

# Bootstrap Estimation of a Non-Parametric Information Divergence Measure

Pradyumna (Prad) Kadambi and Visar Berisha

Arizona State University

Department of Electrical, Computer and Energy Engineering

## Abstract

This work details the bootstrap estimation of a nonparametric information divergence measure, the  $D_p$  divergence measure, applied to the binary classification problem. To address the challenge posed by computing accurate divergence estimates given finite size data, the bootstrap approach is used in conjunction with a power law curve to calculate an asymptotic value of the divergence measure estimate. Monte Carlo estimates of  $D_p$  are found for increasing values of sample data size, and a power law fit is used to find the asymptotic value of the divergence measure as a function of sample size. The fit is also used to generate a confidence interval for the estimate to characterize the quality of the estimator, and the result obtained for the divergence measure is then compared to the result using other resampling methods. Using the inherent relation between divergence measures and classification error rate, an analysis of the Bayes error rate of several test data sets is conducted via this power law estimation approach for  $D_p$ .

## 1 Introduction

Information divergence measures have a wide variety of applications in machine learning, pattern recognition, feature extraction, and big data analysis [8]. The two main classes of information divergence measures are parametric and nonparametric measures. Nonparametric divergence measures, notably including  $f$ -divergences such as the Kullback-Leibler (KL) divergence, measure the difference between two distributions  $F_0$  and  $F_1$ . Arguably the most well known  $f$ -divergence, the KL Divergence is a measure of relative entropy and has applications in coding theory, feature selection, and hypothesis testing [20]. Given these wide variety of applications, there is great interest in estimation of  $f$ -divergences.

Normally, when estimating the divergence between two distributions, we have access to independent and identically distributed (i.i.d) training data from each distribution  $X_i \in c_0$  and  $Y_i \in c_1$  (where  $c_0, c_1$  correspond to two classes of data). The challenge in estimating the divergence measure between two datasets is that the distributions of the data  $F_0$  and  $F_1$  are usually unknown. An  $f$ -divergence,  $D_\phi$ , is of the form:

$$D_\phi(F_0, F_1) = \int_{\Omega} \phi\left(\frac{dF_0}{dF_1}\right) dF_0 \quad (1)$$

given a convex function  $\phi(x)$ , and feature space  $\Omega$  [20]. As we lack knowledge of the distribution functions  $F_0$  and  $F_1$ , a direct computation of  $D_\phi$  is not possible.

A naive method to calculate the divergence between the data is to first find the densities for  $X_i$  and  $Y_i$ , and then calculate the divergence from the computed density estimates. However, as noted in [5] density estimation adds an undesirable intermediate step before the computation of the divergence measure, introduces additional error, and can be difficult for cases of high dimensionality.

In this paper, we perform a bootstrap estimation of a minimum spanning tree based  $f$ -divergence derived in [25] using a power law. From data of size  $N$ , we compute Monte Carlo iterations at  $i$  sample sizes  $n \in \{n_1, n_2, \dots, n_i\} < N$ , and apply the unproven, but reasonable assumption that a power law fit can be used to relate the value of the divergence estimator as a function of sample size. We exploit the unique ability to estimate this divergence measure directly from data, and bypass computing the densities. Utilizing this curve we extrapolate as sample size  $n \rightarrow \infty$ , and find the asymptotic value of the divergence estimate directly from a finite length data set. As  $f$ -divergences are related to the classification error rate [19], this estimation scheme is applied to binary classification examples to find Bayes error rates for several datasets.

The work is organized as follows: the remainder of Section 1 is devoted to background and previous work. Section 1.1-1.2 discuss  $f$ -divergences, their connection to the Bayes optimal error rate, and introduce the specific divergence measure used. Section 1.3 discusses the motivation for the bootstrapped power law estimation method, which is formally introduced in Section 2. In Section 3, examples of the estimation approach are given. In 3.1 we consider generated datasets with known divergence values to demonstrate the accuracy of the estimation algorithm. In 3.2 we perform analysis on the Pima Indians data set and the Banknote data set and compare the calculated Bayes error rate to the classification error rates reported in the literature.

## Background and Previous Work

### 1.1 Divergences Measures

#### 1.1.1 $f$ -divergences

From equation (1), it is clear that  $f$ -divergences are a function of the distributions of the data from each class. In terms of the probability densities  $f_0(x)$  and  $f_1(x)$ , the equation may be rewritten as follows:

$$D_f(f_0, f_1) = \int_{\Omega} f\left(\frac{f_0(\mathbf{x})}{f_1(\mathbf{x})}\right) f_1(\mathbf{x}) d\mathbf{x} \quad (2)$$

The resultant divergence is dependent on the choice of  $f(x)$ . For example, the K-L divergence corresponds to  $f(x) = -\ln(x)$  [6]. A table of commonly used divergences is given below.

Table 1: Commonly Used  $f$ -Divergences

Divergence Measure	$D_f$
K-L Divergence	$\int f_1(x) \ln\left(\frac{f_0(x)}{f_1(x)}\right) dx$
$L^2$ Divergence	$\int (f_0(x) - f_1(x))^2 dx$
Total Variation Distance	$\frac{1}{2} \int  f_0(x) - f_1(x)  dx$
Bhattacharya Distance	$\int \sqrt{f_0(x)f_1(x)} dx$

Note that for some cases the divergence may yield values that are not bounded depending on the choice of  $f(x)$ .

Since in most cases, direct evaluation of the integrals is not possible due to unknown densities, a number of estimation methods have been used to make the problem more tractable. Wang *et al.* [27] derived a nonparametric divergence estimator based on estimating the density ratio  $\frac{dF_0}{dF_1}$ , and in [28] defined a  $k$ -Nearest-Neighbors based divergence estimator that also requires estimates of a density ratio. But, calculation of  $\frac{dF_0}{dF_1}$  rather than  $f_0(\mathbf{x})$  and  $f_1(\mathbf{x})$  still poses the same drawback: it is undesirable to estimate the divergence by performing the intermediate step of estimating a quantity related to the probability distributions.

