Permutation Tests for Classification

Polina Golland POLINA@CSAIL.MIT.EDU
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Feng Liang FENG@STAT.DUKE.EDU
Institute of Statistics and Decision Sciences
Duke University
Durham, NC 27708, USA

Sayan Mukherjee SAYAN@STAT.DUKE.EDU
Institute for Genome Sciences and Policy
Institute of Statistics and Decision Sciences
Duke University
Durham, NC 27708, USA

Dmitry Panchenko PANCHENK@MATH.MIT.EDU
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02451, USA

Editor:

Abstract

Permutation tests have been proposed for a variety of problems going back to the early works of Fisher. We describe a permutation procedure used extensively in classification problems in computational biology and medical imaging. We empirically study the procedure on simulated data and real examples from neuroimaging studies and DNA microarray analysis. A theoretical analysis is suggested to assess the asymptotic behavior of the test. An interesting observation is that concentration of the permutation procedure is controlled by a Rademacher average which also controls the concentration of empirical errors to expected errors. A byproduct of the analysis is a uniform central limit theorem for a permutation procedure.

Keywords: Classification, Permutation testing, Statistical significance, Non-parametric tests, Rademacher processes.

1. Introduction

Many scientific studies involve detection and characterization of predictive patterns in high dimensional measurements, which can often be reduced to training a binary classifier or a regression model. We will use two examples of such applications to illustrate the techniques in this paper: medical image studies and gene expression analysis. Image-based clinical studies of brain disorders attempt to detect neuroanatomical changes induced by diseases, as well as predict development of the disease. The goals of gene expression analysis include classification of the tissue morphology and prediction of the treatment outcome from DNA microarray data. In both fields, training a classifier to reliably label new examples into the healthy population or one of the disease sub-groups can help to improve screening and early diagnostics, as well as provide an insight into the nature of the disorder. Both imaging data and DNA microarray measurements are characterized by high dimensionality of the input space (thousands of features) and small datasets (tens of independent examples), typical of many biological applications.

A basic question is in this setting is how can one have any modicum of faith in the accuracy of the trained classifier. One approach to this problem would be to estimate the test error on a hold-out set – or by applying a cross-validation procedure, such as a jackknife (Efron, 1982) – which, in conjunction with a variance-based convergence bound, provides a confidence interval for the expected error. Small sample sizes render this approach ineffective as the variance of the error on a hold-out set is often too large to provide a meaningful estimate on how close we are to the true error. Applying variance-based bounds to the cross-validation error estimates produces misleading results as the cross-validation iterations are not independent, causing us to underestimate the variance. Classical generalization bounds are also not appropriate in this regime due to the high dimensionality and small sample size. In addition, even if a consistent algorithm is used that produces a classifier with low variance the data itself may have no structure. Neither cross-validation nor classical generalization bounds address this issue.

Recently, several research groups, including ours, proposed using the classical idea of permutation tests (Good, 1994) to assess the reliability of the classifier's accuracy via a notion of statistical significance (Golub et al., 1999, Slonim et al., 2000, Mukherjee et al., 2003, Pomeroy et al., 2002, Golland and Fischl, 2003). Intuitively, statistical significance is a measure of how likely the observed accuracy would be obtained by chance, only because the training algorithm identified some pattern in the high-dimensional data that happened to correlate with the class labels as an artifact of a small data set size. A significant classifier would reject the null hypothesis that the features and the labels are independent, that is, there is no difference between the two classes. Typically a cross-validation error is used as a test statistic that measures how different the two classes are with respect to the family of classifiers we use in training, and its distribution under the null hypothesis is estimated by permuting the labels.

The objective of this paper is to examine with some care permutation tests for classification both empirically and theoretically so as to provide users with some practical recommendations and suggest a theoretical basis to the procedure. After a quick overview of permutation testing the paper is organized as follows. The next section describes the permutation procedure to estimate statistical significance of classification results. Section 3 applies the procedure to simulated data as well as real data from the fields of brain imaging and gene expression analysis and offers practical guidelines for applying the procedure. In Section 4, we suggest a theoretical analysis of the procedure that leads to convergence bounds governed by similar quantities to those that control standard empirical error bounds, closing with a brief discussion of open questions.

1.1 Related work and some history

The idea of permutation tests date to at least the 1930s in the works of Fisher (1936) and Pitman (1937). The following quote by Fisher on permutation tests in 1936 expresses the centrality of permutation testing in statistics:

> *"Actually, the statistician does not carry out this very tedious process but his conclusions have no justification beyond the fact that they could have been arrived at by this very elementary method."*

With the advances of computational tools and techniques since the 1930s this "tedious process" can be carried out by the statistician's computer with little effort.

Theoretical results on the concentration of permuted sequences as well as permutation tests for i.i.d. draws from a sequence or fixed function were developed in the 1940s and 1950s (Lehman and Stein, 1949, Hoeffding, 1951, Wald and Wolfowitz, 1944). The application of permutation tests to problems where the null distributions was unknown or distribution free statistics were needed was developed and analyzed from the 1950s (Dwass, 1957, Box and Anderson, 1955, Arnold, 1964, W. Albers, 1976, Bell and Doksum, 1965). The permutation procedure is closely related to the bootstrap procedure (Efron and Tibshirani, 1993) and often results and observations for the two procedures are related. The above historical description is by no means complete and overlooks many other seminal papers on the topic of permutation tests. For an excellent book on the topic see Good (1994).

Recent advances in technologies in a variety of biological and physical sciences have resulted in data that is very high-dimensional. In the statistical analysis of these data sets it is natural to use permutation tests since the underlying model upon which to base the null distribution maybe very complicated, unknown, or not computable. One such application area is the selection of a relevant feature from thousands of features: identification of significant changes in individual gene expression levels (Ben-Dor et al., 2000, Troyanskaya et al., 2002), localization of voxels with significantly different MRI signal intensities in two populations (Bullmore et al., 1999, Nichols and Holmes, 2001). Another application area where several research groups used permutation tests is in the assessment of statistical significance of cross-validation error (Golub et al., 1999, Slonim et al., 2000, Mukherjee et al., 2003, Pomeroy et al., 2002, Golland and Fischl, 2003). Our paper will focus on this later setting.

A notion of statistical significance or of variance does not always add more information to the classification problem than the classification error. For example, for a fixed classifier Hsing et al. (2003) shows that statistical significance estimates carry at most as much information as the classification error. This is due to the fact that a fixed classifier can be modeled as a Bernoulli distribution and the variance will be determined by the mean, which is an estimate of the classifiers accuracy. However, this will not hold for a family of classifiers, the family needs to be restricted to control the variance and for a uniform law of large numbers or central limit theorem to hold. We will derive necessary and sufficient conditions for such a central limit theorem in Section 4.

2. Permutations Tests for Classification

In two-class comparison hypothesis testing, the differences between two data distributions are measured using a data set statistic

$$\mathcal{T} : (\mathbb{R}^n \times \{-1, 1\})^l \mapsto \mathbb{R},$$

such that for a given data set $S = \{(\mathbf{x}_k, y_k)\}_{k=1}^l$, where $\mathbf{x}_k \in \mathbb{R}^n$ are observations and $y_k \in \{-1, 1\}$ are the corresponding class labels, $\mathcal{T}(\mathbf{x}_1, y_1, \ldots, \mathbf{x}_l, y_l)$ is a measure of the similarity of the subsets $\{\mathbf{x}_k | y_k{=}1\}$ and $\{\mathbf{x}_k | y_k{=}{-1}\}$. The null hypothesis typically assumes that the two conditional probability distributions are identical, $p(\mathbf{x}|y{=}1) = p(\mathbf{x}|y{=}{-1})$, or equivalently, that the data and the labels are independent, $p(\mathbf{x}, y) = p(\mathbf{x})p(y)$. The goal of the hypothesis test is to reject the null hypothesis at a certain level of significance α which sets the maximal acceptable probability of false positive (declaring that the classes are different when the null hypothesis is true). For any value of the statistic, the corresponding *p-value* is the highest level of significance at which the null hypothesis can still be rejected.

The test statistics used in this paper are training errors, cross-validation errors, or jackknife estimates. Here we give as an example the jackknife estimate

$$\mathcal{T}(\mathbf{x}_1, y_1, \ldots, \mathbf{x}_l, y_l) = \frac{1}{l} \sum_{i=1}^{l} I(f_{S^i}(x_i) \neq y_i),$$

where S^i is the data set with the ith sample removed and f_{S^i} is the function obtained by the classification algorithm given the data set S^i and $I(\cdot)$ is the indicator function.

Suppose we have chosen an appropriate statistic $\mathcal{T}$ and the acceptable significance level α. Let Π_l be the set of all permutations of the samples $(\mathbf{x}_i)_{i=1}^l$, where for the permutation π, $\mathbf{x}_i^\pi$ is the i-th sample after permutation. The permutation test procedure is described as follows:

- Repeat M times (with index $m = 1, \ldots, M$):
  - sample a permutation $\boldsymbol{\pi}^m$ from a uniform distribution over Π_l,
  - compute the statistic value for this permutation of samples

$$t^m = \mathcal{T}(\mathbf{x}_1^m, y_1, \ldots, \mathbf{x}_l^m, y_l).$$

- Construct an empirical cumulative distribution (ecdf)

$$\hat{P}(T \leq t) = \frac{1}{M} \sum_{m=1}^{M} \Theta(t - t^m),$$

  where the step function $\Theta(x - y) = 1$ if $x \geq y$ and otherwise is 0.

- Compute $t_0 = \mathcal{T}(\mathbf{x}_1, y_1, \ldots, \mathbf{x}_l, y_l)$ and the corresponding p-value $\hat{p}_0 = \hat{P}(t_0)$. If $\hat{p}_0 \leq \alpha$, then reject the null hypothesis.

Ideally, we would like to use the entire set of permutations Π_l to calculate the corresponding p-value p_0, but it might be not feasible for computational reasons. Instead, we resort to sampling from Π_l and use Monte Carlo methods to approximate p_0. The Monte Carlo approximation $\hat{p}_0$ has a standard deviation given by $\sqrt{\frac{p_0(1-p_0)}{M}}$ (Efron and Tibshirani, 1993). Since p_0 is unknown in practice, the corresponding upper bound $\frac{1}{2\sqrt{M}}$ is often used to determine the number of iterations required to achieve desired precision of the test.

3. Application of the Test

We use permutation testing in our work to assess the significance of the observed classification accuracy before we conclude that the results obtained in the cross-validation procedure are robust, or decide that more data are needed before we can trust the detected pattern or trend in the biological data.

In this section, we demonstrate the procedure in detail on simulated data and then on two different real data sets: a study of changes in the cortical thickness due to Alzheimer's disease using MRI scans for measurement and a discrimination between two types of leukemia based on DNA microarray data. These studies involve very different types of data, but both fall into the the "large p, small n" paradigm of few samples and thousands of features (West, 2003). We then report experimental results on a variety of real biological studies and offer practical guidelines in the application of the test.

3.1 Simulated data

The simulated data was generated as follows: 160 samples were generated from two normal distributions in $\mathbb{R}^2$ with means $(\pm 1, 0)$ and identity covariance with half the samples drawn from each distribution. Samples from group one were assigned a label $y = +1$ with probability p and $y = -1$ with probability $(1 - p)$. The opposite was done for group two. The probability $p \in [0, .5]$ denotes the noise level. We used linear discriminant analysis to train the classifier. The results are shown in Figures (1, 2, 3) for training error, leave-one-out error, and test error (the hold-out set is 20 samples per group), respectively. The black lines in the graphs plot the ecdfs of various errors for 5000 permutations of the data. As the noise parameter p is scanned over $\{.1, .2, .3, .4, .5\}$ the value of the unpermuted statistic, the red bar, shifts right. The value at which the red bar meets the black line determines the p-value (given in the caption for each figure). When the noise level increases, that is, the labels and features become more independent, the p-value increases as shown in these figures.

3.2 Two real data sets

For all subsequent experiments, we used linear Support Vector Machines (Vapnik, 1998) to train a classifier, and jackknifing (i.e., sampling without replacement) for cross-validation. The number of cross-validation iterations was $1,000$, and the number of permutation iterations was $10,000$.

The first example compares the thickness of the cortex in 50 patients diagnosed with dementia of the Alzheimer type and 50 normal controls of matched age. The cortical sheet was automatically segmented from each MRI scan, followed by a registration step that brought the surfaces into correspondence by mapping them onto a unit sphere while

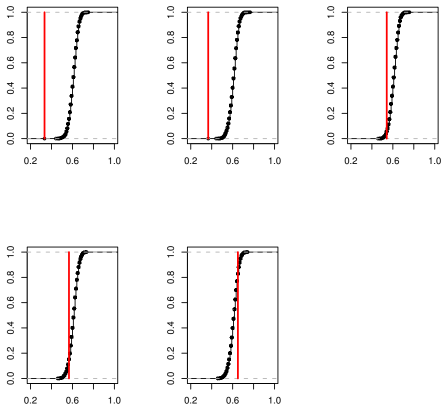


Figure 1: Training error: p-values = $\{0.0002, 0.0002, 0.0574, 0.1504, 0.8290\}$.

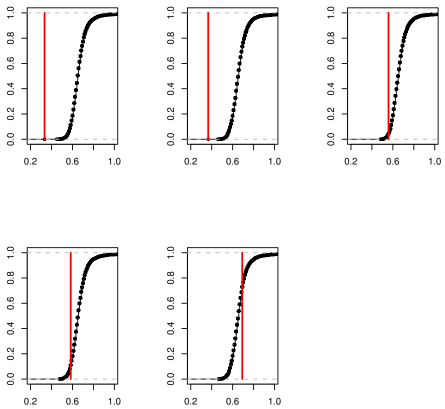


Figure 2: Leave-one-out error: p-values = $\{0.0002, 0.0002, 0.0430, 0.1096, 0.7298\}$.

minimizing distortions and then aligning the cortical folding patterns (Fischl et al., 1999, Fischl and Dale, 2000). The cortical thickness was densely sampled on a 1mm grid at corresponding locations for all subjects, resulting in over 300,000 thickness measurements. The measurements in neighboring locations are highly correlated, as both the pattern of thickness and the pattern of its change are smooth over the surface of the cortex, leading us to believe that learning the differences between the two groups might be possible with a reasonable number of examples. In summary, the dimensionality of the input space was $300,000$ and we had 50 samples from each class.

