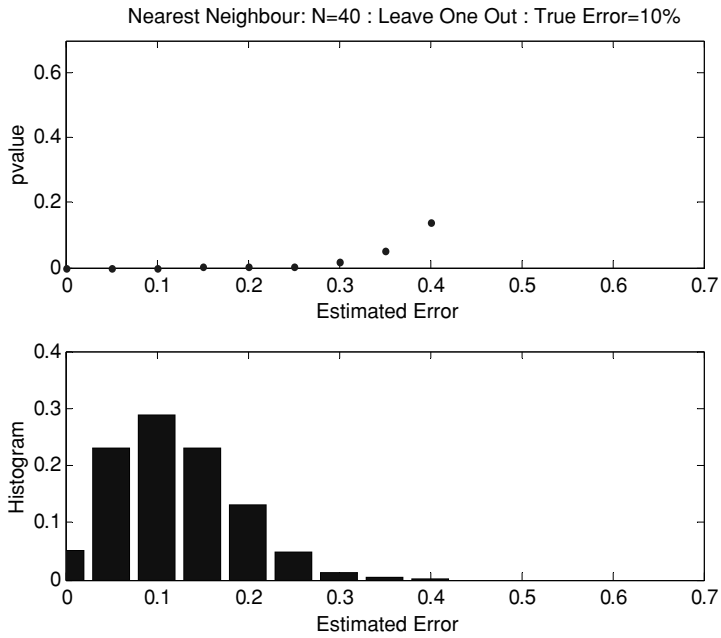


Figure 10. Error histogram and mean p value for leave-one-out, 3NN, $n = 20$, model $M_1(0.10)$.

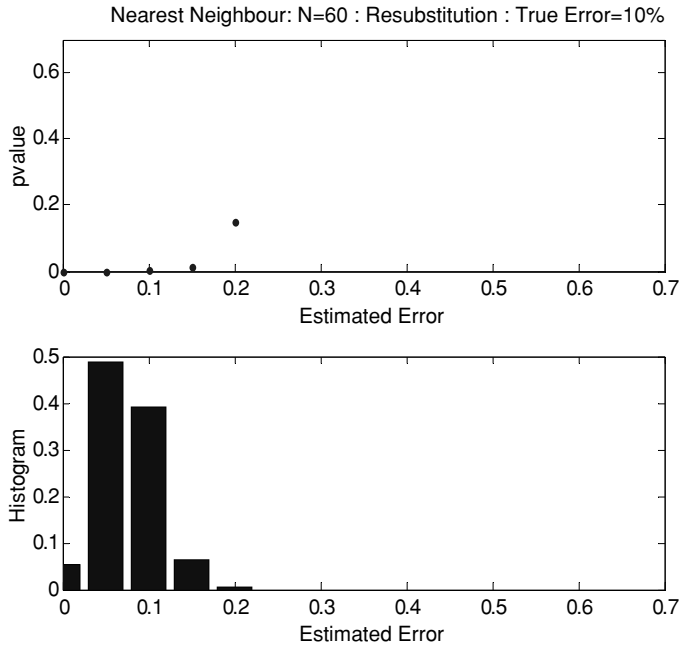


Figure 11. Error histogram and mean p value for resubstitution, 3NN, $n = 60$, model $M_1(0.10)$.

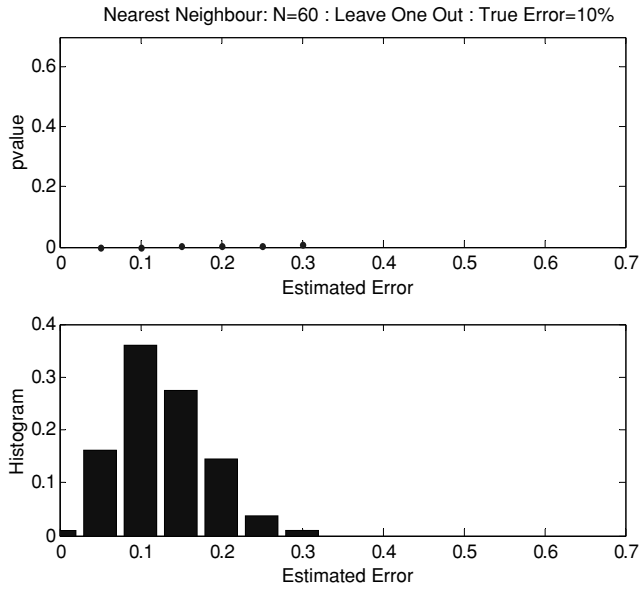


Figure 12. Error histogram and mean p value for leave-one-out, 3NN, $n = 20$, model $M_1(0.10)$.