A key advantage of the  $f$ -divergence we consider is that it can be estimated from the data samples themselves, without intermediate density estimation steps. Towards this end, Hero *et al.* derive a divergence estimator assuming one of the distributions was known. Póczos *et al.* [29] derive estimators for Rényi and  $L_2$  divergences based on  $k$ -Nearest Neighbors statistics, and apply the estimate for classifying astronomical data. We consider the  $f$ -divergence described in [25], which allows for nonparametric estimation directly from sample data via a minimum spanning tree (MST).

### 1.1.2 The $D_p$ Divergence Measure

The aforementioned divergence for probabilities  $p \in (0, 1)$ ,  $q = 1 - p$ , and probability densities  $f_0$  and  $f_1$  is:

$$D_p(f_0, f_1) = \frac{1}{4pq} \left[ \int \frac{(pf_0(\mathbf{x}) - qf_1(\mathbf{x}))^2}{pf_0(\mathbf{x}) + qf_1(\mathbf{x})} d\mathbf{x} - (p - q)^2 \right] \quad (3)$$

To classify  $D_p$  as a statistical distance, it must satisfy the following properties. Firstly,  $0 \leq D_p$ , the divergence must be non-negative. Secondly,  $D_p = 0$  when  $f_0(x) = f_1(x)$ ; the measure between identical distributions must vanish. Third,  $D_p(f_0, f_1) = D_p(f_1, f_0)$ , it must be symmetric. Fourth,  $D_p(f_0, f_2) \leq D_p(f_0, f_1) + D_p(f_1, f_2)$ , the divergence must obey the triangle inequality.  $D_p$  is shown in [25] to have the following properties: it is non-negative ( $0 \leq D_p \leq 1$ ), satisfies the identity property, and is symmetric. However, the triangle inequality has not been proved for the measure, so therefore, we label  $D_p$  as a pseudo-distance.

The estimator for this divergence relies on finding the Friedman-Rafsky (F-R) test statistic:  $\mathcal{C}(\mathbf{X}_f, \mathbf{X}_g)$  from the  $d$ -dimensional class data  $\mathbf{X}_{f_0}$  and  $\mathbf{X}_{f_1}$ . The F-R test statistic is calculated by generating a data set containing both  $\mathbf{X}_{f_0}$  and  $\mathbf{X}_{f_1}$ , finding the Euclidean MST for the data, and counting the number of edges of the MST that connect a point from  $\mathbf{X}_{f_0}$  and  $\mathbf{X}_{f_1}$ . The figure below graphically illustrates how the F-R test statistic is calculated:

In terms of the F-R test statistic, the estimator for  $D_p$  is:

$$1 - \mathcal{C}(\mathbf{X}_{f_0}, \mathbf{X}_{f_1}) \frac{N_{f_0} + N_{f_1}}{2N_{f_0}N_{f_1}} \rightarrow D_p \quad (4)$$

as  $N_{f_0} \rightarrow \infty$  and  $N_{f_1} \rightarrow \infty$ . Given that  $\frac{N_{f_0}}{N_{f_0} + N_{f_1}} \rightarrow p$  and  $\frac{N_{f_1}}{N_{f_0} + N_{f_1}} \rightarrow q$ . Note that  $N_{f_0}$  and  $N_{f_1}$  are the number of samples of data from each class. Using this method,  $D_p$  is estimated from the data samples without any density estimation.

In [2] a modified version of this distance is proposed for implementation in binary classification tasks. As binary classification problems are considered in this work, the modified form of the

distance, and its estimator are used. Notationally,  $\tilde{D}_p$  is used to refer to the modified divergence, and  $D_p$  is used to refer to the distance itself. The same condition that  $N_{f_0} \rightarrow \infty$  and  $N_{f_1} \rightarrow \infty$  is imposed:

$$\tilde{D}_p(f_0, f_1) = \int \frac{(pf_0(\mathbf{x}) - qf_1(\mathbf{x}))^2}{pf_0(\mathbf{x}) + qf_1(\mathbf{x})} d\mathbf{x} \quad (5)$$

$$1 - 2 \frac{\mathcal{C}(\mathbf{X}_{f_0}, \mathbf{X}_{f_1})}{N_{f_0} + N_{f_1}} \rightarrow \tilde{D}_p(f_0, f_1) \quad (6)$$

Note that this quantity is not a distance, as in the case of  $f_0(\mathbf{x}) = f_1(\mathbf{x})$ , it does not satisfy the identity property. However, (5) is estimated rather than (3) as it leads to Bayes error rate bounds that are simpler. Additionally, it is easily seen that when  $p = q = 0.5$ , the identity condition *is* met for  $\tilde{D}_p$ , and for that case  $\tilde{D}_p = D_p$ . For all the cases we consider,  $p = q = 0.5$ . Therefore,  $\tilde{D}_p$  and  $D_p$  are equivalent in the context of this work.

## 1.2 Bayes Error Rate and Divergence Measures

A common problem in machine learning is binary classification, in which data  $\mathbf{X}_i \in \mathbf{R}^{n \times d}$  are assigned a class label  $c_i \in \{0, 1\}$ . Given  $c_0$  and  $c_1$  correspond to data with respective probability distributions  $f_0(\mathbf{x})$  and  $f_1(\mathbf{x})$ , prior probabilities  $p \in (0, 1)$  and  $q = 1 - p$ , the Bayes optimal classifier assigns class labels to  $x_i$  such that the posterior probability is maximized [4]. The error rate of this optimal classifier, the Bayes error rate (BER), provides an absolute lower bound on the classification error rate. Accurate estimation of the BER makes it possible to quantify the performance of a classifier with respect to this optimal lower bound, or apply improved BER bounds to feature selection algorithms [1].

Given the two conditional density functions,  $f_0(\mathbf{x})$  and  $f_1(\mathbf{x})$ , it is possible to write the Bayes error rate in terms of the prior probabilities  $p$  and  $q$ :

$$E_{Bayes} = \int_{r_1} pf_0(\mathbf{x}) d\mathbf{x} + \int_{r_0} qf_1(\mathbf{x}) d\mathbf{x} \quad (7)$$

Here,  $r_1$  and  $r_0$  refer to the regions where the respective posterior probabilities are larger. Direct evaluation of this integral can be quite involved and impractical, and poses similar problems to that of estimation of  $f$ -divergences: it is challenging to create an exact model for the distributions  $f_0(\mathbf{x})$  and  $f_1(\mathbf{x})$ . As an alternative to direct evaluation of the integral, it is possible to derive bounds for the Bayes error rate in terms of divergences measures [5].