The statistic and its null distribution as a function of training set and hold-out set size is plotted in Figure (4). Every point in Figure (4a,b) is characterized by a corresponding training set size N and hold-out set size K, drawn from the original data set. It is not surprising that increasing the number of training examples improves the robustness of classification as exhibited by both the accuracy and the significance estimates. By examining

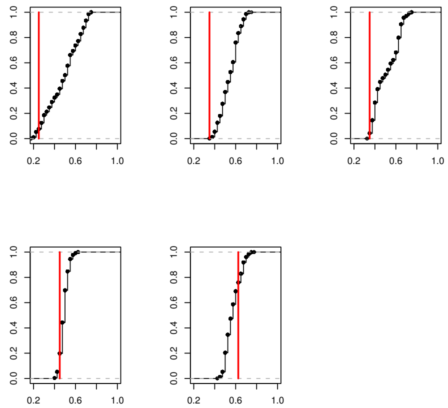
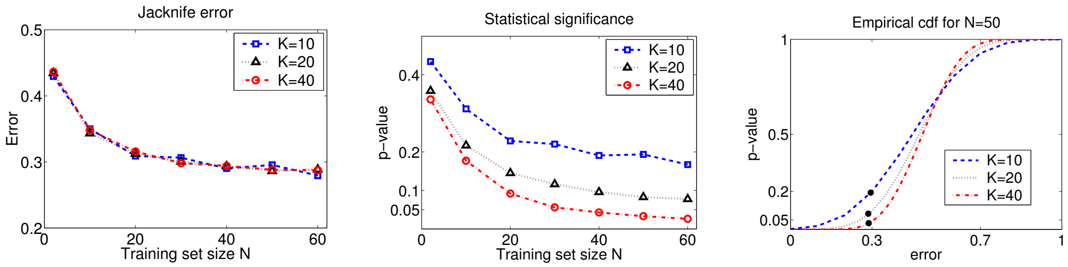


Figure 3: Test error: p-values $= \{0.0764, 0.0012, 0.0422, 0.1982, 0.7594\}$.



(a)

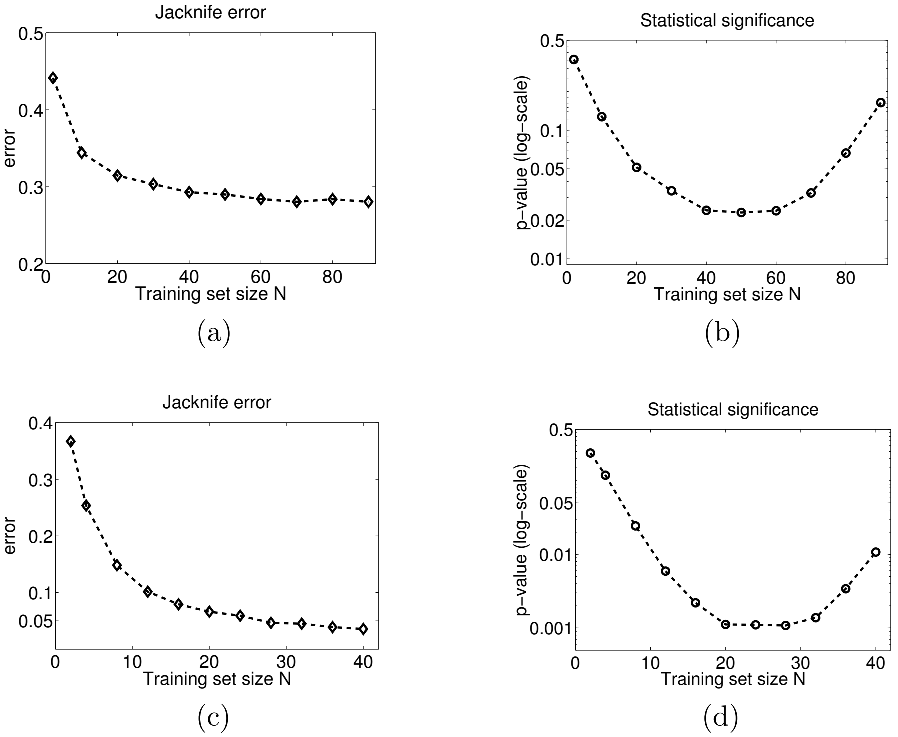
(b)

(c)

Figure 4: (a) Estimated test error (b) and statistical significance computed for different training set sizes N and test set sizes K, and (c) the empirical error distribution constructed for $N = 50$ and different test set sizes K in the cortical thickness study. Filled circles on the right graph indicate the classifier performance on the true labels ($K = 10$: $e = .30$, $p = .19$; $K = 20$: $e = .29$, $p = .08$; $K = 40$: $e = .29$, $p = .03$).

Figure (4a), we conclude that at approximately $N = 40$, the accuracy of the classification saturates at 71% ($e = .29$). After this point decreasing the number of hold-out samples does not significantly affect the estimated classification error, but does substantially decrease the statistical significance of the same error value. Figure (4c) illustrates this point for a particular training set size of $N = 50$.

Figure (5a,b) shows the estimated classification error and the corresponding p-values that were estimated using all of the examples left out in the training step in the hold-out set. While the error graph looks very similar to that in Figure (4), the behavior of significance estimates is quite different. The p-values originally decrease as the training set size increases, but after a certain point, they start growing. Two conflicting factors control p-value estimates as the number of training examples increases: improved accuracy of the



(a) (b)

(c) (d)

Figure 5: Estimated test error and statistical significance for different training set sizes N for (a,b) the cortical thickness study and (c,d) the leukemia morphology study. Unlike the experiments in Figure (4), all of the examples unused in training were used to test the classifier. The p-values are shown on a logarithmic scale.

classification, which causes the point of interest to slide to the left – and as a result, down – on the ecdf curve, and the decreasing number of test examples, which causes the ecdf curve to become more shallow.

The second example compares DNA microarray expression data from two types of leukemia acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) Golub et al. (1999), Slonim et al. (2000). The data set contains 48 samples of AML and 25 samples of ALL. Expression levels of $7,129$ genes and expressed sequence tags (ESTs) were measured via an oligonucleotide microarray for each sample, resulting in a dimensionality of $7,129$. Figure (5c,d) shows the results for this study. The cross-validation error reduces rapidly as we increase the number of training examples, dropping below 5% at $N = 26$ training examples. The p-values also decrease very quickly as we increase the number of training examples, achieving minimum of .001 at $N = 28$ training examples. Like the previous example, the most statistically significant result lies in the range of relatively slow error change.

As both examples demonstrate, using all of the data for training, for example a leave-one-out estimator, may lead to poor estimates of significance since an accurate variance estimate is not available. The additional data might not improve the testing accuracy in a substantial way and could be much more useful in obtaining more accurate estimates of the variance and the p-value.

3.3 Experimental results over several data sets

We report experimental results for nine different studies involving classification of biological data derived from imaging of two different types, as well as microarray measurements. Table 1 summarizes the number of features, which determines the input space dimensionality, the number of examples in each class for each study and the results of the statistical analysis. The studies are loosely sorted in the ascending strength of the results (decreasing error and p-values).

Imaging Data. In addition to the MRI study of cortical thickness described above (“MRI” in Table 1), we include the results of two fMRI (functional MRI) experiments that compare the patterns of brain activations in response to different visual stimuli in a single subject. We present the results of comparing activations in response to face images to those induced by house images, as these categories are believed to have special representation in the cortex. The feature extraction step was similar to that of the MRI study, treating the activation signal as the measurement to be measured over the cortical surface. The first experiment (“fMRI full”) used the entire cortical surface for feature extraction, while the second experiment (“fMRI reduced”) considered only the visually active region of the cortex. The mask for the visually active voxels was obtained using a separate visual task. The goal of using the mask was to test if removing irrelevant voxels from consideration improves the classification performance.

Microarray Data. In addition to the leukemia morphology study (“AML *vs.* ALL”), we include the results of five other expression datasets where either the morphology or the treatment outcome was predicted (Mukherjee et al., 2003). Three studies involved predicting treatment outcome: survival of lymphoma patients, survival of patients with brain cancer and predicting metastasis of breast cancers. Three other studies involved predicting morphological properties of the tissue: medulloblastomas (medullo) *vs.* glioblastomas (glio)[1] AML *vs.* ALL and tumor tissue *vs.* normal.

For each experiment, we analyzed the behavior of the cross-validation error and the statistical significance similarly to the detailed examples presented earlier. Table 1 summarizes the results for three important events: the first time the p-value plot crosses .05 threshold (thus achieving statistical significance at that level), the point of lowest p-value, and the lowest classification error. For the first two events, we report the number of training examples, the error and the p-value. The lowest error is typically achieved for the largest training set size and is shown here mainly for comparison with the other two error values reported. We observe that the error values corresponding to the lowest p-values are very close to the smallest errors reported, implying that the p-values bottom out in the region of a relatively slow change in the error estimates.

The first three studies in the table did not produce statistically significant results. In the first two studies, the rest of the indicators are extremely weak[2]. In the third study,

[1] Gliobastomas are tumors of glial cells in the brain while medulloblastomas are tumors of neural tissue.

[2] In (Mukherjee et al., 2003), a gene selection procedure led to greater accuracy and smaller p-values.

				p-value < .05			lowest p-value			
Experiment	features	pos	neg	$N/2$	e	p	$N/2$	e	p	e_{min}
Lymphoma outcome	7,129	32	26	–			15	.47	.47	.47
Brain cancer outcome	7,129	22	28	–			17	.46	.47	.45
Breast cancer outcome	24,624	44	34	–			15	.39	.15	.38
MRI	327,684	50	50	15	.30	.03	25	.29	.02	.28
Medullo *vs.* glioma	7,129	45	15	6	.12	.04	6	.12	.04	.11
AML *vs.* ALL	7,129	48	25	4	.15	.02	14	.05	.001	.04
fMRI full	303,865	15	15	6	.14	.02	8	.08	.01	.06
fMRI reduced	95,122	15	15	4	.08	.02	8	.009	.007	.003
Tumor *vs.* norm	16.063	190	90	10	.30	.008	45	.16	10^{-6}	.14

Table 1: Summary of the experimental data and results. The first four columns describe the study: the name of the experiment, the number of features and the number of examples in each class. Columns 5-7 report the number of training examples from each class $N/2$, the jackknife error e and the p-value p for the smallest training set size that achieves significance at $\alpha = .05$. Columns 8-10 report the results for the training set size that yields the smallest p-value. The last column contains the lowest error e observed in the experiment.

predicting whether a breast tumor will metastasize, the error stabilizes fairly early (training on 30 examples leads to 39% error, while the smallest error observed is 38%, obtained by training on 58 examples), but the p-values are too high. This leads us to believe that more data could help establish the significance of the result, similarly to the MRI study. Unfortunately, the error in these two studies is too high to be useful in a diagnostic application. The rest of the studies achieve relatively low errors and p-values. The last study in the table predicting cancerous tissue from normal tissue yields a highly significant result ($p < 10^{-6}$), with the error staying very stable for training on 90 examples to training on 170 examples. We also observe that the significance threshold of .05 is probably too high for these experiments, as the corresponding error values are significantly higher than the ones reported for the smallest p-value. A more stringent threshold of .01 would cause most experiments to produce a more realistic estimates of the cross-validation error attainable on the given data set.

3.4 Summary And Heuristics

In this section, we show how permutation testing in conjunction with cross-validation can be used to analyze the quality of classification results on scientific data. Here, we provide a list of practical lessons learned from the empirical studies that we hope will be useful to readers applying this methodology.

1. *Interpreting the p-value.* Two factors affect statistical significance: the separation between the classes and the amount of data we have to support it. We can achieve low p-values in a situation where the two classes are very far apart and we have a few

data points from each group, or when they are much closer, but we have substantially more data. The p-value by itself does not indicate which of these two situations is true. However, looking at both the p-value and the classifier accuracy gives us an indication of how easy it is to separate the classes.

2. *Size of the holdout set.* In our experience, small test sizes in cross-validation and permutation testing leads to noisy estimates of the classification accuracy and the significance. We typically limit the training set size to allow at least 10 test examples in each iteration of the resampling procedures.

3. *Size of the training set.* Since we are interested in a robust estimates of the test error, one should utilize sufficient number of training examples to be working in the region where the cross-validation error does not vary much. Cross-checking this with the region of lowest p-values provides another useful indication of the acceptable training set size. In our experiments with artificial data, the number of training examples at which the p-values stop decreasing dramatically remains almost constant as we add more data. This could mean that acquiring more experimental data will not change the "optimal" training set size substantially, only lower the resulting p-values.

4. *Performance on future samples.* As we pointed out in earlier sections, the permutation test does not provide a guarantee on how close the observed classification error is to the true expected error. Thus we might achieve statistical significance, but still have an inaccurate estimate of the expected error. However, asymptotically if the null distribution of the permutation procedure concentrates then the empirical error of the classifier will converge to the expected error independent of the underlying distribution. This surprising result is a result of the theory developed in the next section.

4. A Theoretical Motivation for the Permutation Procedure

The point of the permutation procedure is to examine if a classifier selected from a family of classifiers given a data set is predictive. By predictive we mean that the dependence relationship between y and $\mathbf{x}$ learned by the classifier is significantly different from the independent one. In the examples shown previously in the paper we used the training error as well as the leave-one-out or cross-validation error as the statistic used in the permutation procedure. Our theoretical motivation will focus on the training error. We will remark on generalizations to the leave-one-out error.

We first state conditions under which the permutation procedure concentrates, Section 4.1. In Section 4.2 we relate the concentration of the permutation procedure to p-values and comment on generalizing the proof to account for the leave-one-out error as the statistic used in the permutation procedure. In Section 4.3 we note that for classifiers, finite VC dimension is a necessary and sufficient condition for the concentration of the permutation procedure. We close with a discussion of the theoretical results.

4.1 Concentration of the permutation procedure

We are given a class of classifiers $\mathcal{C}$. Since there are only two classes, any classifier $c \in \mathcal{C}$ can be regarded as a subset of $\mathbb{R}^n$ to which class label $\{+1\}$ is assigned. Without loss of generality we will assume $\emptyset \in \mathcal{C}$. The unknown concept from which the data is sampled is defined as c_0. When there is no noise in this sampling process then there exists a c_0: $y = +1$ if $\mathbf{x} \in c_0$ and $y = -1$ otherwise. If the underlying distribution $\mu(\mathbf{x}, y)$ giving rise to the concept is noisy then we define c_0 as a Bayes optimal classifier for $\mu(\mathbf{x}, y)$ and c_0: $y = +1$ if $\mu(y = +1|\mathbf{x}) \geq \mu(y = -1|\mathbf{x})$ and $y = -1$ otherwise. This simple observation generalizes the results of Golland et al. (2005).

