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, a bootstrap approach is used in conjunction with a power law curve to calculate an asymptotic estimate of the divergence measure in question. 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 the 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_ϕ, is of the form:

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

given a convex function $\phi(x)$, and feature space Ω [20]. As we lack knowledge of the distribution functions, a direct computation of D_ϕ 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] adding this naive approach 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 estimator as a function of sample size. We exploit the unique ability to estimate this divergence measure directly from data, and bypass density estimation. Utilizing this curve we extrapolate as sample size $n \to \infty$, and find the asymptotic value of the divergence estimate directly from a finite length dataset. As f-divergences may be related to the classification error rate, this estimation scheme is applied to binary classification to find Bayes error rates.

The work is organized as follows: the remainder of Section 1 is devoted to background and previous work. Section 2 introduces the bootstrap sampling method, and the power law used in this estimation method. In Section 3 we apply the method to several generated and real-world datasets to show that the power law method can successfully be used to calculate the divergence and classification error rate of several distributions. In 3.1 we consider generated datasets with known divergence values, to demonstrate the accuracy of the divergence estimates. In 3.2 we perform analysis on the Pima Indians dataset and the Banknote dataset 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), 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:

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

The resultant divergence (such as K-L 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	$\int f_1(x) ln(\frac{f_0(x)}{f_1(x)}) dx$
L^2 Divergence	$\int (f_0(x) - f_1(x))^2 dx$
Total Variation Distance	$\frac{1}{2} \int \lvert f_0(x) - f_1(x) \rvert dx$
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 an k-Nearest-Neighbors based divergence estimator that also requires estimates of a density ratio. But, estimation of $\frac{dF_0}{dF_1}$ instead of $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ independently still poses the same drawback: it is undesirable to estimate the divergence by performing the intermediate step of estimating the probability distribution.

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 densities was known. Póczos *et al.* [29] derive estimators for Rényi and L_2 divergences based on k-Nearest Neighbors statistics, and apply the estimate to 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:

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

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 done by generating a dataset 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:

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

as $N_{f0} \to \infty$ and $N_{f_1} \to \infty$. Given that $\frac{N_{f_0}}{N_{f_0}+N_{f_1}} \to p$ and $\frac{N_{f_1}}{N_{f_0}+N_{f_1}} \to q$. Note that N_{f0} and N_{f1} 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, $\widetilde{D}_p$ is used to refer to the modified divergence, and D_p is used to refer to the distance itself. The same condition that $N_{f0} \to \infty$ and $N_{f_1} \to \infty$ is imposed:

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

$$1 - 2\frac{\mathcal{C}(\mathbf{X}_{f_0}, \mathbf{X}_{f_1})}{N_{f_0} + N_{f_1}} \to \widetilde{D}_p(f_0, f_1) \tag{6}$$

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 $\widetilde{D}_p$, and for that case $\widetilde{D}_p = D_p$.

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:

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

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 given in terms of the K-L divergence [30], [31]. The Pinkser 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 ∞. 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 Bhattacharya distance in [33]. In [2] the Bayes error rate is given in terms $\widetilde{D}_p$:

$$\frac{1}{2} - \frac{1}{2}\sqrt{\widetilde{D}_p(f_0, f_1)} \le E_{Bayes} \le \frac{1}{2} - \frac{1}{2}\widetilde{D}_p(f_0, f_1) \tag{8}$$

1.3 Bootstrap Estimation Based on Power Law

As we have just shown, the method for empirically calculating a specific $\widetilde{D}_p$ value for a dataset 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 $\widetilde{D}_p$ 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 dataset. 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 $\widetilde{D}_p$. Now, we have an estimate of $\widetilde{D}_p$ along with a confidence interval. But, this estimate is for finite data size, and the estimator for $\widetilde{D}_p$, equation (6), specifies an asymptotic condition of $N_{f0} \to \infty$ and $N_{f_1} \to \infty$. Obtaining this estimate of $\widetilde{D}_p$ for $N \to \infty$ 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 $n_1<n_2<...<n_i<N$. Then they apply a parametric power law curve to calculate the bootstrapped Bayes error rate estimates as a function of sample size, n:

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

with power law fit constants a, b, c, and sample size n. Given that this model is valid, $b < 0$, and as $n \to \infty$, $\bar{L}_n(k) \to c$ 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:

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

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 $\bar{L}_n(k)$).