The Bayes error rate can be related to the total variation distance (shown in Table 1), which itself can be given in terms of the K-L divergence [30], [31]. The Pinsker inequality [32] and related bounds are one such method to arrive at the total variation distance from the K-L divergence. However, as noted previously, in certain cases the K-L divergence may not be bounded, and can result in a value that tends to  $\infty$ . Vajda [33] modified the relation between the K-L divergence and the total variation distance to account for this problem.

Bounds for the classification error rate have been given in terms of the Bhattacharyya distance in [33]. In [2] the Bayes error rate is given in terms  $\tilde{D}_p$ :

$$\frac{1}{2} - \frac{1}{2} \sqrt{\tilde{D}_p(f_0, f_1)} \leq E_{Bayes} \leq \frac{1}{2} - \frac{1}{2} \tilde{D}_p(f_0, f_1) \quad (8)$$

As expected, when there is no overlap between the two distributions,  $\tilde{D}_p = 1$ , and the BER is lower bounded by zero. In other words, if the two classes are highly separated, it should be possible to design a classifier that has a very low probability of error. On the other hand, if there is full overlap between the two distributions,  $\tilde{D}_p = 0$ , and the BER is 0.5. The optimal error rate is equivalent to the error in randomly assigning class.

### 1.3 Bootstrap Estimation Based on Power Law

As we have just shown, the method for empirically calculating a specific  $D_p$  value for a data set of length  $N$ , and obtaining an estimate for the BER is quite straight forward, but it leaves much to be desired. Specifically, it is necessary to characterize the quality of the  $D_p$  estimate. A direct calculation of the divergence measure using all  $N$  data points yields only a single value, and does not provide any insight into the error or spread of the statistic. Indeed, in many cases knowledge of the spread of the estimate is as important as the estimate itself.

Bootstrap resampling, first introduced by Efron in [10], is a powerful tool to find the spread of an estimator. From a data set  $\mathbf{X}_i$  of size  $N$ , the bootstrap method functions by repeatedly and randomly sampling, with replacement,  $b$  subsets of size  $n < N$  from the original data set. Then estimates are computed for all  $b$  generated subsets. This Monte Carlo approach gives a powerful way to analyze some measure of estimator quality from  $b$  estimates. However, the bootstrap with replacement fails when applied to the F-R test statistic based estimator. Because the F-R test statistic requires the generation of unique distances between data points when computing the minimum spanning tree, it is not desirable to sample with replacement [2].

To satisfy this requirement, we consider another bootstrap resampling technique, the  $m$  out of  $N$  bootstrap, that generates  $b$  randomly sampled subsets of size  $m < N$ , *without replacement*, in order to obtain a sense of the distribution of the estimator. Particularly, we consider the confidence interval of  $D_p$ . Now, we have an estimate of  $D_p$  along with a confidence interval. But, this estimate is for finite data size, and the estimator for  $D_p$ , equation (6), specifies an asymptotic condition of  $N_{f0} \rightarrow \infty$  and  $N_{f1} \rightarrow \infty$ . Obtaining this estimate of  $D_p$  for  $N \rightarrow \infty$  is desirable in order to minimize the bias.

Hawes and Priebe [1] applied a  $k$ -Nearest Neighbors rule to find the upper and lower bound on the asymptotic Bayes error rate as a function of sample size. They perform bootstrap estimates of the BER (which they denote as  $\bar{L}_n(k)$ ) at sample sizes  $n_1 < n_2 < \dots < n_i < N$ . Then they apply a parametric power law curve to calculate the bootstrapped Bayes error rate estimates as a function of sample size,  $n$ :

$$\bar{L}_n(k) = an^b + c \quad (9)$$

with power law fit constants  $a$ ,  $b$ ,  $c$ , and sample size  $n$ . Given that this model is valid,  $b < 0$ , and as  $n \rightarrow \infty$ ,  $\bar{L}_n(k) \rightarrow c$  with  $c = \bar{L}_\infty(k)$ . In [34] it is shown that  $|\bar{L}_n(1) - \bar{L}_\infty(1)| \leq an^{-2}$ ; the absolute error of the BER estimate for a 1-dimensional data, with  $k = 1$  rule, converges in the form given by equation (9).

This result was generalized in [35] for  $d$ -dimensional data. In [36] was generalized to any choice

of  $k$ , and produced the following expression for the BER:

$$\bar{L}_n(k) \approx \bar{L}_\infty(k) + \sum_{j=2}^{\infty} c_j n^{-j/d} \quad (10)$$

As  $n$  increases, the term that dominates happens to be  $cn^{-2/d}$ . This is in agreement with the earlier described result for the  $d = 1$  case. (Please note that for the remainder of this paper, the Bayes error rate will be referred to as  $E_{Bayes}$ , not  $\bar{L}_n(k)$ ).

## 2 Methods

While Hawes and Priebe focus on obtaining asymptotic bounds of the BER, this work focuses on finding the asymptotic value for the  $D_p$  estimator. As shown in equation (8) of Section 1.2, it is possible to simply and directly relate the Bayes error rate to  $D_p$ . Therefore, the motivation behind the power law method for bounding the BER can also motivate an approach to find  $D_p$ . Though it has not been proven, it is a sensible assumption that the divergence estimates follow a similar power law for increasing sample size, and that an asymptotic estimate,  $\bar{D}_p^*$ , may be generated using this formulation. The following power law is used:

$$\bar{D}_p(f_0, f_1) = an^b + c \quad (11)$$

Notice that under the sound assumption of  $b < 0$ ,  $\bar{D}_p^* \rightarrow c$  as  $n \rightarrow \infty$ . So, we have good reason to believe that from a size  $N$  finite length data set, it is possible to obtain asymptotic estimates for the divergence. To find a measure of spread for the divergence estimator, the 95% confidence interval calculated from the curve fitting process. Reviewing notation,  $D_p$  refers to the distance in equation (3),  $\tilde{D}_p$  is the modified version of the distance suited to binary classification given in equation (5), and is equivalent to  $D_p$  for our cases.  $\bar{D}_p$  is the power law curve describing the estimator of  $D_p$  as a function of sample size from the equation above. The asymptotic value of the divergence is denoted as  $\bar{D}_p^*$ .