We note three critical points. First, the fact that the variance of p for a fixed error estimate is typically so small that the standard-deviation bars are not visible means that p is essentially a function of the error estimate. Second, $E[p] \approx 0$ for most of the trials, the exceptions being the minority of larger error estimates, and in the case of leave-one-out estimation, these exceptions corresponding to substantially high-biased estimates relative to the Bayes error. Third, not only is p virtually a function of the error estimate, but that function is also non-invertible because $E[p] \approx 0$ for all moderate error estimates. Hence, not only is p non-informative beyond the error estimate, it is less informative.

All of the cases studied for the kNN classifier with model $M_1(\epsilon)$ have also been studied for model $M_2(\epsilon)$, for which classification is more difficult. Here we only show two cases, for Bayes error 0.10, leave-one-out estimation, and samples sizes 40 and 60 (figures 13 and 14, respectively). All other cases are on the website. The difficulty of this model is reflected in figure 13, where $n = 40$ and the unbiased error estimate averages close to 0.30. But even here, $E[p] \approx 0$ for 0.30, which is significantly in excess of the Bayes error. When $n = 60$ (figure 14), the average error estimate is slightly lower, but again $E[p] \approx 0$ for 0.30.

Lastly, figures 15 and 16 show results for a linear support vector machine (SVM) applied to model $M_1(\epsilon)$ [for which the Bayes classifier is a perceptron], with Bayes error 0.10, leave-one-out estimation, and sample sizes 20 and 40. Similar observations can be made as those pertaining to kNN classification in regard to the relation between the p value and the error estimate. The full body of SVM simulations appears on the website. We note that SVMs have been used for expression-based classification (Yeang et al., 2001; Ramaswamy et al., 2001).

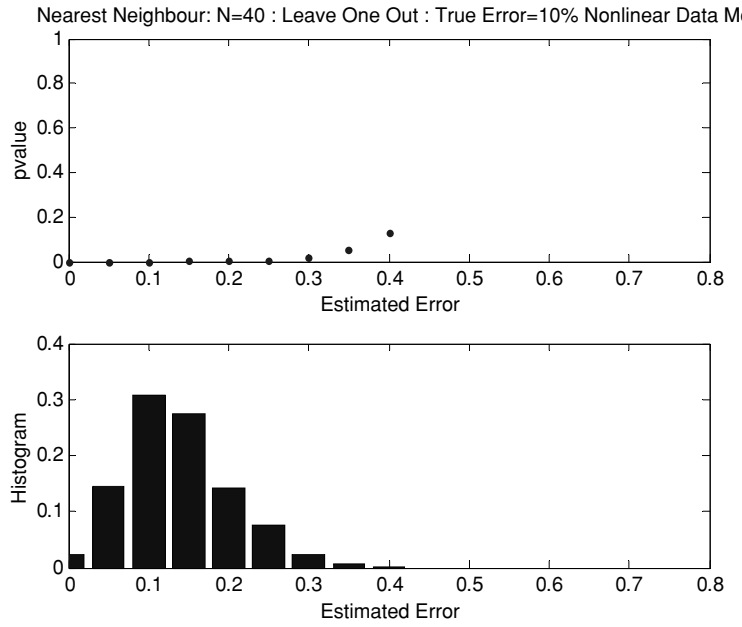


Figure 13. Error histogram and mean p value for leave-one-out, 3NN, $n = 40$, model $M_2(0.10)$.