While Hawes and Priebe focus on obtaining asymptotic bounds of the BER, this works focuses on asymptotic bounds for $\widetilde{D}_p$. As shown in equation (8) of Section 1.2, it is possible to simply and directly relate the Bayes error rate in terms of $\widetilde{D}_p$. Therefore, the motivation behind the estimation method for BER may also motivate an approach to find $\widetilde{D}_p$. 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 asymptotic estimates may be generated using this formulation. In this work, the following power law is used:

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

Notice that under the good assumption of $b < 0$, $\bar{D}_{p\infty} \to c$ as $n \to \infty$, . So, we have shown that from a size N finite length dataset, it is possible to obtain asymptotic estimates for the divergence. Reviewing notation, D_p refers to the distance in equation (3), $\widetilde{D}_p$ is the modified version of the distance suited to binary classification given in equation (5), and $\bar{D}_p$ is the power law curve describing the estimator of $\widetilde{D}_p$ as a function of sample size from the equation above. The asymptotic value of the divergence is denoted as $\bar{D}_{p\infty}$.

2 Methods

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

Result: Asymptotic estimate of $\widetilde{D}_p$: $\bar{D}_{p_\infty}$
Power law curve: $\bar{D}_p(f_0, f_1) = an^b + c$

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

for $i \in n_1, n_2, ..., n_i$ **do**
  Define empty array $\widetilde{\mathbf{D}}_p = \{\widetilde{D}_{p_1}, \widetilde{D}_{p_2}, ..., \widetilde{D}_{p_m}\}$, containing the m Monte Carlo estimates

  for $k \in 1...m$ **do**
    Randomly sample a subset $\mathbf{S} = \{x_1, ..., x_j\}$ from $\mathbf{X}_S$ without replacement
    Ensure $N_{S,0} = N_{S,1}$ // number of data samples from each class must be equal

    $\widetilde{D}_{p_k} = 1 - 2\frac{\mathcal{C}(\mathbf{S}_0, \mathbf{S}_1)}{N_{S,0}+N_{S,1}}$ // compute k^{th} Monte Carlo estimate
  end

  $\bar{D}_{p_i} = \frac{1}{m}\sum_{k=1}^{m} \widetilde{D}_{p_k}$ // the Bootstrapped estimate $\bar{D}_{p_i}$ is the average of the $\widetilde{D}_{p_k}$
end

Algorithm 1: Algorithm for finding asymptotic divergence value $\bar{D}_{p\infty}$

In this section we outline the method of estimating D_p.

3 Results

3.1 Uniform Dataset

To test the operation of the estimation algorithm, we generate a dataset with a known divergence in order to ensure that the bootstrapped, asymptotic value of D_p matches with the analytically computed divergence. For this purpose, the uniform distribution shown in Table 1 is chosen (as opposed to a distribution like a Gaussian) due to the ease of performing the analytical computation. We define an 8 dimensional dataset, where each dimension of data has variance $\sigma^2 = frac112$ and is uniformly distributed along $[-0.5, 0.5]$ with the exception of one dimension from class 1. That dimension has mean offset to $\mu_0 = \frac{1}{2}$, and a direct application of equation 3 result in a divergence value of $D_p = 0.5$.

Table 2: Uniform Dataset for Bootstrap Analysis of D_p

D_0								
μ_0	0	0	0	0	0	0	0	0
σ_0^2	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
D_1								
μ_1	$\frac{1}{2}$	0	0	0	0	0	0	0
σ_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, N = 200 trials

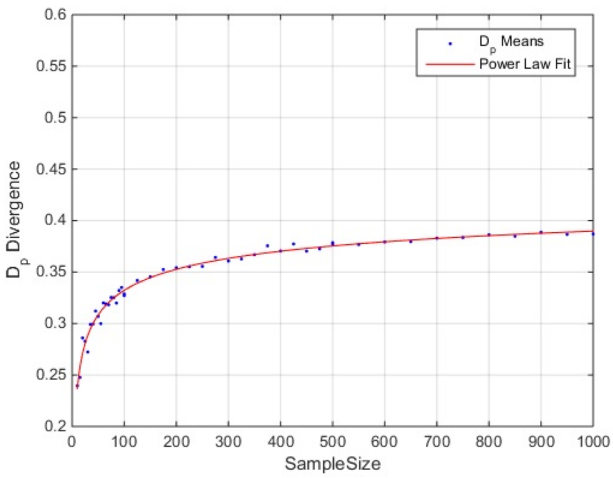
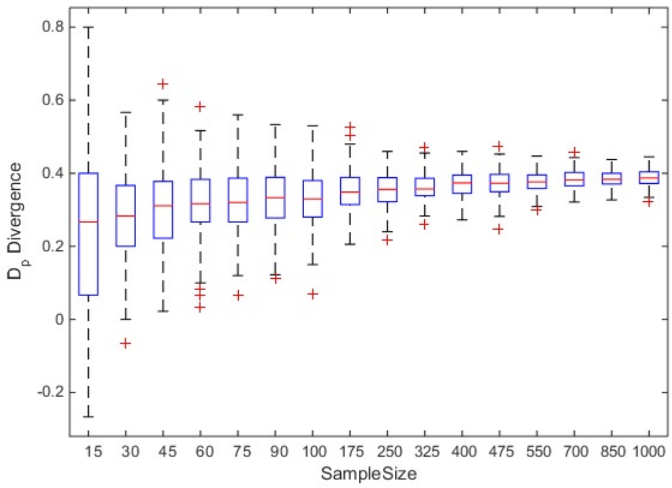