## 2.1 Algorithm for $\bar{D}_p^*$ Calculation

**Input:** Data  $\mathbf{X}_0, \mathbf{X}_1 \in \mathbf{R}^{n \times d}$  of length  $N$ , dimensionality  $d$   
 $m$ : number of Monte Carlo iterations  
 $i$ : number of bootstrap subsample sizes  $\mathbf{n}_i \in \{n_1, n_2, \dots, n_i < N\}$   
 $\mathbf{X}_S = \mathbf{X}_0 \cup \mathbf{X}_1$

**Result:** Asymptotic estimate of  $D_p$  :  $\bar{D}_p^*$   
Power law curve:  $\mathcal{P}(\bar{\mathbf{D}}_{p_i}, \mathbf{n}_i) = \bar{D}_p(f_0, f_1) = an^b + c$

**Define:**  $\bar{\mathbf{D}}_{p_i} = \{\bar{D}_{p_1}, \bar{D}_{p_2}, \dots, \bar{D}_{p_i}\}$ , bootstrapped estimate for each sample size  $n_i$

**for**  $i \in n_1, n_2, \dots, n_i$  **do**

    Define empty array  $\mathbf{D}_p = \{D_{p_1}, D_{p_2}, \dots, D_{p_m}\}$ , containing the  $m$  Monte Carlo estimates

**for**  $k \in 1 \dots m$  **do**

        Randomly sample a length  $n_i$  subset:  $\mathbf{S} = \{x_1, \dots, x_{n_i}\}$  from  $\mathbf{X}_S$ , without replacement  
        // Ensure  $N_{S,0} = N_{S,1}$ , number of data samples from each class must be equal

        // Compute  $k^{th}$  Monte Carlo estimate

$$D_{p_k} = 1 - 2 \frac{c(\mathbf{S}_0, \mathbf{S}_1)}{N_{S,0} + N_{S,1}}$$

**end**

    // Bootstrapped estimate  $\bar{D}_{p_i}$  is the average of the  $D_{p_k}$

$$\bar{D}_{p_i} = \frac{1}{m} \sum_{k=1}^m D_{p_k}$$

**end**

    // Apply the power law

$$\{a, b, c\} = \mathcal{P}(\bar{\mathbf{D}}_{p_i}, \mathbf{n}_i)$$

$$\bar{D}_p^* = c$$

**Algorithm 1:** Algorithm for finding asymptotic divergence value  $\bar{D}_p^*$

The algorithm for finding the  $\bar{D}_p^*$  value for a two class data set, follows from the overview of bootstrap sampling in 1.3. Then  $m$ , the number of Monte Carlo iterations, must be defined. Choose,  $i$  and  $\mathbf{n}_i$ , the number of bootstrap subsamples and the bootstrap subsample sizes. Begin with the outer loop, and iterate through the number bootstrap subsample sizes,  $i$ . Create a randomly sampled subset  $\mathbf{S}$  of length  $n_i$  from the data  $\mathbf{X}_S$  containing an equal number of elements from each class, and compute the divergence estimate for the subset  $\mathbf{S}$ . Repeat the subset creation and divergence estimation  $m$  times (this is the inner loop). Upon returning to the outer loop, find the mean of the  $m$   $D_{p_k}$  values. Once the mean value of  $m$  estimates for all  $i$  bootstrap subsample sizes has been found, apply the power law fit,  $\mathcal{P}$ , to the mean values and subsample sizes. The asymptotic value of the divergence estimator  $\bar{D}_p^*$  is equal to  $c$ .

We note several restrictions on input parameters. Define maximum value of subsample size as  $n_{max}$ . This value must be less than  $N$ . Also,  $N$  choose  $n_{max}$  must be greater than  $m$ . This is a requirement for sensible Monte Carlo iterations: there must be at least  $m$  unique subsets of size  $n_{max}$ . From the lower extreme of subsample size,  $n_1$  must be greater than the number of dimensions of the data set.

### 3 Results

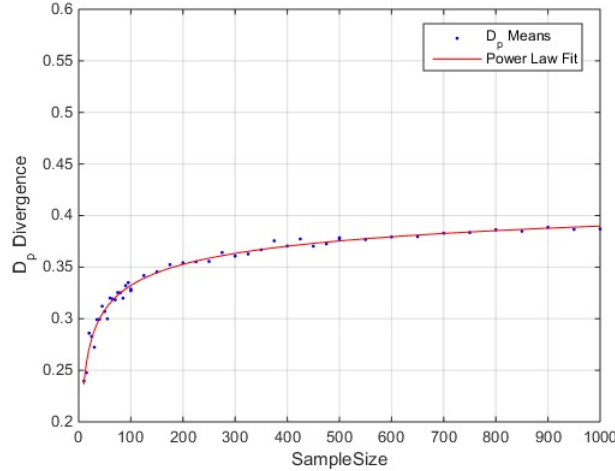
#### 3.1 Uniform Dataset

To test the operation of the estimation algorithm, a data set with a known divergence is constructed in order to ensure that the computed value of  $\bar{D}_p^*$  matches with the known divergence. For this purpose, the uniform distribution shown in Table 2 is defined. The data set contains 8 dimensions, all of which have variance  $\sigma^2 = \frac{1}{12}$ , and are uniformly distributed along  $[-0.5, 0.5]$ , with the exception of one dimension from  $c_1$ . That dimension has an offset mean of  $\mu_1 = \frac{1}{2}$  rather than  $\mu_1 = 0$ . It is easy to see that a direct application of equation (3) or (5) results in a divergence value of  $D_p = 0.5$ . Refer to the Appendix for this computation.

Table 2: Uniform Dataset for Analysis of  $D_p$

$c_0$								
$\mu_0$	0	0	0	0	0	0	0	0
$\sigma_0^2$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
$c_1$								
$\mu_1$	$\frac{1}{2}$	0	0	0	0	0	0	0
$\sigma_1^2$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

Figure 1: Asymptotic Convergence of  $D_p$  for 8-Dimensional Uniform Data Set,  $m = 200$  trials



To find  $\bar{D}_p^*$ , a 10000 point dataset containing an equal number of instances from both classes  $c_0$  and  $c_1$  was created. With respect to Algorithm 1, the parameters of the simulation were:  $N = 10000$ ,  $m = 200$  Monte Carlo iterations,  $i = 50$  bootstrap subsample sizes, and  $n_{50} = 1000$  as the maximum bootstrap sample size. The results of the simulation are shown in Figure 1. The plot shows computed estimates of  $D_p$  as a function of sample size and displays the resulting power law fit. Each blue point on the figure is a  $D_p$  mean - the mean of 200 Monte Carlo trials at each bootstrap sample size  $n_i$ .