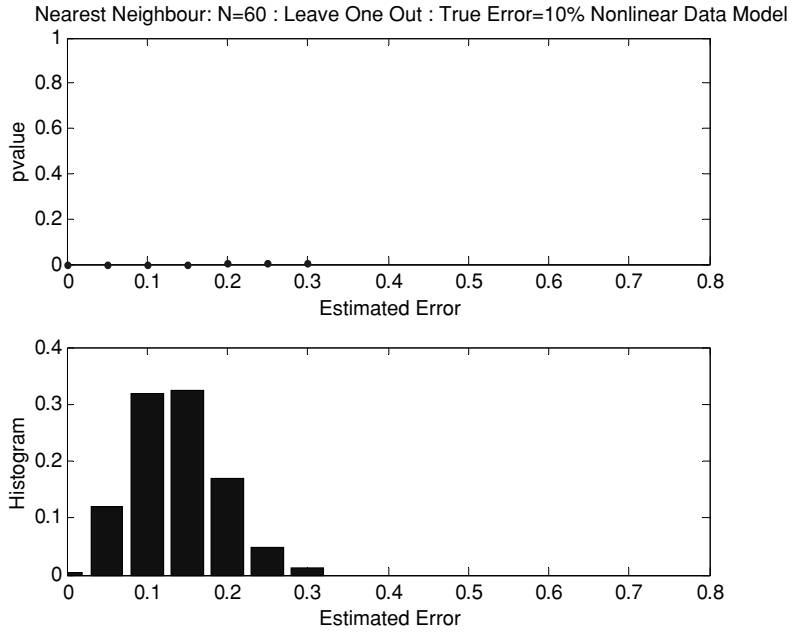


Figure 14. Error histogram and mean p value for leave-one-out, 3NN, $n = 60$, model $M_2(0.10)$.

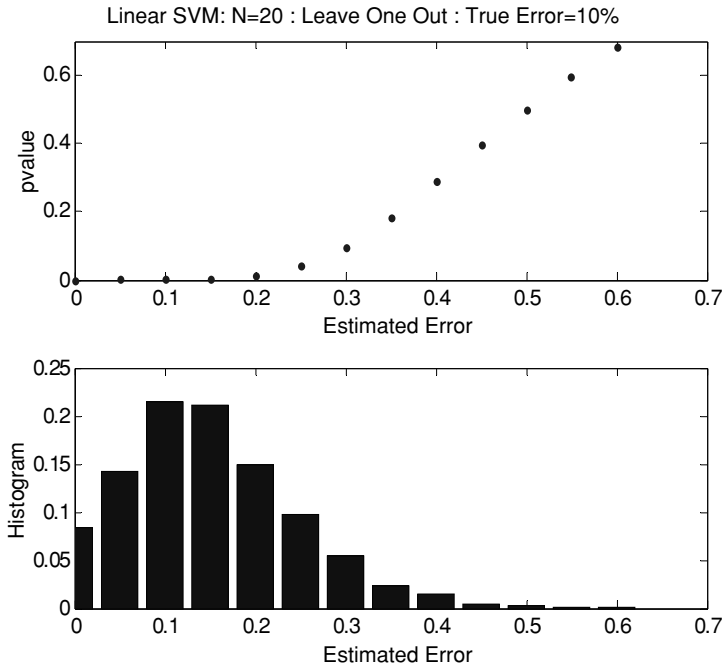


Figure 15. Error histogram and mean p value for leave-one-out, SVM, $n = 20$, model $M_1(0.10)$.

4. Theoretical considerations

In this section we wish to achieve two goals. First we derive an approximate analytic formulation of the permutation-test p value as a function of the resubstitution error for single-variable linear discrimination. With that we intend to argue that p value computation in that case does not provide insight into the true labeling distribution beyond $\bar{\varepsilon}_n$. The second goal is to demonstrate that the permutation distribution of the sample generally does not yield useful information on the true distribution from which the sample is drawn. To illustrate the generality of this statement, we consider the permutation distribution of the sample correlation of a bivariate normal sample.

An observation x is classified as 0 or 1 depending on whether it is closer to mean$(x^{(0)})$ or mean$(x^{(1)})$, where mean$(x^{(0)})$ and mean$(x^{(1)})$ are the averages of the x's for which $y = 0$ and $y = 1$, respectively. Letting I denote the indicator function,

$$\begin{aligned}\psi(x) = {} & I\big(\text{mean}\big(x^{(0)}\big) \le \text{mean}\big(x^{(1)}\big), x > \text{mean}(x)\big) \\ & + I\big(\text{mean}\big(x^{(0)}\big) > \text{mean}\big(x^{(1)}\big), x \le \text{mean}(x)\big)\end{aligned}$$

where

$$\text{mean}(x) = \big[\text{mean}\big(x^{(0)}\big) + \text{mean}\big(x^{(1)}\big)\big]/2.$$