For a permutation $\boldsymbol{\pi}$ of the training data, the smallest training error on the permuted set is

$$\begin{aligned} e_l(\boldsymbol{\pi}) &= \min_{c\in\mathcal{C}} P_l(c \triangle c_0) \qquad (1) \\ &= \min_{c\in\mathcal{C}} \left[\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c, \mathbf{x}_i^{\pi} \notin c_0) + I(\mathbf{x}_i \notin c, \mathbf{x}_i^{\pi} \in c_0) \right], \end{aligned}$$

where $\mathbf{x}_i$ is the i-th sample and $\mathbf{x}_i^{\pi}$ is the i-th sample after permutation. For a fixed classifier $c \in \mathcal{C}$ the average error is

$$\begin{aligned} \mathbb{E} P_l(c \triangle c_0) &= \left(1 - \frac{1}{l}\right) [P(c)(1 - P(c_0)) + (1 - P(c))P(c_0)] + \\ &\quad \frac{1}{l}[P(c) + P(c_o) - 2P(c \cap c_0)], \end{aligned}$$

where the expectation is taken over the data $\mathbf{x}$ and permutations $\boldsymbol{\pi}$. As l gets large the average error is approximately $P(c)(1 - P(c_0)) + (1 - P(c))P(c_0)$ and since we can assume $P(c_0) \leq 1/2$ taking $c = \emptyset$ minimizes the average error at $P(c_0)$. We later refer to $P(c_0)$ as the random error because, a classifier such as $c = \emptyset$ is not informatively at all. Our goal is to show that under some complexity assumptions on class $\mathcal{C}$ the smallest training error $e_l(\boldsymbol{\pi})$ is close to the random error $P(c_0)$.

Minimizing (1) is equivalent to the following maximization problem

$$\max_{c\in\mathcal{C}} \left[\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c)(2I(\mathbf{x}_i^{\pi} \in c_0) - 1) \right],$$

since

$$e_l(\boldsymbol{\pi}) = P_l(\mathbf{x} \in c_0) - \max_{c\in\mathcal{C}} \left[\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c)(2I(\mathbf{x}_i^{\pi} \in c_0) - 1) \right],$$

and $P_l(\mathbf{x} \in c_0)$ is the empirical measure of the target concept. We would like to show that $e_l(\boldsymbol{\pi})$ is close to the random error $P(\mathbf{x} \in c_0)$ and give rates of convergence. We will do this by bounding the process

$$G_l(\boldsymbol{\pi}) = \sup_{c\in\mathcal{C}} \left[\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c)(2I(\mathbf{x}_i^{\pi} \in c_0) - 1) \right]$$

and using the fact that, by Chernoff's inequality, $P_l(\mathbf{x} \in c_0)$ is close to $P(\mathbf{x} \in c_0)$:

$$\mathbb{P}\left(P(\mathbf{x} \in c_0) - P_l(\mathbf{x} \in c_0) \leq \sqrt{\frac{2P(c_0)(1-P(c_0))t}{l}} \right) \geq 1 - e^{-t}. \tag{2}$$

Theorem 1 *If the concept class $\mathcal{C}$ has VC dimension V then with probability $1 - Ke^{-t/K}$*

$$G_l(\boldsymbol{\pi}) \leq K \min\left(\sqrt{\frac{V \log l}{l}}, \frac{V \log l}{l(1-2P(c_0))^2} \right) + \sqrt{\frac{Kt}{l}}.$$

Remark

The second quantity in the above bound comes from the application of Chernoff's inequality similar to (2) and, thus, has a "one dimensional nature" in a sense that it doesn't depend on the complexity (VC dimension) of class $\mathcal{C}$. An interesting property of this result is that if $P(c_0) < 1/2$ then first term that depends on the VC dimension V will be of order $\frac{V \log l}{l}$ which, ignoring the "one dimensional terms", gives the zero-error type rate of convergence of $e_l(\boldsymbol{\pi})$ to $P(\mathbf{x} \in c_0)$. Combining this theorem and equation (2) we can state that with probability $1 - Ke^{-t/K}$.

$$P(\mathbf{x} \in c_0) \leq P_l(\mathbf{x} \in c_0) + K \min\left(\sqrt{\frac{V \log l}{l}}, \frac{V \log l}{l(1-2P(c_0))^2} \right) + \sqrt{\frac{Kt}{l}}.$$

Throughout this paper K designates a constant the value of which can change over the equations.

In order to prove Theorem 1, we require several preliminary results. We first prove the following useful lemma.

Lemma 2 *It is possible to construct on the same probability space two i.i.d Bernoulli sequences $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ and $\varepsilon' = (\varepsilon'_1, \ldots, \varepsilon'_n)$ such that ε is independent of $\varepsilon'_1 + \ldots + \varepsilon'_n$ and $\sum_{i=1}^n |\varepsilon_i - \varepsilon'_i| = |\sum_{i=1}^n \varepsilon_i - \sum_{i=1}^n \varepsilon'_i|$.*

Proof

For $k = 0, \ldots, n$, let us consider the following probability space $\mathcal{E}_k$. Each element w of $\mathcal{E}_k$ consists of two coordinates $w = (\varepsilon, \pi)$. The first coordinate $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ has the marginal distribution of an i.i.d. Bernoulli sequence. The second coordinate π implements the following randomization. Given the first coordinate ε, consider a set $\mathcal{I}(\varepsilon) = \{i : \varepsilon_i = 1\}$ and denote its cardinality $m = \text{card}\{\mathcal{I}(\varepsilon)\}$. If $m \geq k$, then π picks a subset $\mathcal{I}(\pi, \varepsilon)$ of $\mathcal{I}(\varepsilon)$ with cardinality k uniformly, and if $m < k$, then π picks a subset $\mathcal{I}(\pi, \varepsilon)$ of the complement $I^c(\varepsilon)$ with cardinality $n - k$ also uniformly. On this probability space $\mathcal{E}_k$, we construct a sequence $\varepsilon' = \varepsilon'(\varepsilon, \pi)$ in the following way. If $k \leq m = \text{card}\{\mathcal{I}(\varepsilon)\}$ then we set $\varepsilon'_i = 1$ if $i \in \mathcal{I}(\pi, \varepsilon)$ and $\varepsilon'_i = -1$ otherwise. If $k > m = \text{card}\{\mathcal{I}(\varepsilon)\}$ then we set $\varepsilon'_i = -1$ if $i \in \mathcal{I}(\pi, \varepsilon)$ and $\varepsilon'_i = 1$ otherwise. Next, we consider a space $\mathcal{E} = \cup_{k \leq n} \mathcal{E}_k$ with probability measure $\mathbb{P}(\mathcal{A}) = \sum_{k=0}^n B(n, p, k) \mathbb{P}(\mathcal{A} \cap \mathcal{E}_k)$, where $B(n, p, k) = \binom{n}{k} p^k (1-p)^{n-k}$. On this probability space the sequence ε and ε' will satisfy the conditions of the lemma. First of all, $X = \varepsilon'_1 + \ldots + \varepsilon'_n$ has binomial distribution since by construction $\mathbb{P}(X = k) = \mathbb{P}(\mathcal{E}_k) = B(n, p, k)$. Also, by construction, the distribution of ε' is invariant under the permutation of coordinates. This, clearly, implies that ε' is i.i.d. Bernoulli. Also, obviously, ε is independent of $\varepsilon'_1 + \ldots + \varepsilon'_n$. Finally, by construction $\sum_{i=1}^n |\varepsilon_i - \varepsilon'_i| = |\sum_{i=1}^n \varepsilon_i - \sum_{i=1}^n \varepsilon'_i|$.□

Definition 3 *Let $u > 0$ and let $\mathcal{C}$ be a set of classifiers. Every finite set of concepts $c_1, ..., c_n$ with the property that for all $c \in \mathcal{C}$ there is a c_j such that*

$$\frac{1}{l} \sum_{i=1}^{l} |c_j(x_i) - c(x_i)|^2 \leq u$$

is called a u-cover with respect to $|| \cdot ||_{L_2(\mathbf{x}_l)}$. The covering number $\mathcal{N}(\mathcal{C}, u, \{\mathbf{x}_1, ..\mathbf{x}_l\})$ is the smallest number for which the above holds.

Definition 4 *The uniform metric entropy is $\log \mathcal{N}(\mathcal{C}, u)$ where $\mathcal{N}(\mathcal{C}, u)$ is the smallest integer for which*

$$\forall l, \quad \forall (\mathbf{x}_1, ..., \mathbf{x}_l), \quad \mathcal{N}(\mathcal{C}, u, \{\mathbf{x}_1, ..\mathbf{x}_l\}) \leq \mathcal{N}(\mathcal{C}, u).$$

Lemma 5 *The following holds with probability greater than $1 - Ke^{-t/K}$*

$$\begin{aligned} G_l(\boldsymbol{\pi}) \quad &\leq \quad \sup_r \left[K \frac{1}{\sqrt{l}} \int_0^{\sqrt{\mu_r}} \sqrt{\log \mathcal{N}(u, \mathcal{C})} du - \frac{\mu_r}{2} (1 - 2P(c_0)) + \sqrt{\frac{\mu_r (t + 2 \log(r+1))}{l}} \right] \\ &\quad + 2 \sqrt{\frac{2tP(c_0)(1 - P(c_0))}{l}}, \end{aligned}$$

where $\mu_r = 2^{-r}$ and $\log \mathcal{N}(\mathcal{C}, u)$ is the uniform metric entropy for the class $\mathcal{C}$.

Proof

The process

$$G_l(\boldsymbol{\pi}) = \sup_{c \in \mathcal{C}} \left[\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c)(2I(\mathbf{x}_i^{\pi} \in c_0) - 1) \right].$$

can be rewritten as

$$G_l(\boldsymbol{\pi}) = \sup_{c \in \mathcal{C}} \left[\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c) \varepsilon_i \right],$$

where $\varepsilon_i = 2I(\mathbf{x}_i^{\pi} \in c_0) - 1 = \pm 1$ are Bernoulli random variables with $P(\varepsilon_i = 1) = P(c_0)$. Due to permutations the random variables (ε_i) depend on $(\mathbf{x}_i)$ only through the cardinality of $\{\mathbf{x}_i \in c_0\}$. By lemma 2 we can construct a random Bernoulli sequence (ε_i') that is independent of $\mathbf{x}$ and for which

$$G_l(\boldsymbol{\pi}) \leq \sup_{c \in \mathcal{C}} \left[\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c) \varepsilon_i' \right] + \left| \frac{1}{l} \sum_{i=1}^{l} \varepsilon_i - \frac{1}{l} \sum_{i=1}^{l} \varepsilon_i' \right|.$$

We first control the second term

$$\left| \frac{1}{l} \sum_{i=1}^{l} \varepsilon_i - \frac{1}{l} \sum_{i=1}^{l} \varepsilon_i' \right| \leq \left| \frac{1}{l} \sum_{i=1}^{l} \varepsilon_i' - (2P(c_0) - 1) \right| + \left| \frac{1}{l} \sum_{i=1}^{l} \varepsilon_i - (2P(c_0) - 1) \right|,$$

then using Chernoff's inequality twice we get with probability $1 - 2e^{-t}$

$$\left| \frac{1}{l} \sum_{i=1}^{l} \varepsilon_i - \frac{1}{l} \sum_{i=1}^{l} \varepsilon_i' \right| \leq 2 \sqrt{\frac{2tP(c_0)(1 - P(c_0))}{l}}.$$

We block concepts in $\mathcal{C}$ into levels

$$\mathcal{C}_r = \left\{ c \in \mathcal{C} : \frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c) \in (2^{-r-1}, 2^{-r}] \right\}$$

and denote $\mu_r = 2^{-r}$. We define the processes

$$R(r) = \sup_{c \in \mathcal{C}_r} \left[\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c) \varepsilon'_i \right],$$

and obtain

$$\sup_{c \in \mathcal{C}} \left[\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c) \varepsilon'_i \right] \leq \sup_r R(r).$$

By Talagrand's convex hull inequality on the two point space (Talagrand, 1995), we have for each level r

$$\mathbb{P}_{\varepsilon'} \left(R(r) \leq \mathbb{E}_{\varepsilon'} \sup_{c \in \mathcal{C}_r} \left[\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c) \varepsilon'_i \right] + \sqrt{\frac{\mu_r t}{l}} \right) \geq 1 - K e^{-t/K}.$$

Note that for this inequality to hold, the random variables (ε') need only be independent, they do not need to be symmetric. This bound is conditioned on a given $\{\mathbf{x}_i\}_{i=1}^{l}$ and by taking the expectation w.r.t. $\{\mathbf{x}_i\}$ we get,

$$\mathbb{P} \left(R(r) \leq \mathbb{E}_{\varepsilon'} \sup_{c \in \mathcal{C}_r} \left[\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c) \varepsilon'_i \right] + \sqrt{\frac{\mu_r t}{l}} \right) \geq 1 - K e^{-t/K}.$$

If, for each r, we set $t \to t + 2 \log(r+1)$, we can write

$$\mathbb{P} \left(\forall r \;\; R(r) \leq \mathbb{E}_{\varepsilon'} \sup_{c \in \mathcal{C}_r} \left[\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c) \varepsilon'_i \right] + \sqrt{\frac{\mu_r (t + 2\log(r+1))}{l}} \right)$$
$$\geq 1 - \sum_{r=0}^{\infty} \frac{1}{(r+1)^2} e^{-t/4} \geq 1 - 2e^{-t/4}.$$

Using standard symmetrization techniques we add and subtract an independent sequence ε''_i such that $\mathbb{E}\varepsilon''_i = \mathbb{E}\varepsilon'_i = (2P(c_0) - 1)$:

$$\begin{aligned}
&\mathbb{E}_{\varepsilon'} \sup_{c \in \mathcal{C}_r} \left[\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c) \varepsilon'_i \right] \\
&\leq \mathbb{E}_{\varepsilon'} \sup_{c \in \mathcal{C}_r} \left[\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c) \varepsilon'_i - \frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c) \mathbb{E}\varepsilon''_i + \frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c)(2P(c_0) - 1) \right] \\
&\leq \mathbb{E}_{\varepsilon' \varepsilon''} \sup_{c \in \mathcal{C}_r} \left[\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c)(\varepsilon' - \varepsilon'') \right] - (1 - 2P(c_0)) \inf_{c \in \mathcal{C}_r} \left(\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c) \right) \\
&\leq 2\mathbb{E}_{\eta_i} \sup_{c \in \mathcal{C}_r} \left[\frac{1}{l} \sum_{i=1}^{l} I(\mathbf{x}_i \in c) \eta_i \right] - \frac{\mu_r (1 - 2P(c_0))}{2},
\end{aligned}$$

where $\eta_i = (\varepsilon'_i - \varepsilon''_i)/2$ takes values $\{-1, 0, 1\}$ with probability $P(\eta_i = 1) = P(\eta_i = -1)$. One can easily check that the random variables η_i are subgaussian, i.e.

$$\mathbb{P}\left(\sum_{i=1}^{l} \eta_i a_i > t\right) \leq e^{-\frac{t^2}{2\sum_{i=1}^{l} a_i^2}},$$

which is the only prerequisite for the chaining method. Thus, one can write Dudley's entropy integral bound, (van der Vaart and Wellner, 1996)

$$\mathbb{E}_{\eta_i} \sup_{c \in \mathcal{C}_r} \left[\frac{1}{l}\sum_{i=1}^{l} I(\mathbf{x}_i \in c)\eta_i\right] \leq K\frac{1}{\sqrt{l}}\int_0^{\sqrt{\mu_r}} \sqrt{\log \mathcal{N}(u, \mathcal{C})}du.$$

We finally get

$$\mathbb{P}\left(\forall r \;\; R(r) \leq K\frac{1}{\sqrt{l}}\int_0^{\sqrt{\mu_r}} \sqrt{\log \mathcal{N}(u, \mathcal{C})}du + \sqrt{\frac{\mu_r(t + 2\log(r+1))}{l}} - \frac{\mu_r(1 - 2P(c_0))}{2}\right)$$
$$\geq 1 - 2e^{-t/4}.$$

This completes the proof of Lemma 5. □

Proof of Theorem 1

For a class with VC dimension V, it is well known that (van der Vaart and Wellner, 1996)

$$\frac{1}{\sqrt{l}}\int_0^{\sqrt{\mu_r}} \sqrt{\log \mathcal{N}(u, \mathcal{C})}du \leq K\sqrt{\frac{V\mu_r \log\frac{2}{\mu_r}}{l}}.$$

Since without loss of generality we only need to consider $\mu_r > 1/l$, it remains to apply lemma 5 and notice that

$$\sup_r \left[K\sqrt{\frac{V\mu_r \log l}{l}} - \frac{\mu_r}{2}\left(1 - 2P(c_0)\right)\right] \leq K \min\left(\sqrt{\frac{V\log l}{l}}, \frac{V\log l}{l(1 - 2P(c_0))^2}\right).$$

All other terms that do not depend on the VC dimension V can be combined to give $\sqrt{Kt/l}$.□

4.2 Relating p-values to concentration and the leave-one-out error as the permutation statistic

The result of the previous section states that for VC classes the training error concentrates around $q = \min\{P(y = 1), P(y = -1)\}$.

We can relate this concentration result to the p-value computed by the permutation procedure. The purpose of this is to give a theoretical justification for the empirical procedure outlined in section 2. We do not recommend replacing the empirical procedure with the theoretical bound in practical applications. We assume the statistic used in the permutation procedure is the training error

$$\tau = \frac{1}{l}\sum_{i=1}^{l} I(f_S(x_i) \neq y_i),$$

and f_S is the function obtained by the classification algorithm given the data set S. If we were given the distribution of the training errors over random draws and random label permutations we would have the distribution under the null hypothesis $P_{\text{null}}(\xi)$ and the p-value of the statistic τ would simply be $P_{\text{null}}(\xi \leq \tau)$. In the empirical procedure outlined in section 2 we used an empirical estimate $\hat{P}(\xi)$ to compute the p-value.

The results of section 4.1 give us a bound of the deviation of the training error of the permuted data from $P(c_0)$ under the null hypothesis, namely,

$$\mathbb{P}\left(|e_l(\boldsymbol{\pi}) - P(c_0)| \geq \varepsilon\right) \leq Ke^{-\varepsilon^2 \mathcal{O}(l)}, \tag{3}$$

where $\mathcal{O}(l)$ ignores $\log l$ terms. We assume that we know $P(c_0)$, otherwise it can be accurately estimated by a frequency count of $y = \pm 1$. We can bound the p-value by setting $|t - P(c_0)| = \varepsilon$ and computing $Ke^{-\varepsilon^2 \mathcal{O}(l)}$.

The difference between the p-value computed using the inequality (3) and that outlined in section 2 is the later is a one-sided test and is based upon empirical approximations rather than bounds. A one-sided test can be derived from the results in section 4.1 in a similar fashion as the two-sided test.

We can also use the leave-one-out error as the statistic used in the permutation procedure

$$\tau = \frac{1}{l} \sum_{i=1}^{l} I(f_{S^i}(x_i) \neq y_i),$$

where S^i is the data set with the ith sample removed and f_{S^i} is the function obtained by the classification algorithm given the data set S^i. In this case, for certain algorithms we can make the same theoretical arguments for the leave-one-out estimator as we did for the training error since with high probability the training error is close to the leave-one-out error.

Proposition 6 *If independent of measure $\mu(\mathbf{x}, y)$ with probability greater than $1 - Ke^{-t/K}$*

$$\left| \frac{1}{l} \sum_{i=1}^{l} I(f_{S^i}(x_i) \neq y_i) - \frac{1}{l} \sum_{i=1}^{l} I(f_S(x_i) \neq y_i) \right| \leq K \sqrt{\frac{t \log l}{l}},$$

then the leave-one-out estimate on the permuted data will concentrate around $P(c_0)$ with the same rates of convergence as the training error.

The proof is obvious in that if the deviation between the leave-one-out estimator and the training error is of the same order as that of the deviation between the training error and $P(c_0)$ and both hold with exponential probability then we can simply replace the leave-one-out error with the training error and maintain the same rate of convergence.

The condition in Proposition 6 holds for empirical risk minimization on a VC class in the realizable setting (Kutin and Niyogi, 2002) and for Tikhonov regularization with Lipschitz loss functions (Bousquet and Elisseeff, 2002).

4.3 A necessary and sufficient condition for the concentration of the permutation procedure

In this section we note that for a class of classifiers finite VC dimension is a necessary and sufficient condition for the concentration of the training error on the permuted data.

The proof of Lemma 5 makes no assumptions of the class $\mathcal{C}$ except that it is a class of indicator functions and the bounds used in the proof are tight in that the equality can be achieved under certain distributions. A step in the proof of the lemma involved upper-bounding the Rademacher process by Dudley's entropy integral. Here assumptions on the class $\mathcal{C}$ are introduced to control the Rademacher process in the inequality in Lemma 5. For finite VC dimension the process can be upper bounded by $\mathcal{O}\left(\sqrt{\frac{1}{l}}\right)$ using Dudley's entropy integral which proves sufficiency. The Rademacher process can also be lower bounded by a function of the metric entropy via Sudakov minorization (van der Vaart and Wellner, 1996). If $\mathcal{C}$ has infinite VC dimension this lower bound is a constant and the process does not concentrate which proves necessity.

4.4 Comments on the Theoretical Results

The theoretical results in this section were to give an analysis and motivation for the permutation tests. The bounds derived are not meant to replace the empirical permutation procedure. The bounds derived are from a worst-case analysis and are similar to VC-style generalization bounds in appearance and magnitude. For this reason these bounds would only be practically applicable for sample sizes far beyond the limited range of samples for which the empirical procedure has proven success.

Note that Theorem 1 is a uniform central limit theorem for permutation procedures and generalizes the central limit theorem for permutation of sequences derived by Hoeffding (1951) in the same spirit as the results of Vapnik and Červonenkis (1971) generalized the central limit theorem for sums of independent random variables derived by Hoeffding (1963).

5. Conclusion And Open Problems

This paper describes and explores an approach to estimating statistical significance of a classifier given a small sample size based on permutation testing. The following is the list of open problems related to this methodology:

1. *Size of the training/test set.* We provide a heuristic to select the size of the training and the holdout sets. A more rigorous formulation of this problem might suggest a more principled methodology for setting the training set size. This problem is clearly an example of the ubiquitous bias-variance trade-off dilemma.