The power law found for  $D_p$  for this uniform dataset is:

$$\bar{D}_p = -0.39n^{-0.22} + 0.4775 \quad (12)$$

The asymptotic estimate  $\bar{D}_p^* = 0.4775$  is in approximate agreement with the analytically calculated value for the dataset,  $\mathbf{D}_p = 0.5$ . To understand the true capability of the power law based, asymptotic estimation method consider Table 3.

Table 3: Estimated  $D_p$  for Uniform Data Set for  $n_{max} = 1000$

Value	Result (95% Confidence Interval)
$\mathbf{D}_p$ (true value)	0.5
$D_p$ (no Bootstrap)	0.3370
$D_{p\_mean}$	0.3870 (0.3422, 0.4288)
$\bar{D}_p^*$	<b>0.4775</b> (0.4378, 0.5173)

When a direct computation of the divergence measure is performed for 1000 data points an estimate of  $D_p = 0.337$  is obtained. This is problematic for two reasons. As explained earlier, there is no information about the distribution of the estimate. Additionally, the calculated value  $D_p = 0.337$ , is far from the true value of  $D_p = 0.5$ . The result is of little use for any application.

$\bar{D}_{p\_mean}$  is the average of  $m = 200$  Monte Carlo estimates for a subsample size of 1000. Because the distribution of the estimates are approximately Gaussian, a crude way to characterize the 95% confidence interval of the estimator is to consider values within  $2\sigma$  of the mean. But, the resulting confidence interval and value  $D_{p\_mean} = 0.3870$  are only marginally better than the value found without bootstrapping iterations, and still have large errors.

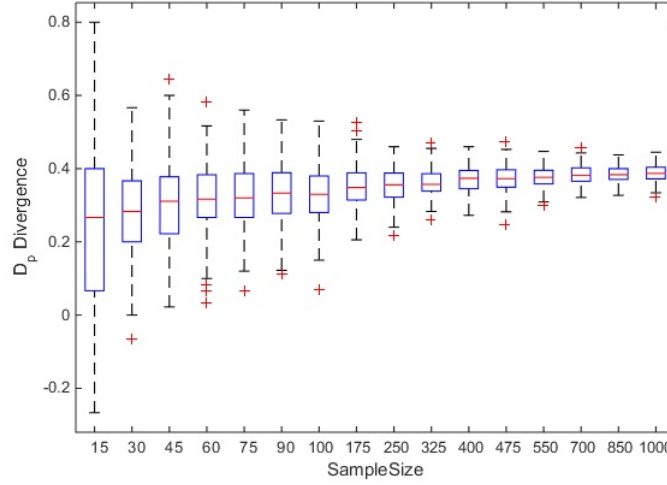
If we then consider the asymptotic power law estimate  $\bar{D}_p^* = 0.4775$  with confidence interval (0.4378, 0.5173), we see that the value of  $\bar{D}_p^* = 0.4775$  is fairly close to the true value of  $\mathbf{D}_p = 0.5$ . This result is far more favorable than the previous results. Additionally, the confidence interval found via the power law fit process includes the true value. Through this example, we have confirmed our assumption that the power law can be applied to find the asymptotic divergence estimate.

### Proper Selection of $\mathbf{n}_i$

While selecting the set of subsample sizes  $\mathbf{n}_i \in \{n_1, n_2, \dots, n_i < N\}$ , it is vital that a significant portion of the  $i$  subsample sizes are concentrated within the rapidly rising portion of the power law curve. In Figure 1, notice that for a sample size of up to  $n = 200$ , the estimates of divergence change rapidly for increasing sample size. But, for  $n > 200$ , the convergence of the estimates slows - the divergence estimates change slowly for increasing sample size. In this case the  $n_i$  are chosen so that  $n_1$  to  $n_{20}$  are spaced evenly on the interval  $[8, 100]$  ( $n_1$  should not be smaller than the number of dimensions). Then,  $n_{21}$  to  $n_{40}$  are evenly spaced for  $[100, 500]$ . Finally,  $n_{41}$  to  $n_{50}$  are evenly spaced between  $[500, 1000]$ .

Although the exact choice of  $\mathbf{n}_i$  may differ between each use case, a useful heuristic to ensure a good power law fit is described. Take the maximum bootstrap subsample size to be  $n_{max} < N$ . In this case,  $n_{max} = 1000$ . Choose approximately  $\frac{1}{3}$  of the  $n_i$  subsamples on the interval  $(0, 0.1n_{max})$ , choose  $\frac{1}{3}$  of the subsamples between  $(0.1n_{max}, 0.5n_{max})$ , and choose the final  $\frac{1}{3}$  in the interval  $(0.5n_{max}, n_{max})$ . If there are fewer number of subsamples  $n_i$  that are small relative to  $n_{max}$ , or if  $n_i$  are evenly spaced along  $(0, n_{max})$ , the goodness of fit for the power law is likely to be compromised. If  $n_i$  must be evenly spaced, we may increase the number of subsamples,  $i$ , and decrease the space between each subsample size to try and preserve a good curve fit.

Figure 2: Distribution of  $D_p$  Values for 8-Dimensional Uniform Data Set,  $m = 200$  trials



An additional benefit of increasing the subsample size, is that the spread of estimator decreases. The same data used to create Figure 1 are shown in Figure 2 to emphasize the decrease in estimator's spread. Recognize that the x-axis is not linearly scaled, and that the y-axis does not have the same scale as Figure 1. For every  $D_p$  point plotted in Figure 1 (every blue point),  $m = 200$  Monte Carlo estimates have been averaged. In Figure 2, box plots of the Monte Carlo iterations are shown for select values of subsample size. Although every single average  $D_p$  value plotted in Figure 1 has a corresponding box plot, only a select number of box plots are shown due to limited space, and to avoid cluttering the Figure. Here, the estimator's bias for small sample sizes is clearly visible in the  $n = 15$  case, as negative values are produced. But, as sample size increases, a dramatic reduction in the interquartile range of the  $n$ .