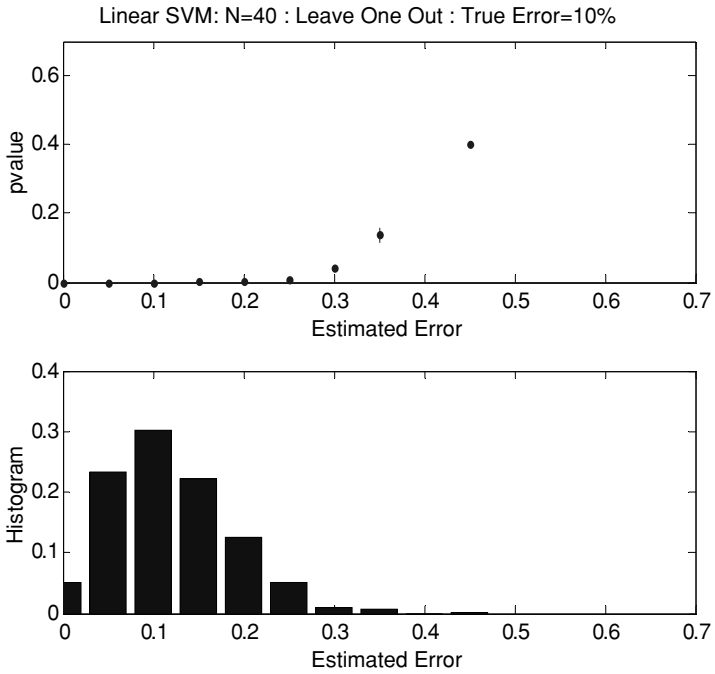


Figure 16. Error histogram and mean p value for leave-one-out, SVM, $n = 40$, model $M_1(0.10)$.

We compute the probability density for the randomization resubstitution error

$$P(\tilde{\epsilon}_n = k/n), \quad 1 \le k \le n.$$

Let mean$(\tilde{x}^{(0)})$, mean$(\tilde{x}^{(1)})$, mean$(\tilde{x})$ be the same averages defined above but based on the randomized labels $\tilde{Y}_1, \ldots, \tilde{Y}_n$. Let

$$U = \sum_{i=1}^{n} I(x_i \le \text{mean}(\tilde{x}), \tilde{Y}_i = 0), \ \bar{U} = \sum_{i=1}^{n} I(x_i \le \text{mean}(\tilde{x})) - U$$

and

$$V = \sum_{i=1}^{n} I(x_i > \text{mean}(\tilde{x}), \tilde{Y}_i = 1), \quad \bar{V} = \sum_{i=1}^{n} I(x_i > \text{mean}(\tilde{x})) - V.$$

It is easy to see that $\tilde{\epsilon}_n = k/n$ if and only if

$$U + V = k \ \text{ and } \text{mean}\big(\tilde{x}^{(0)}\big) > \text{mean}\big(\tilde{x}^{(1)}\big)$$

or

$$\bar{U} + \bar{V} = k \text{ and } \text{mean}\big(\tilde{x}^{(0)}\big) \le \text{mean}\big(\tilde{x}^{(1)}\big)$$

Hence

$$\begin{aligned} P(\tilde{\epsilon}_n = k/n) = {} & P\big(U + V = k \text{ and } \text{mean}\big(\tilde{x}^{(0)}\big) > \text{mean}\big(\tilde{x}^{(1)}\big)\big) \\ & + P\big(\bar{U} + \bar{V} = k \text{ and } \text{mean}\big(\tilde{x}^{(0)}\big) \le \text{mean}\big(\tilde{x}^{(1)}\big)\big). \end{aligned}$$

We focus on the first term since the computations are identical. In general, the exact computations of these quantities are tedious. To illustrate our point, we resort to simplification and approximations. Expressing the probability by the product rule $P(A \cap B) = P(A)P(B \mid A)$,

$$\begin{aligned} & P\big(U + V = k \text{ and } \text{mean}\big(\tilde{x}^{(0)}\big) > \text{mean}\big(\tilde{x}^{(1)}\big)\big) \\ & \quad = P(U + V = k) P\big(\text{mean}\big(\tilde{x}^{(0)}\big) > \text{mean}\big(\tilde{x}^{(1)}\big) \,\big|\, U + V = k\big). \end{aligned}$$