2. *Leave-one-out error and training error.* In the theoretical motivation, we relate the leave-one-out error to the training error for certain algorithms. The result would be stronger if proposition 6 held for VC classes in the nonrealizable setting.

3. *Feature selection.* Both in neuroimaging studies and in DNA microarray analysis, finding the features which most accurately classify the data is very important. Permutation procedures similar to the one described in this paper have been used to

address this problem (Golub et al., 1999, Slonim et al., 2000, Nichols and Holmes, 2001). The analysis of permutation procedures for selecting discriminative features seems to be more difficult than the analysis of the permutation procedure for classification. It would be very interesting to extend the type of analysis here to the feature selection problem.

4. *Multi-class classification.* Extending the methodology and theoretical motivation for the multi-class problem has not been done.

To conclude, we hope other researchers in the community will find the technique useful in assessing statistical significance of observed results when the data are high dimensional and are not necessarily generated by a known distribution.

Acknowledgments

Authors are listed in alphabetical order. The authors would like thank Pablo Tamayo, Vladimir Koltchinskii, Jill Mesirov, Todd Golub and Bruce Fischl for useful discussions. Sayan Mukherjee was supported by a SLOAN/DOE grant. Polina Golland was supported in part by NSF IIS 9610249 grant and Athinoula A. Martinos Center for Biomedical Imaging collaborative research grant. Dmitry Panchenko was partially supported by AT&T Buchsbaum grant.

The authors would like to acknowledge Dr. M. Spiridon and Dr. N. Kanwisher for providing the fMRI data, Dr. R. Buckner for providing the cortical thickness data and Dr. D. Greve and Dr. B. Fischl for help with registration and feature extraction in the MRI experiments discussed in this paper. Dr. Kanwisher would like to acknowledge EY 13455 and MH 59150 grants. Dr. Buckner would like to acknowledge the assistance of the Washington University ADRC, James S McDonnell Foundation, the Alzheimer's Association, and NIA grants AG05682 and AG03991. Dr. B. Fischl would like to acknowledge NIH R01 RR16594-01A1 grant. The Human Brain Project/Neuroinformatics research is funded jointly by the NINDS, the NIMH and the NCI (R01-NS39581). Further support was provided by the NCRR (P41-RR14075 and R01-RR13609).

Support was also provided by grants from: Office of Naval Research (DARPA) Contract No. N00014-00-1-0907, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation (ITR/IM) Contract No. IIS-0085836, National Science Foundation (ITR/SYS) Contract No. IIS-0112991, National Science Foundation (ITR) Contract No. IIS-0209289, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218693, and National Science Foundation-NIH (CRCNS) Contract No. EIA-0218506.

Additional support was provided by: AT&T, Central Research Institute of Electric Power Industry, Center for e-Business (MIT), DaimlerChrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda R&D Co., Ltd., ITRI, Komatsu Ltd., The Eugene McDermott Foundation, Merrill-Lynch, Mitsubishi Corporation, NEC Fund, Nippon Telegraph & Telephone, Oxygen, Siemens Corporate Research, Inc., Sony MOU, Sumitomo Metal Industries, Toyota Motor Corporation, and WatchVision Co., Ltd.

References

H.J. Arnold. Permutation support for multivariate techniques. *Biometrika*, 51:65–70, 1964.

C.B. Bell and K.A. Doksum. Some new distribution free statistics. *Annal Math Statist*, 36: 455–467, 1965.

A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *Journal Computational Biology*, 7:559–584, 2000.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal Machine Learning Research*, 2:499–526, 2002.

G.E.P. Box and S.L. Anderson. Permutation theory in the development of robust criteria and the study of departures from assumptions. *J Roy Statist Soc B*, 17:1–34, 1955.

E.T. Bullmore, J. Suckling, S. Overmeyer, S. Robe-Hesketh, E. Taylor, and M. J. Brammer. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural mr images of the brain. *IEEE Transactions on Medical Imaging*, 18 (1):32–42, 1999.

M. Dwass. Modified randomization tests for non-parametric hypotheses. *Annal Math Statist*, 28:181–187, 1957.

B. Efron. *The Jackknife, The Bootstrap, and Other Resampling Plans.* SIAM, Philadelphia, PA, 1982.

B. Efron and R. Tibshirani. *An introduction to the bootstrap.* Chapman & Hall Ltd, 1993.

B. Fischl and A.M. Dale. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *PNAS*, 26:11050–11055, 2000.

B. Fischl, M.I. Sereno, R.B.H. Tootell, and A.M. Dale. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, 8:262–284, 1999.

R.A. Fisher. Coefficient of racial likeness and the future of craniometry. *J Roy Anthrop Soc*, 66:57–63, 1936.

P. Golland and B. Fischl. Permutation tests for classification: Towards statistical significance in image-based studies. In *IPMI'2003: The 18th International Conference on Information Processing and Medical Imaging*, volume LNCS 2732, pages 330–341, 2003.

P. Golland, F. Liang, S. Mukherjee, and D. Panchenko. Permutation tests for classification. In P Auer and R Meir, editors, *Conference on Learning Theory*, pages 501–515. Spinger-Verlag, 2005.

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

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

W. Hoeffding. Combinatorial central limit theorem. *Annal Math Statist*, 22:558–566, 1951.

W. Hoeffding. Probability Inequalities for Sums of Bounded random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

T. Hsing, S. Attoor, and E. Dougherty. Relation between permutation-test p values and classifier error estimates. *Machine Learning*, 52:11–30, 2003.

S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. Technical report TR-2002-03, University of Chicago, 2002.

E.L. Lehman and C. Stein. On the theory of some nonparametric hypothesis. *Annal Math Statist*, 20:28–45, 1949.

S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T.R. Golub, and J.P. Mesirov. Estimating dataset size requirements for classifying dna microarray data. *Journal Computational Biology*, 10(2):119–142, 2003.

T.E. Nichols and A.P. Holmes. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15:1–25, 2001.

E.J.G. Pitman. Significance tests which may be applied to samples from any population. *Roy Statist Soc Suppl*, 4:119–130,225–232, 1937.

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

D. Slonim, P. Tamayo, J.P. Mesirov, T.R. Golub, and E. Lander. Class prediction and discovery using gene expression data. In *Proceedings of the Fourth Annual Conference on Computational Molecular Biology (RECOMB)*, pages 263–272, 2000.

M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'I.H.E.S.*, 81:73–205, 1995.

O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein, and R. B. Altman. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18(11):1454–1461, 2002.

A. van der Vaart and J. Wellner. *Weak convergence and Empirical Processes With Applications to Statistics.* Springer-Verlag, 1996.

V.N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, 1998.

V.N. Vapnik and A. Ya. Červonenkis. On the uniform convergence of relative frequencies of events to their probabilites. *Theory Probab Appl*, 16:264–280, 1971.

W.R. Van Zwet W. Albers, P.J. Bickel. Asymptotic expansions for the power of distribution-free tests in the one-sample problem. *Annal Statist*, 4:108–156, 1976.

A. Wald and J. Wolfowitz. Statistical tests based upon permutations of observations. *Annal Math Statist*, 15:358–372, 1944.

M. West. Bayesian factor regression models in the "large p, small n" paradigm. In J.M. Bernardo et al., editor, *Bayesian Statistics 7*, pages 723–732. Oxford, 2003.