### 3.2 Gaussian Dataset

Table 4: Gaussian Dataset for Analysis of  $D_p$

$c_0$								
$\mu_0$	0	0	0	0	0	0	0	0
$\sigma_0$	1	1	1	1	1	1	1	1
$c_1$								
$\mu_1$	2.56	0	0	0	0	0	0	0
$\sigma_1$	1	1	1	1	1	1	1	1

We wish to show that this estimation method is valid for many types of distributions. Therefore, we now consider the 8-dimensional Gaussian dataset given in Table 4 [37]. All dimensions of the data are zero mean and unit variance except for the first dimension of class 1. The mean of one of the dimensions of  $c_1$  is shifted to  $\mu_1 = 2.56$ . It is not possible to analytically calculate  $D_p$  for a Gaussian dataset. But, the Bayes error rate for this data set is known (BER=10%). So,  $\bar{D}_p^*$  can be validated by calculating the bounds on the BER from  $\bar{D}_p^*$ , and comparing to the known BER value.

The same conditions as Section 3.1 are applied. A 10000 instance dataset is created containing an equal number of points from both classes. The number of Monte Carlo iterations  $m = 200$ , and bootstrap subsample sizes  $\mathbf{n}_i$  are selected in the same manner with  $n_{max} = 1000$ . The only difference in this case is that a different dataset is evaluated, and an additional step of computing the BER is performed. The resulting power law curve for this dataset is:

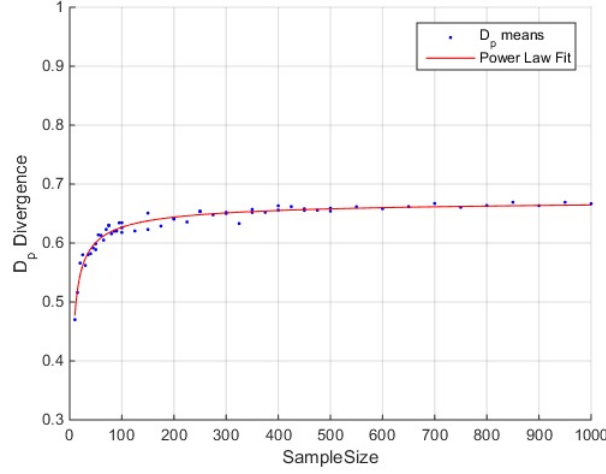
$$\bar{D}_p = -0.79n^{-0.59} + 0.6773 \quad (13)$$

with asymptotic divergence estimate and confidence interval:  $\bar{D}_p^* = 0.6773$  (0.6687, 0.686). Applying equation (8), Table 5 contains resultant bounds on the BER. The true value of the Bayes error rate is within the upper and lower bounds found via the value  $\bar{D}_p^*$ . So,  $\bar{D}_p^*$  has been successfully estimated for the Gaussian case. Refer to Figure 3 for a plot of equation (13).

Table 5: Estimated Bayes Error Rate for Gaussian Data Set for  $n_{max} = 1000$  [4]

Quantity	Bayes Error
True Value	10 %
Estimated Lower Bound	$8.85\% \pm 0.26\%$
Estimated Upper Bound	$16.13\% \pm 0.43\%$

Figure 3: Asymptotic Convergence of  $D_p$  for Gaussian Data Set,  $m = 200$  trials



### 3.3 Banknote Dataset

The first real world example considered is the Banknote Authentication Data Set, taken from the University of California, Irvine Machine Learning Repository [7]. The dataset is 4-dimensional, and has  $N = 1372$  instances. The features of the dataset are extracted from images of genuine and forged banknotes, and the classification task is to label a data vector as either forged or genuine. The dataset consists of a relatively small number of dimensions, and highly separated data, so the convergence is rapid, even for relatively small sample size. The following parameters for Algorithm 1 are set:  $m = 50$ ,  $i = 50$ ,  $n_{max} = 600$ , and most subsamples sizes  $n_i$  are less than  $0.5n_{max}$ . For a sensitive task such as authenticating banknotes, it should not be surprising to see an asymptotic value for  $D_p$  that is almost equal to 1:

$$\bar{D}_p = -3.18n^{-0.98} + 1.001 \quad (14)$$

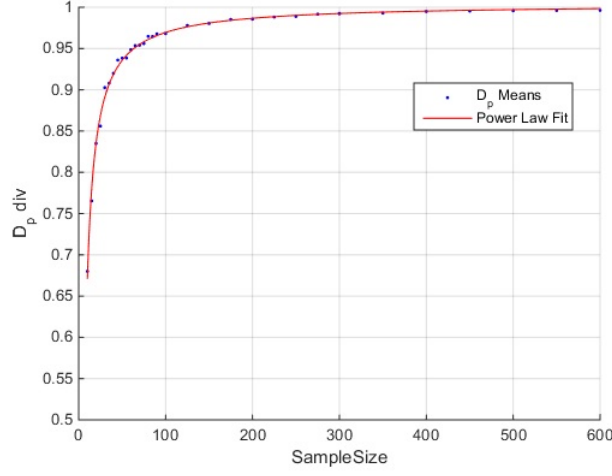
This curve is plotted in Figure 4, along with the means of the  $m = 50$  Monte Carlo trials. The asymptotic value of the divergence estimate and its 95% confidence interval is  $\bar{D}_p^* = 1.000(0.997, 1.005)$ .

Table 6: Estimated Bayes Error Rate For Banknote Data Set

Quantity	Bayes Error
Estimated Lower Bound	$-0.02\% \pm 0.09\%$
Estimated Upper Bound	$-0.05\% \pm 0.20\%$

If equation (8) is applied, the lower and upper bound of the BER are both negative:  $(-0.02\%, -0.05\%)$ . Because a negative error rate is nonsensical, for this dataset the Bayes error rate is taken as 0%. This means that an optimal classifier for this dataset could hope to make no errors at all in sorting banknotes as forged or genuine. This is certainly good news!

Figure 4: Convergence of  $D_p$  for Banknote Authentication Data Set,  $m = 50$  trials



### 3.4 Pima Indian Dataset

The second real world dataset analyzed is the Pima Indian Dataset, also sourced from the UCI Machine learning repository [38]. The dataset has 8-dimensions containing clinical information such as age, blood pressure, BMI, and plasma glucose concentration about female patients of Pima Indian heritage who are age 21 or older. Due to the relatively low number of instances in the dataset, it is of particular interest to find an asymptotic value of the divergence estimator.

Of the  $N = 768$  instances, 500 belong to class 0, corresponding to diabetes, and 268 instances are from class 1, corresponding to diabetic patients. In the discussion of equations (5) and (6) it was noted that in order for the condition  $\tilde{D}_p = D_p$  to hold,  $p = q = 0.5$ . Therefore,  $n_{max} < 2 * 268$  because calculation of the divergence estimate requires an equal number of data samples from each class to ensure  $p = q = 0.5$ . For our purposes  $n_{max} = 500$  is chosen.

For the uniform dataset, it was shown that as the value  $n_i$  increases, the variance in the

Figure 5: Asymptotic Convergence for Pima Indian Data Set,  $m = 50$  trials

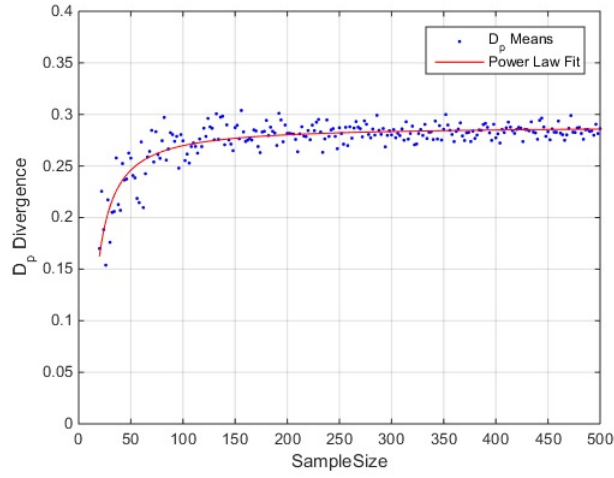
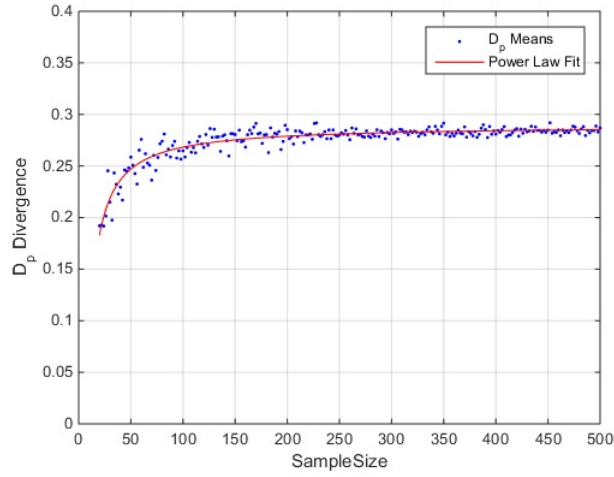


Figure 6: Asymptotic Convergence for Pima Indian Data Set,  $m = 200$  trials



## References

- [1] Hawes, Chad M., and Carey E. Priebe. "A Bootstrap Interval Estimator for Bayes' Classification Error." 2012 IEEE Statistical Signal Processing Workshop, 2012
- [2] V. Berisha, A. Wisler, A.O. Hero, and A. Spanias, "Empirically Estimable Classification Bounds Based on a Nonparametric Divergence Measure" IEEE Transactions on Signal Processing, vol. 64, no. 3, pp.580-591, Feb. 2016.
- [3] A. O. Hero, B. Ma, O. Michel, and J. Gorman, Alpha-divergence for classification, indexing

Figure 7: Asymptotic Convergence for Pima Indian Data Set,  $m = 5000$  trials

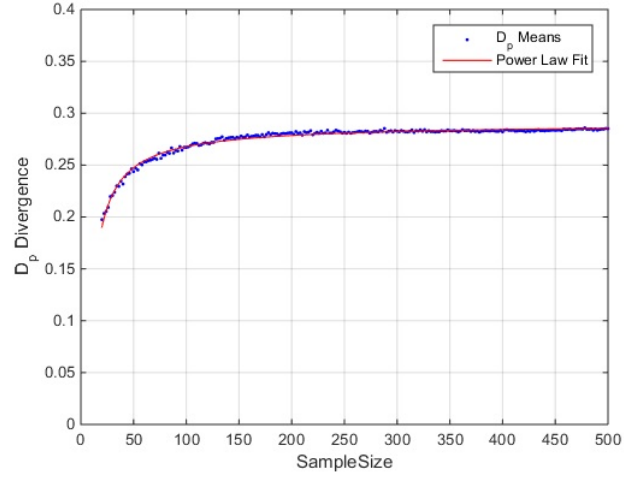


Table 7: Bayes Error Rates in Literature for Pima Indians Data Set [4]

Algorithm	Bayes Error Rate (%)
Discrim	22.50
Quadisc	26.20
Logdisc	22.30
SMART	23.20
ALLOC80	30.10
K-NN	32.40
CASTLE	25.80
CART	25.50
IndCART	27.10
NewID	28.90
AC2	27.60
Baytree	27.10
NaiveBay	26.20
CN2	28.90
C4.5	27.00
Itrule	24.50
Cal5	25.00
Kohonen	27.30
DIPOL92	22.40
Backprob	24.80
RBF	24.30
LVQ	27.20

Table 8: Bootstrap Estimated Bayes Error Rates for Pima Indians Data Set [4]

Algorithm	Bayes Error Rate (%)
$D_p$ (no Bootstrap)	$29.32 \pm 6.22$ *
Efron Bootstrap	$14.87 \pm 2.465$ **
$m < n$ Bootstrap, $m = 200$	$23.13 \pm 4.13$
$D_p$ <b>Asymptotic Power Law</b>	<b><math>22.78 \pm 0.11</math></b>

Table 9:  $D_p$  and Bayes Error Rate for the Pima Indian Data Set for Increasing Sample Size, and Increasing Monte Carlo Iterations

Sample Size	Monte Carlo Iterations	$D_p$ Asymptotic Value (95% Confidence Interval)	Bayes Error Rate (%), ( $\pm$ 95% CI) Lower Bound	Bayes Error Rate (%), ( $\pm$ 95% CI) Upper Bound
100	50	0.2725 (0.245, 0.3)	$23.90 \pm 1.32$	$36.38 \pm 1.38$
100	200	0.2958 (0.265, 0.3267)	$22.81 \pm 1.42$	$35.21 \pm 1.54$
100	5000	0.3107 (0.2959, 0.3254)	$22.13 \pm 0.67$	$34.47 \pm 0.75$
200	50	0.2946 (0.2732, 0.3161)	$22.86 \pm 0.99$	$35.27 \pm 1.07$
200	200	0.3029 (0.288, 0.3178)	$22.48 \pm 0.68$	$34.86 \pm 0.74$
200	5000	0.3162 (0.3114, 0.3209)	$21.88 \pm 0.21$	$34.19 \pm 0.24$
300	50	0.3118 (0.2827, 0.3409)	$22.08 \pm 1.31$	$34.41 \pm 1.46$
300	200	0.3073 (0.2926, 0.3219)	$22.28 \pm 0.66$	$34.63 \pm 0.74$
300	5000	0.3041 (0.3006, 0.3075)	$22.43 \pm 0.16$	$34.79 \pm 0.18$
500	50	0.2886 (0.2855, 0.2917)	$23.14 \pm 0.14$	$35.57 \pm 0.15$
500	200	0.2895 (0.2871, 0.2918)	$23.10 \pm 0.11$	$35.53 \pm 0.12$
500	5000	0.2963 (0.2939, 0.2987)	$22.78 \pm 0.11$	$35.19 \pm 0.12$

and retrieval, Communication and Signal Processing Laboratory, Technical Report CSPL-328, U. Mich, 2001

[4] K. Tumer, K. (1996) "Estimating the Bayes error rate through classifier combining" in Proceedings of the 13th International Conference on Pattern Recognition, Volume 2, 695699

[5] Tumer, Kagan, and Joydeep Ghosh. "Bayes Error Rate Estimation Using Classifier Ensembles." International Journal of Smart Engineering System Design 5.2 (2003): 95-109.

[6] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[7] V. Lohweg, Banknote Authentication Data Set, 2012. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ba



- [8] K. Pranesh, and L. Hunter. "On an Information Divergence Measure and Information Inequalities." (n.d.): n. pag. University of Northern British Columbia.
- [9] Tukey, J.W. 1958. Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics* 29: 614
- [10] Efron, B. "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics* 7.1 (1979)
- [19] S. Ali and S. D. Silvey, A general class of coefficients of divergence of one distribution from another, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 131-142, 1966.
- [20] Nguyen, Xuanlong, Martin J. Wainwright, and Michael I. Jordan. "Nonparametric Estimation of the Likelihood Ratio and Divergence Functionals." 2007 IEEE International Symposium on Information Theory (2007)
- [21] Sugiyama, Masashi, Song Liu, Marthinus Christoffel Du Plessis, Masao Yamanaka, Makoto Yamada, Taiji Suzuki, and Takafumi Kanamori. *Journal of Computing Science and Engineering* 7.2 (2013)
- [22] Nguyen, Xuanlong, Martin J. Wainwright, and Michael I. Jordan. "Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization." *IEEE Trans. Inform. Theory* *IEEE Transactions on Information Theory* 56.11 (2010)
- [23] L. Song, M.D. Reid, A.J. Smola, and R.C. Williamson. Discriminative estimation of  $f$ -divergence. Submitted to AISTATS09, October 2008.
- [24] Tumer, Kagan, and Joydeep Ghosh. "Bayes Error Rate Estimation Using Classifier Ensembles." *International Journal of Smart Engineering System Design* 5.2 (2003)
- [25] Berisha, Visar, and Alfred O. Hero. "Empirical Non-Parametric Estimation of the Fisher Information." *IEEE Signal Processing Letters* *IEEE Signal Process. Lett.* 22.7 (2015)
- [26] S. Kullback and R. A. Leibler, On information and sufficiency, *The Annals of Mathematical Statistics*, pp. 79-86, 1951.
- [27] Wang, Q., S.r. Kulkarni, and S. Verdu. "Divergence Estimation of Continuous Distributions Based on Data-Dependent Partitions." *IEEE Trans. Inform. Theory* *IEEE Transactions on Information Theory* 51.9 (2005)
- [28] Wang, Qing, Sanjeev R. Kulkarni, and Sergio Verdu. "Divergence Estimation for Multidimensional Densities Via K-Nearest-Neighbor Distances." *IEEE Trans. Inform. Theory* *IEEE Transactions on Information Theory* 55.5 (2009)
- [29] Barnabs Pczos, Liang Xiong, Jeff G. Schneider, Nonparametric Divergence Estimation with Applications to Machine Learning on Distributions. *UAI 2011*: 599-608

- [30] T. Kailath, The divergence and Bhattacharyya distance measures in signal selection, *Communication Technology, IEEE Transactions on*, vol. 15, no. 1, pp. 5260, 1967.
- [31] I. Csisz et al., Information-type measures of difference of probability distributions and indirect observations, *Studia Sci. Math. Hungar.*, vol. 2, pp. 299318, 1967
- [32] I. Vajda, Note on discrimination information and variation (corresp.), *Information Theory, IEEE Transactions on*, vol. 16, no. 6, pp. 771773, 1970.
- [33] A. Bhattacharyya, On a measure of divergence between two multinomial populations, *Sankhya: The Indian Journal of Statistics* , pp. 401 406, 1946.
- [34] Thomas M. Cover, Rates of convergence of nearest neighbor decision procedures, in *Proceedings of 1st Annual Hawaii Conference on Systems Theory*, 1968, pp. 413415
- [35] Demetri Psaltis, Robert R. Snapp, and Santosh S. Venkatesh, On the finite sample performance of the nearest neighbor classifier, *IEEE Transactions on Information Theory*, vol. 40, no. 3, pp. 820837, 1994
- [36] Robert R. Snapp and Santosh S. Venkatesh, Asymptotic expansions of the k nearest neighbor risk, *The Annals of Statistics*, vol. 26, no. 3, pp. 850878, 1998
- [37] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic press, 1990
- [38] A. Frank and A. Asuncion, UCI machine learning repository, <http://archive.ics.uci.edu/ml>, 2010.