Figure 2: Distribution of D_p Values for 8-Dimensional Uniform Data Set, N = 200 trials



3.2 Gaussian Dataset

Table 3: Gaussian Dataset for Bootstrap Analysis of D_p

D_0								
μ_0	0	0	0	0	0	0	0	0
σ_0	1	1	1	1	1	1	1	1
D_1								
μ_1	0	0	0	0	0	0	0	0
σ_1	2.56	1	1	1	1	1	1	1

Figure 3: Asymptotic Convergence of D_p for Gaussian Data Set, N = 50 trials

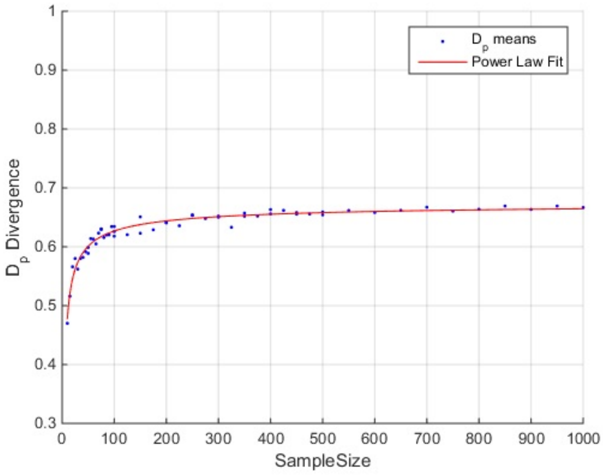
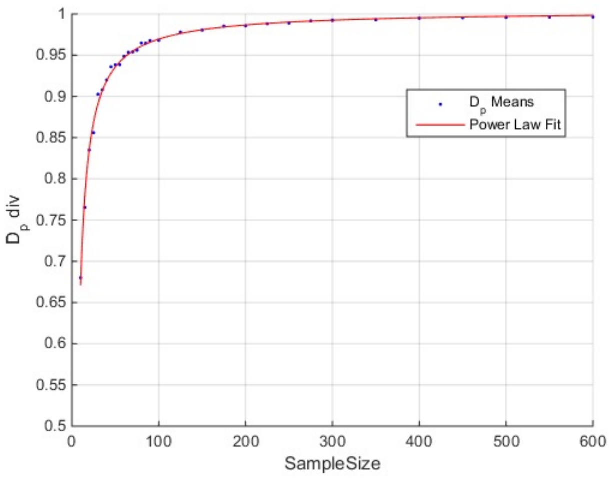


Figure 4: Convergence of D_p for Banknote Authentication Data Set, N = 50 trials



3.3 Banknote Dataset

The empirical example we consider is the Banknote Authentication Data Set taken from the University of California, Irvine Machine Learning Repository [7]. The 4-dimensional dataset contains data extracted from images of banknotes. The data set consists of a relatively small number of dimensions, and highly separated data, so the convergence is rapid, even for relatively small sample size. We note that for a sensitive task such as authenticating banknotes, it should not be surprising to see an asymptotic value for D_p that is close to 1, indicating that the classes are well separated.