First we deal with the conditional probability. We focus on those k which are small relative to n since such values are of the most relevance in p value computation. Observe that if we take a randomized sample $\tilde{y}_1, \ldots, \tilde{y}_n$ for which $U + V = k$ where k is small relative to n, then it is almost always true (so long as there is not an extraordinary number of extreme outliers) that $\text{mean}(\tilde{x}^{(0)}) > \text{mean}(\tilde{x}^{(1)})$. This leads to the approximation

$$P\big(\text{mean}\big(\tilde{x}^{(0)}\big) > \text{mean}\big(\tilde{x}^{(1)}\big) \,\big|\, U + V = k\big) \approx 1 \text{ if } k/n \text{ is small.}$$

We conclude that

$$P(\tilde{\epsilon}_n = k/n) \approx P(U + V = k) + P(\bar{U} + \bar{V} = k) \text{ if } k/n \text{ is small,}$$

and therefore

$$P(\tilde{\epsilon}_n \le u) \approx \sum_{k \le nu} [P(U + V = k) + P(\bar{U} + \bar{V} = k)] \text{ for small } u.$$

In particular, if the resubstitution error, $\bar{\epsilon}_n$, for the actual labeling is small, then we have

$$p \approx \sum_{k \le n\bar{\epsilon}_n} [P(U + V = k) + P(\bar{U} + \bar{V} = k)]. \tag{1}$$

We will see below in a simulation that this approximation works very well. With this formula we intend to make the point that for the classifier considered in this section, p value computation does not lead to any useful information on labeling beyond what is in $\bar{\epsilon}_n$, as we do now. In order to make our point more easily understood, we make the somewhat

restrictive assumption that there is an equal number of 0's and 1's in the y_i so that mean($\tilde{x}$) $\equiv$ mean(x). Then straightforward combinatorics show that

$$P(U+V=k)=P(\bar{U}+\bar{V}=k)=\frac{\binom{n_-}{(n/2-n_++k)/2}\binom{n_+}{(n/2+n_+-k)/2}}{\binom{n}{n/2}}$$

where $\binom{m_1}{m_2}=\frac{m_1!}{m_1!(m_2-m_1)!}$ whenever $0\le m_2\le m_1$ and m_1, m_2 are nonnegative integers, and 0 otherwise, where $n_-=\sum_{i=1}^{n} I(x_i<\text{mean}(x))$ and $n_+=n-n_-$. Thus we obtain from (1) that

$$p\approx 2\sum_{k\le n\bar{\epsilon}_n}\frac{\binom{n_-}{(n/2-n_++k)/2}\binom{n_+}{(n/2+n_+-k)/2}}{\binom{n}{n/2}}.$$

Observe that to compute p value we only need $\bar{\epsilon}_n$, n_- and n_+. Since n_- and n_+ contain no information on labeling, we conclude that the only useful information in this computation with regard to true labeling distribution comes solely from $\bar{\epsilon}_n$. To further illustrate our point, note that whenever the average of x is close to the median of x we have $n_+\approx n_-\approx n/2$, in which case

$$p\approx 2\sum_{k\le n\bar{\epsilon}_n}\frac{\binom{n/2}{k/2}^2}{\binom{n}{n/2}}.$$

To evaluate the quality of this approximation, we generated a random sample of size 10 from Normal(0,1): .0808, −1.5264, 1.4507, .1027, −1.2031, .7720, −1.3606, −.2431, 1.4761, −1.1099, and another sample of size 10 from Normal(2,1): 1.1903, 3.1520, 1.9497, 2.6552, 3.9009, 3.1682, 1.2221, 1.4481, .7861, 2.1931. For the combined sample the mean is equal to the median. Assuming that half of the labels are equal to 0 and the other half equal to 1, we simulated 1,000,000 permutations. Then $P(\bar{\epsilon}_n=k/n)$ are computed based on simulated value $P_s(\bar{\epsilon}_n=k/n)$ as well as on the approximated values $P_a(\bar{\epsilon}_n=k/n)=\binom{n/2}{k/2}^2/\binom{n}{n/2}$:

k	0	2	4	6	8
$P_s(\bar{\epsilon}_n=k/n)$	.000014	.001036	.022013	.155865	.433360
$P_a(\bar{\epsilon}_n=k/n)$	.000011	.001082	.021921	.155881	.477386

