SLAC-PUB-10325
January 2004

# On Multivariate Goodness–of–Fit and Two–Sample Testing

Jerome H. Friedman
Department of Statistics and
Stanford Linear Accelerator Center,
Stanford University, Stanford, CA 94305
(jhf@stanford.edu)

It is shown how classification learning machines can be used to do multivariate goodness–of–fit and two–sample testing.

## 1. Introduction

In the goodness–of–fit testing problem one is given a data set of $N$ measured observations $\{\mathbf{x}_i\}_{i=1}^N$ each of which is presumed to be randomly drawn independently from some probability distribution with density $p(\mathbf{x})$. The goal is to test the hypothesis that $p(\mathbf{x}) = p_0(\mathbf{x})$, where $p_0(\mathbf{x})$ is some specified reference probability density. Ideally, the test should have power against all alternatives. That is as the sample size $N$ becomes arbitrarily large, $N \to \infty$, the test will reject the hypothesis for all distributions $p \neq p_0$ at any non zero significance $\alpha$ level.

A related problem is two–sample testing. Here one has two data sets: $\{\mathbf{x}_i\}_{i=1}^N$ drawn from $p(\mathbf{x})$, and $\{\mathbf{z}_i\}_{i=1}^M$ drawn from $q(\mathbf{z})$. The goal is to test the hypothesis that $p = q$, again with power against all alternatives; as $N \to \infty$ and $M \to \infty$ the test will always reject when $p \neq q$. Two–sample testing can be used to do goodness–of–fit testing. A random sample $\{\mathbf{z}_i\}_{i=1}^M$ is drawn from the reference distribution $q = p_0$ and then a two–sample test is performed on $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{z}_i\}_{i=1}^M$.

In univariate (one–dimensional) problems each observation $\mathbf{x}_i$ (and $\mathbf{z}_i$) consists of only a single measurement. In this case there are a wide variety of useful and powerful goodness–of–fit and two–sample testing procedures. Some of these can be extended to two or perhaps three dimensions if the sample size is large enough. However, when each observation consists of many measured attributes $\mathbf{x}_i = \{x_{i1}, x_{i2}, \cdots, x_{in}\}$ (and $\mathbf{z}_i = \{z_{i1}, z_{i2}, \cdots, z_{in}\}$) , for large $n$, these tests rapidly loose power because all finite samples are sparse in high dimensional settings owing to the "curse–of–dimensionality" (Bellman 1961).

## 2. Machine learning classification

The purpose of a learning machine is to predict (estimate) the unknown value of an attribute $y$ given a set of jointly measured values $\mathbf{x}$ of other attributes. The quantity $y$ is called the "output" or "response" variable, and $\mathbf{x} = \{x_1, \cdots, x_n\}$ are referred to as the "input" or "predictor" variables. In the binary classification problem, the response variable realizes two values, i.e. $y \in \{-1, 1\}$, respectively labeling the observations from each of two classes. The goal is to produce a model $F(\mathbf{x})$ that represents a score reflecting confidence that $y = 1$, given a set of joint values for the predictor variables $\mathbf{x}$. This score can then be used in a decision rule to obtain a corresponding prediction

$$\hat{y}(\mathbf{x}) = \begin{cases} 1 & \text{if } F(\mathbf{x}) > t^* \\ -1 & \text{otherwise.} \end{cases}$$

Here $t^*$ is a threshold whose value is chosen to minimize the error rate.

There are a variety of ways one can go about trying to find a good predicting function $F(\mathbf{x})$. In predictive or machine learning a "training" data base $\{y_i, \mathbf{x}_i\}_{i=1}^N$ of $N$ previously solved cases is used for which the values of all variables (response and predictors) have been jointly measured. A "learning machine" is applied to these data in order to extract (estimate) a good scoring function $F(\mathbf{x})$. There are a great many commonly used learning machines. These include linear/logistic regression, neural networks, kernel methods, decision trees, support vector machines, etc. Many are intended for use with large numbers of predictor variables. For descriptions of a wide variety of such learning procedures see Hastie, Tibshirani and Friedman 2001.

## 3. Two–sample testing

Binary classification procedures can be used for two–sample testing. A predictor variable training data set is created by pooling the two samples

$$\{\mathbf{u}_i\}_{i=1}^{N+M} = \{\mathbf{x}_i\}_{i=1}^N \cup \{\mathbf{z}_i\}_{i=1}^M.$$

Those observations that originated from the first sample ($1 \leq i \leq N$) are assigned a response value $y_i = 1$ while those from the second sample ($N + 1 \leq i \leq N + M$) are assigned $y_i = -1$. A binary classification learning machine is applied to this training data to

Work supported in part by the Department of Energy contract DE-AC03-76SF00515.

produce a scoring function $F(\mathbf{u})$. This is then used to score each of the observations $\{s_i = F(\mathbf{u}_i)\}_{i=1}^{N+M}$.