3.4 Pima Indians Dataset

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

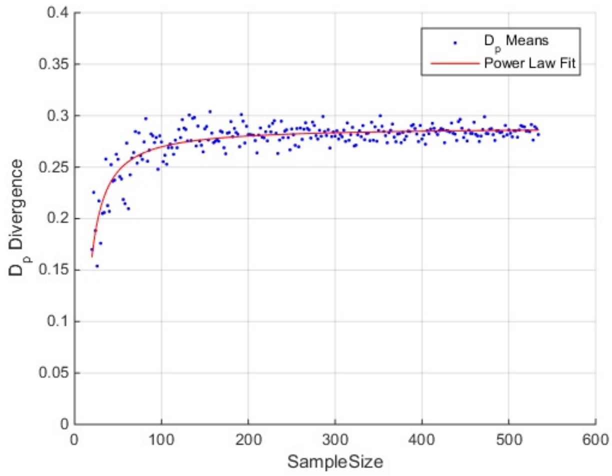


Figure 6: Asymptotic Convergence for Pima Indian Data Set, N = 200 trials

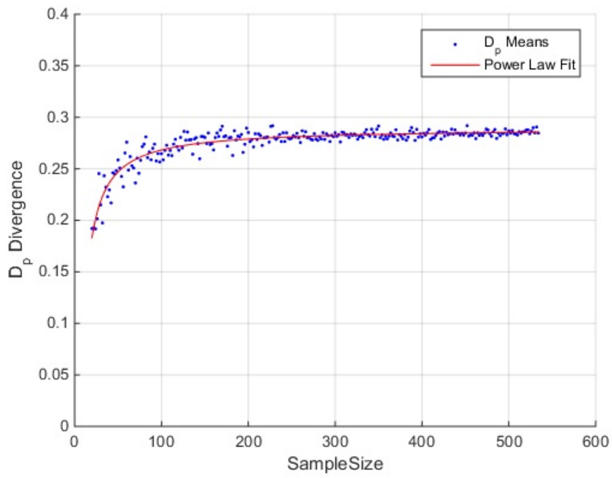


Figure 7: Asymptotic Convergence for Pima Indian Data Set, N = 5000 trials

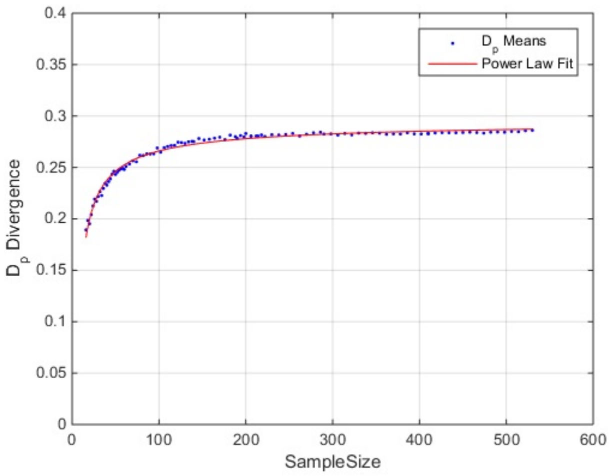


Table 4: Bayes Error Rates in Literature for Pima Indians Data Set [4]

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

Algorithm	Bayes Error Rate (%)
D_p (no Bootstrap)	29.32 ± 6.22 *
Efron Bootstrap	14.87 ± 2.465 **
$m < n$ Bootstrap, $m = 200$	23.13 ± 4.13
D_p **Asymptotic Power Law**	$\mathbf{23.95 \pm 0.11}$

Table 6: 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)	Classification Error Rate (%), (95% CI)
100	50	0.2725 (0.245, 0.3)	23.90 ± 1.32
100	200	0.2958 (0.265, 0.3267)	22.81 ± 1.42
100	5000	0.3107 (0.2959, 0.3254)	22.13 ± 0.67
200	50	0.2946 (0.2732, 0.3161)	22.86 ± 0.99
200	200	0.3029 (0.288, 0.3178)	22.48 ± 0.68
200	5000	0.3162 (0.3114, 0.3209)	21.88 ± 0.21
300	50	0.3118 (0.2827, 0.3409)	22.08 ± 1.31
300	200	0.3073 (0.2926, 0.3219)	22.28 ± 0.66
300	5000	0.3041 (0.3006, 0.3075)	22.43 ± 0.16

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 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)

[11] https://www.princeton.edu/ verdu/reprints/WanKulVerSep2005.pdf **Create Citation

[12] http://www.princeton.edu/ verdu/reprints/WanKulVer.May2009.pdf?q=tilde/verdu/ reprints/WanKulVer.Ma **Create Citation

[13] http://www.eecs.berkeley.edu/ wainwrig/Papers/NguWaiJor10.pdf **Create Citation

[14] T. Kailath, The divergence and Bhattacharyya distance measures in signal selection, Communication Technology, IEEE Transactions on, vol. 15, no. 1, pp. 5260, 1967.

[16] http://arxiv.org/pdf/1404.6230.pdf** Create Citation

[17] Bootstrap sampling Efron Citation

[18]

[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. 131142, 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. 7986, 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