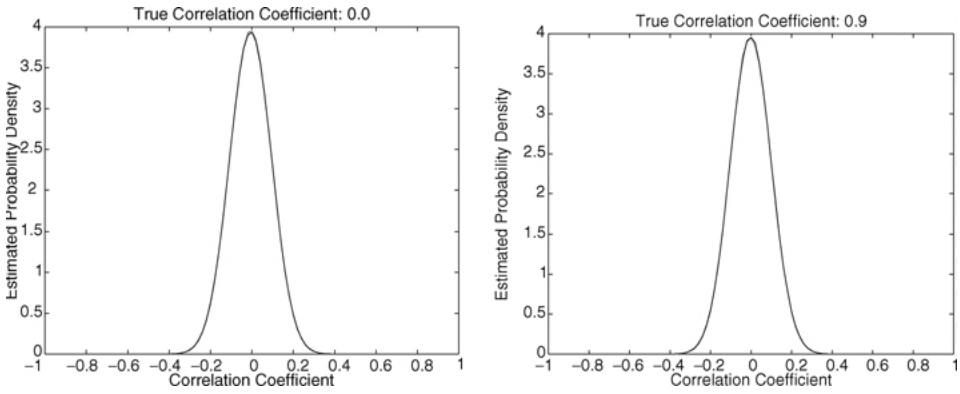
The quality of approximation in general is very similar to what is demonstrate by this set of data.

We now turn to the second goal of this section. For that let us go beyond the classification problem and consider a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ obtained from the bivariate

normal distribution with mean 0, variance 1, and correlation ρ. The sample correlation

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(X_i - \bar{X})^2\right]^{1/2}}$$

is a good estimate of the true correlation ρ. We will show below that the randomized distribution of $\hat{\rho}$ contains no information on ρ (just as the randomized distribution of the error rate contains no information on the true error rate in many classification situations). From the normal distributions with $\rho = 0$ and $\rho = .9$, we simulated 10 random samples each of size 100. For each of these samples we obtained the permutation distribution of $\hat{\rho}$ based on 1,000,000 random permutations, and estimated the probability density function using the histogram method. In the following the graph on the left (right) is the superimposed plot of the estimated probability density functions based on the samples for $\rho = 0$ ($\rho = .9$). Observe that the probability densities are virtually indistinguishable. This phenomenon is quite general and is not limited to the normal distribution and the statistic $\hat{\rho}$. A comprehensive theory that explores these issues from a theoretical perspective can be found in Bai and Hsing.



5. Conclusion

Permutation tests offer the possibility of obtaining useful statistics to discover genes that can be used to discriminate between phenotypes based on expression measurements; however, if a permutation-test p value is to be considered worthwhile, it must carry information beyond the statistic used in its derivation and beyond that carried by the error estimates and their properties that have been historically developed in the theory of pattern recognition. This paper has focused on p values arising directly from the error estimates, and in that sense it represents a natural first step. Other p values based on randomized re-labeling have been proposed in the microarray literature. For instance, one defines a weight θ for each gene to measure its discriminating power between two classes (Allander et al., 2001); another

defines a score θ derived from error measurements (Ben-Dor et al., 2000), and another defines a distance between the vector of class labels and expression measurements for a gene, and then counts the number of genes within a given distance of the gene to obtain a statistic θ (Golub et al., 1999; Slonim et al., 2000). In each case, θ is calculated from the actual data and the p value of θ is computed relative to the scores derived from re-labeling. These and other permutation methods proposed in the future to help find good features and classifiers in the context of gene expression need to be examined to understand the manner and extent to which they are informative. We believe that the results in this paper point to the need of rigorously analyzing all permutation methods proposed to help find good features and classifiers in the context of gene expression so that the manner and extent to which they are informative is well understood.

Appendix

We elucidate on the uniformity assumption for permutation testing commented upon previously. Let $\xi(\mathbf{S})$ be a statistic based on a sample $\mathbf{S} = (S_1, S_2, \ldots, S_n)$, and let the decision rule be to reject the null hypothesis H_0 if $\xi(\mathbf{s}) > c(\mathbf{s})$, where $c(\mathbf{s})$ is the $(1-\alpha)100\%$-th percentile of the permutation distribution of $\xi(\mathbf{s})$. Consider the decomposition of the whole sample space of $\mathbf{s}$ induced by the equivalence relation $\sim$ where $\mathbf{s} \sim \mathbf{r}$ if $\mathbf{s}$ and $\mathbf{r}$ are permutations of one another. Then the type-I error probability is

$$P_0(\text{reject } H_0) = P_0(\xi(\mathbf{S}) > c(\mathbf{S})) = E_0[P_0(\xi(\mathbf{S}) > c(\mathbf{S}) \mid \mathcal{K}(\mathbf{S}))]$$

where $\mathcal{K}(\mathbf{s})$ denotes the equivalence class containing $\mathbf{s}$ and P_0 and E_0 are probability and expectation, respectively, under H_0. Under the assumption that all members of an equivalence class are equally likely under H_0, we have by the definition of $c(\mathbf{s})$

$$P_0(\xi(\mathbf{S}) > c(\mathbf{S}) \mid \mathcal{K}(\mathbf{S}) = \mathcal{K}(\mathbf{s})) = \frac{\sum_{\mathbf{r}\in\mathcal{K}(\mathbf{s})} I(\xi(\mathbf{r}) > c(\mathbf{s}))}{|\mathcal{K}(\mathbf{s})|} = \alpha,$$

where I and $|\mathcal{K}(\mathbf{s})|$ denote the indicator function and the number of elements in $\mathcal{K}(\mathbf{s})$, respectively. Unconditioning yields $P_0(\text{reject } H_0) = \alpha$.

References

Allander, S. V., Nupponen, N. N., Ringner, M., Hostetter, G., Maher, Goldberger, N., Chen, Y., Carpten, J., Elkahloun, A. G., & Meltzer, P. S. (2001). Gastrointestinal stromal tumors with KIT mutations exhibit a remarkably homogeneous gene expression profile. *Cancer Research, 61*, 8624–8628.

Armstrong, S. A. et al. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics, 30*, 41–47.

Bai, Z., & Hsing, T. The broken sample problem. Available at http://stat.tamu.edu/∼thsing/papers/link.pdf.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., & Yakhini, Z. (2000). Tissue classification with gene expression profiles. *Computational Biology, 7*, 559–583.

Bhattacharjee, A. et al. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences, 98:24*, 13790–13795.

De Risi, J. L., Iyer, V. R., & Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science, 278*, 680–686.

Devroye, L., Gyorfi, L., & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag.

Dougherty, E. R. (2001). Small sample issues for microarray-based classification. *Comparative and Functional Genomics, 2*, 28–34.

Duggan, D. J., Bittner, M. L., Chen, Y., Meltzer, P. S., & Trent, J. M. (1999). Expression profiling using cDNA microarrays. *Nature Genetics, 21*, 10–14.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 286*, 531–537.

Good, P. (1994). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypothesis*. New York: Springer-Verlag.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvverger, S., Loman, N., Johannsson, O., Olsson, H., Wifond, B., Sauter, G., Kallioniemi, O. P., Borg, A., & Trent, J. (2001). Gene expression profiles distinguish hereditary breast cancers. *New England Journal of Medicine, 34*, 539–548.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., & Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Natture Medicine, 7*, 673–679.

Lesnick, S. T., Dacwag, C. S., & Golub, T. R. (2002). The Ewing's sarcoma oncoprotein EWS/FLI induces a p53-dependent growth arrest in primary human fibroblasts. *Cancer Cell, 1*, 393–401.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, Kim, J. Y. H., Goumnerova, L. C., Black, P., Lau, C., Allen, J. C., Zagzag, D., Olson, J. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, C., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D. N., Mesirov, J. P., Lander, E. S., & Golub, T. R. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature, 415*, 436–442.

Ramaswamy, S. et al. (2001). Multi-class cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences, 98*, 15149–15154.

Rogers, W., & Wagner, T. (1978). A finite sample distribution-free performance bound for local discrimination rules. *Annals of Statistics, 8*, 506–514.

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science, 270*, 467–470.

Slonim, D. K., Tamayo, P., Mesirov, J. P., Golub, T. R., & Lander, E. S. (2000). Class prediction and discovery using gene expression aata. *Annual Conference on Research in Computational Molecular Biology*, Tokyo.

Vapnik, V., & Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications, 16*, 264–280.

Yeang, C.-H. et al. (2001). Molecular classification of multiple tumor types. Bioinformatics. 17(Supplement 1): S316–S322.

Received June 18, 2002
Revised November 6, 2002
Accepted November 8, 2002
Final manuscript November 12, 2002