Consider the two sets of score values $S_+ = \{s_i\}_{i=1}^{N}$ and $S_- = \{s_i\}_{i=N+1}^{N+M}$. These are the scores respectively assigned by the learning machine $F(\mathbf{u})$ to the first sample $\{\mathbf{x}_i\}_{i=1}^{N}$ and the second sample $\{\mathbf{z}_i\}_{i=1}^{M}$. Each of these sets of numbers $S_\pm$ can be viewed as a random sample from respective probability distributions with densities $p_+(s)$ and $p_-(s)$. Consider a *univariate* two–sample test $T$ for the equality of these densities $p_+(s) \overset{\cdot}{=} p_-(s)$. Let $\hat{t}$ represent the value of the corresponding test statistic

$$\hat{t} = T(\{s_i\}_{i=1}^{N}, \{s_i\}_{i=N+1}^{N+M}). \qquad (1)$$

Examples of commonly applied univariate two–sample tests include chi–squared, Kolmogorov–Smirnov, Mann–Whitney, t–test, etc. This quantity (1) is taken to be the statistic for the *multivariate* two–sample test for the equality of the distributions of $\{\mathbf{x}_i\}_{i=1}^{N}$ and $\{\mathbf{z}_i\}_{i=1}^{M}$ $(p = q)$.

## 3.1. Null distribution

In order to test the "null" hypothesis $p = q$ it is necessary to know the distribution $H_0(t)$ of (1) when the hypothesis is in fact true. One rejects the null hypothesis at significance level $\alpha$ if the value $\hat{t}$ actually observed is greater than or equal to the $1-\alpha$ quantile of $H_0(t)$, assuming smaller values of $t$ represent greater likelihood of $p = q$. For commonly applied *univariate* two–sample tests the corresponding null distributions are known and have been tabulated. These distributions are valid for the multivariate application provided that separate independent data sets are respectively used for training the learning machine and evaluating the scores (1).

When the same data is used for both training and subsequent scoring, these univariate null distributions are not valid. In this case one can perform a permutation ("Fisher's exact") test. Let $\{j(i)\}_{i=1}^{N+M}$ represent a random permutation of the integers $\{1, 2, \cdot\cdot\cdot, N+M\}$. One constructs a data set $\{y_{j(i)}, \mathbf{u}_i\}_{i=1}^{N+M}$ in which the actual response values $\{y_i\}_{i=1}^{N+M}$ are randomly permuted among the predictors $\{\mathbf{u}_i\}_{i=1}^{N+M}$. These data are then used to train the learning machine, score the observations, and compute the test statistic (1). This random permutation process is repeated many (say $P$) times producing a set of test statistic values $\{\hat{t}_l\}_{l=1}^{P}$. One can then reject the null hypothesis with significance level $\alpha$ if the value $\hat{t}$ computed form the original data $\{y_i, \mathbf{u}_i\}_{i=1}^{N+M}$ is greater than or equal to the $1-\alpha$ quantile of $\{\hat{t}_l\}_{l=1}^{P}$. This is valid for any number of random permutations $P$ , but power increases with increasing $P$, reaching a diminishing return for large enough values.

## 4. Goodness–of–fit testing

As noted in Section 1, two–sample testing can be used to perform goodness–of–fit tests. One draws an artificial ("Monte Carlo") sample $\{\mathbf{z}_i\}_{i=1}^{M}$ from the reference distribution $q = p_0$ and tests the hypothesis $p = q$, where $p(\mathbf{x})$ is the unknown probability density of the data sample $\{\mathbf{x}_i\}_{i=1}^{N}$. The test is valid for any size $M$ of the Monte Carlo sample, but power increases with increasing $M$, reaching a diminishing return for $M >> N$.

In two–sample testing a null distribution $H_0(t)$ is constructed by repeated random permutations of the responses $\{y_i\}_{i=1}^{N+M}$ over the predictors $\{\mathbf{u}_i\}_{i=1}^{N+M}$. This is valid for the goodness–of–fit application as well. However in the goodness–of–fit context there is an alternative method for creating a null distribution that can increase power at the expense of increased computation. One repeatedly draws many (say $P$) independent Monte Carlo samples of size $M$ from the reference distribution. Each of these Monte Carlo samples $\{\mathbf{z}_i^{(l)}\}_{i=1}^{M}$ is used, along with the actual data $\{\mathbf{x}_i\}_{i=1}^{N}$, for training the learning machine and subsequent scoring to produce a test statistic value $\hat{t}_l$ from (1). This produces a set of values $\{\hat{t}_l\}_{l=1}^{P}$ that can be used as a null distribution to test the hypothesis $p = p_0$ in the usual manner.

The permutation procedure used with two–sample testing to construct a null distribution conditions on the observed data values $\{\mathbf{x}_i\}_{i=1}^{N}$ and $\{\mathbf{z}_i\}_{i=1}^{M}$; only information from the labels $\{y_i = \pm 1\}_{i=1}^{N+M}$ that identify the sample from which each observation originated is used. When used for goodness–of–fit testing this conditions on the values of the single Monte Carlo sample $\{\mathbf{z}_i\}_{i=1}^{M}$ drawn from the reference distribution $q = p_0$. Goodness–of–fit testing using repeated Monte Carlo samples as described above does not involve such conditioning and thereby uses information from the values of $\{\mathbf{z}_i\}_{i=1}^{M}$, as well as the labels $\{y_i = \pm 1\}_{i=1}^{N+M}$, in testing the null hypothesis. Using this additional information has the potential for increased power at the expense of having to generate many Monte Carlo samples, instead of just one.

## 5. Discussion

As noted in the introduction, a desirable property of goodness–of–fit and two–sample tests is power against all alternatives to the null hypothesis. This will be the case provided that the chosen leaning machine is universal. That is, as the number of observations used to train it grows arbitrarily large, $N, M \to \infty$, an "optimal" scoring function $F(\mathbf{u})$ is produced that is a strictly monotone function of $\Pr(y = +1 \mid \mathbf{u})$. Some examples of universal learning machines are decision trees, neural networks, and support vector machines based on appropriate kernels. Additionally, a consistent univariate test statistic must be used in (1). That is, as $N, M \to \infty$ they will always reject the null hypothesis when $p_+(s) \neq p_-(s)$.

This notion of power against all alternatives applies in the asymptotic limit of infinite data. It has at best limited meaning in actual finite data applications. With finite data, tests based on different types of (even universal) learning machines will have differential power against different alternative distributions $p \neq p_0$ or $p \neq q$. Depending upon the actual data distribution(s) $p(\mathbf{x})$ (and $q(\mathbf{z})$) encountered in a particular application, some learning machines will have more power than others. Thus, the power of these tests can be highly sensitive to the learning machine employed. Particular choices depend on the types of potential differences between the distributions that are deemed most important to detect. For example, if the distributions tend to be different on a large fraction of the variables, near–neighbor or kernel methods will provide high power. On the other hand if they tend to differ on only a relatively small number of variables, decision trees will provide greater sensitivity.

Some multivariate two–sample tests based on near–neighbors have an advantage in that the permutation null distribution can be computed analytically. For these tests repeated learning machine training and scoring based on randomly generated permutations is not required (see Friedman and Rafsky 1979 and 1983).

In contrast to the dependence on the particular learning machine employed, the multivariate procedures described here are not likely to be very sensitive to the choice of a univariate test statistic (1).

It should be noted that as a data analytic procedure hypothesis testing extracts very little information from the data. This summary information can be encoded in a single binary bit: $b = 0/1 \Rightarrow$ accept/reject the null hypothesis. This represents a rather terse summary of a data set often consisting of many millions of bits. Furthermore, such tests will nearly always reject given enough data. Null hypotheses are seldom strictly true. It is unlikely that the hypothesized reference distribution $p_0(\mathbf{x})$, or the distribution of the second sample $q(\mathbf{z})$, will be *exactly* equal to that of the observed data $p(\mathbf{x})$. Especially if a universal learning machine is employed, enough data will detect the differences however small between them.

If the null hypothesis cannot be rejected then, at least for the size of the samples used, little additional information concerning the nature of the differences between the distributions is likely to be obtainable. However, rejection should serve as a signal to examine the data further in a attempt to extract the ways in which the distributions differ. Some learning machines such as neural networks, near–neighbor and kernel methods, and support vector machines are "black box" procedures that produce little or no interpretable information. Thus, they are not appropriate for this part of the exercise. Other methods such as decision trees are highly interpretable. For example, a decision tree produces sequences of simple inequalities ("cuts") that identify joint values of the measured variables $\mathbf{x}$ for which $p(\mathbf{x}) >> p_0(\mathbf{x})$, $p(\mathbf{x}) << p_0(\mathbf{x})$, and $p(\mathbf{x}) \simeq p_0(\mathbf{x})$. Such information might yield considerable insight into the mechanism that produced the data.

## References

[1] Bellman, R. E. (1961). *Adaptive Control Processes.* Princeton University Press.

[2] Friedman, J. and Rafsky, L. (1979). Multivariate analogs of the Wald–Wolfowitz and Smirnov two–sample tests. Annals of Statist. **7**, 697.

[3] Friedman, J. and Rafsky, L. (1983). Graph–theoretic measures of multivariate association and prediction. Annals of Statist. **11**, 377.

[4] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer.