Bootstrap Estimation of a Non-Parametric Information Divergence Measure

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 to 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 convergence value of the divergence measure as a function of sample size. The fit is also used to generate a confidence interval for the estimate which allows us to characterize the quality of the estimator. The results obtained for the divergence measure are then compared to several other resampling methods. Utilizing 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 the 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 P and Q. 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 P and Q are usually unknown. An f-divergence, D_ϕ, is of the form:

$$D_\phi(P||Q) = \int_\Omega \phi\left(\frac{dP}{dQ}\right) dP \tag{1}$$

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 intermediate step before the computation of the divergence measures 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 a power law curve of the form:

$$\bar{D}_\phi(n) = an^b + c \tag{2}$$

Utilizing this curve we extrapolate as sample size $n \to \infty$. Exploiting the ability to estimate this divergence measure directly from data, computation of the densities is bypassed, and an asymptotic value of the estimator is found from a finite length dataset. As f-divergences may be used to bound the classification error rate, the divergence estimate is applied in a binary classification setting.

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 estimation method. In Section 3 we will 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 the generated example datasets, and in 3.2 we perform analysis on the Pima Indians dataset and the Banknote dataset found in the University of California, Irvine machine learning repository [6].

Background and Previous Work

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 $p(x)$ and $g(x)$, the equation may be written as follows:

$$D_f(p(x), g(x)) = \int_\Omega f\left(\frac{p(x)}{g(x)}\right) g(x) dx \tag{3}$$

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.

Divergence Measure	D_f
K-L Divergence	$\int g(x) ln(\frac{p(x)}{g(x)}) dx$
L^2 Divergence	$\int (p(x) - g(x))^2 dx$
Total Variation Distance	$\frac{1}{2} \int \lvert p(x) - g(x) \rvert dx$
Bhattacharya Distance	$\int \sqrt{p(x) g(x)} dx$

Note that for some cases the divergence may yield values that are not bounded depending on the choice of $f(x)$. An undefined or unbounded K-L divergence result can be problematic if we then apply the result to a task, and it may be desirable use a bounded divergence measure.

Since in most cases, direct evaluation of the expression containing these $f(x)$ is not possible, 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 first estimating the density ratio $\frac{dP}{dQ}$, and in [28] defined an k-Nearest-Neighbors based divergence estimator that also requires estimates of a density ratio. But estimation of $\frac{dP}{dQ}$ instead of dP and dQ independently still poses the same drawback as estimating dP and dQ: 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 can be estimated from the training data samples themselves, without independent 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. The f-divergence described in [25] allows for nonparametric estimation directly from data via a minimum spanning tree(MST).

1.2 The D_p Divergence Measure

The aforementioned distance given in [25] for probability $p \in (0,1)$, $q = 1-p$, and probability densities f and g is:

$$D_p = \frac{1}{4pq}\left[\int \frac{(pf(\mathbf{x}) - qg(\mathbf{x}))^2}{pf(\mathbf{x}) + qg(\mathbf{x}))} d\mathbf{x} - (p-q)^2\right] \tag{4}$$

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 $\mathbf{X}_f$ and $\mathbf{X}_g$. The F-R test statistic is calculated done by generating a dataset containing both $\mathbf{X}_f$ and $\mathbf{X}_g$, finding the Euclidean MST for the data, and counting the number of edges of the MST that connect a point from $\mathbf{X}_f$ and $\mathbf{X}_g$ [2]. In terms of the F-R test statistic, the estimator for D_p is shown in [2]:

$$1 - \mathcal{C}(\mathbf{X}_f, \mathbf{X}_g)\frac{N_f + N_g}{2N_f N_g} \to D_p \tag{5}$$

as $N_f \to \infty$ and $N_g \to \infty$. Given that $\frac{N_f}{N_f+N_g} \to p$ and $\frac{N_g}{N_f+N_g} \to q$. Using this method, D_p is estimated from the data samples without any density estimation. As expected, D_p has the following properties : $0 \leq D_p \leq 1$, $D_p = 0$ when $f(x) = g(x)$, and $D_p(f,g) = D_p(g,f)$.

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:

$$\widetilde{D}_p = \int \frac{(pf(\mathbf{x}) - qg(\mathbf{x}))^2}{pf(\mathbf{x}) + qg(\mathbf{x}))} d\mathbf{x} \tag{6}$$

$$1 - 2\frac{\mathcal{C}(\mathbf{X}_f, \mathbf{X}_g)}{N_f + N_g} \to \widetilde{D}_p \tag{7}$$

Although this quantity is not a distance, as in the case of $p \neq q$ and $f(x) = g(x)$, (6) is estimated rather than (4) as it leads to less complex expressions for bounding the classification error rate [2]. Additionally, it is easily seen that $\widetilde{D}_p = D_p$ when the condition $p = q = 0.5$ is met.

1.3 Bayes Error Rate and Divergence Measures

A common problem in machine learning is binary classification, in which data $\mathbf{X}_i \in \mathbf{R}^{\mathbf{n}\times\mathbf{d}}$ are assigned a class label $c_i \in \{0,1\}$. Given c_0 and c_1 contain data with respective probability distributions $f(\mathbf{x})$ and $g(\mathbf{x})$, 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 estimation methods in a feature selection algorithm [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 as given in [2]:

$$E_{Bayes} = \int_{r_1} p f_0(\mathbf{x})\, d\mathbf{x} + \int_{r_0} q f_1(\mathbf{x})\, d\mathbf{x} \tag{8}$$

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].

1.4 Bootstrap Sampling

As we have just shown, the method for empirically calculating a specific D_p value for a dataset of length N is quite straight forward, but it leaves much to be desired. Specifically, it is desirable to characterize the quality of the D_p estimate, and have the ability to obtain a. 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 method to find the spread of an estimator. From a data set $\mathbf{X}_i$ of size D, bootstrap method functions by taking i sets of repeated random samples of size $n < D$ with replacement, and computing estimates for all i generated sets. This Monte Carlo approach gives a powerful way to analyze some measure estimator quality. However, the bootstrap with replacement fails when applied to the F-R test statistic based estimator. Resampling techniques such as the jackknife [9], and the bootstrap [10] can be applied to find the statistical distribution of the estimated quantity in question.

2 Methods

Input : Data $x_0, x_1 \in \mathbf{R^n}$ of length N, B Monte-Carlo iterations,
n_i Bootstrap subsample sizes
Result: Estimate of D_p for
for $j \in n_1 ... n_i$ **do**
  if then
    go to next section;
    current section becomes this one;
  else
    go back to the beginning of current section;
  end
end

Algorithm 1: How to write algorithms

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 1: 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

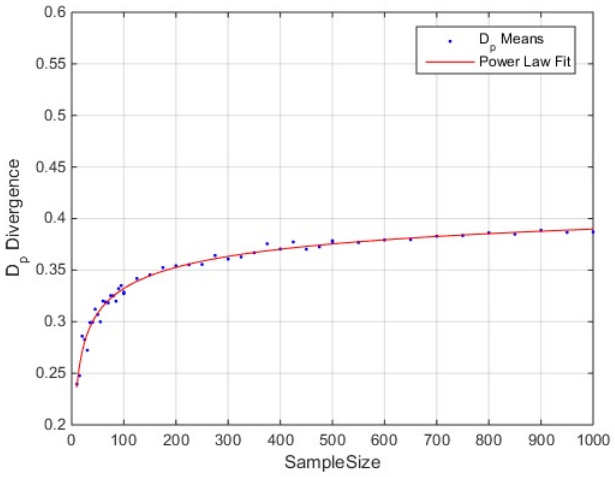
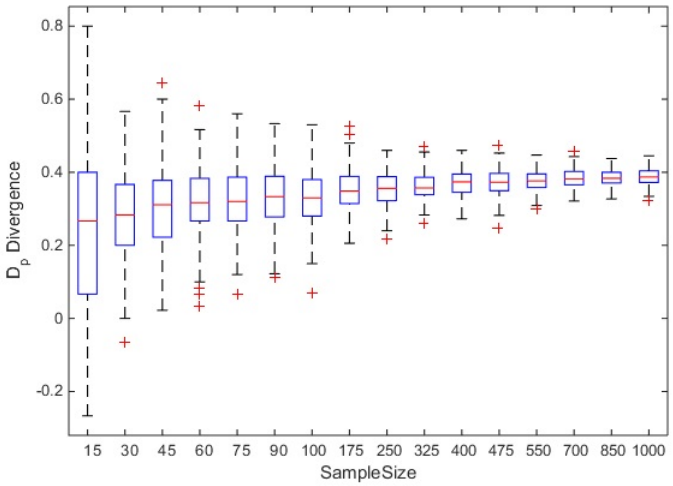


Figure 2: Distribution of D_p Values for 8-Dimensional Uniform Data Set, N = 200 trials



3.2 Gaussian Dataset

Table 2: 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

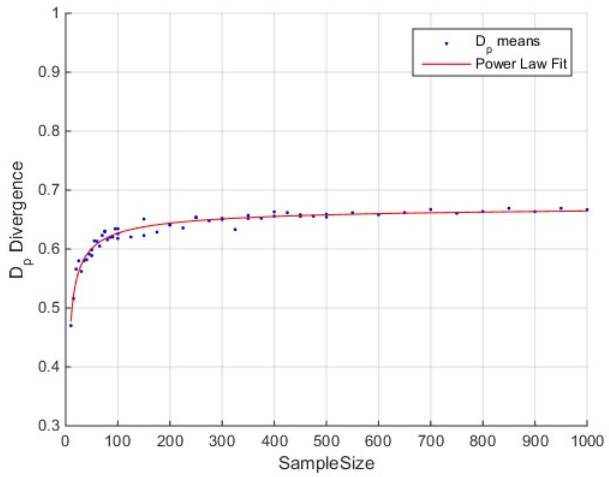
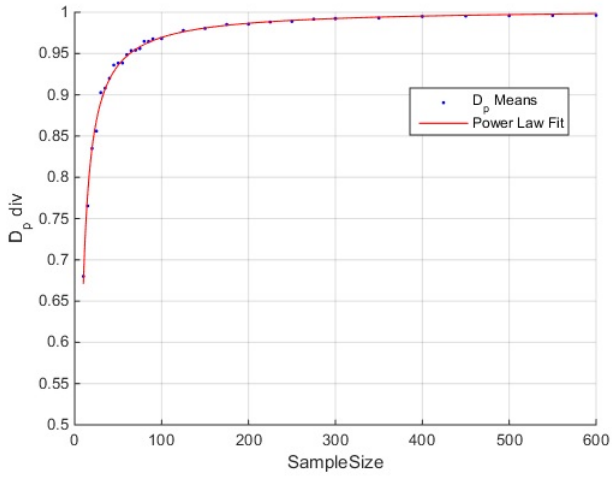


Figure 4: Convergence of D_p for Banknote Authentication Data Set, N = 50 trials



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

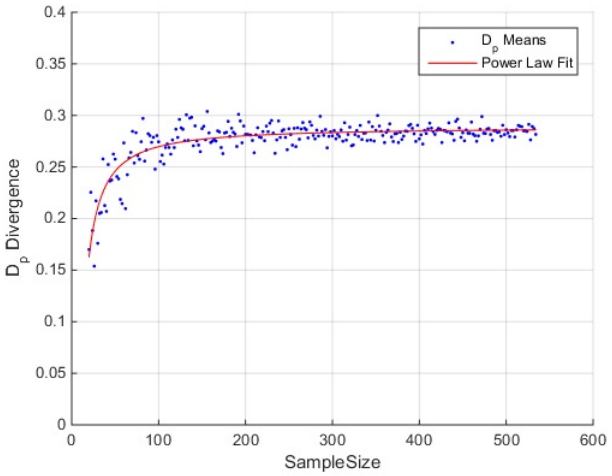


Figure 6: Asymptotic Convergence for Pima Indian Data Set, N = 200 trials

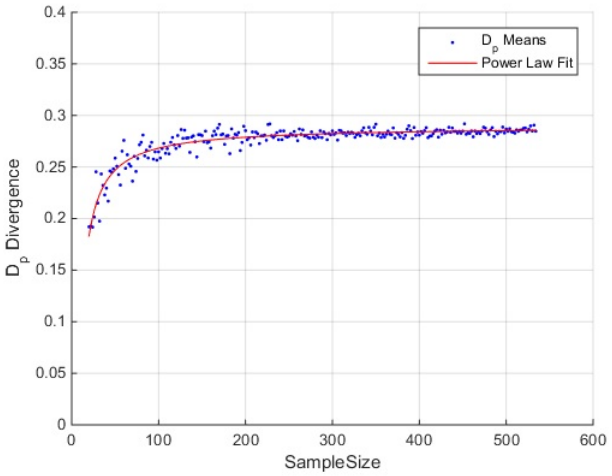


Figure 7: Asymptotic Convergence for Pima Indian Data Set, N = 5000 trials

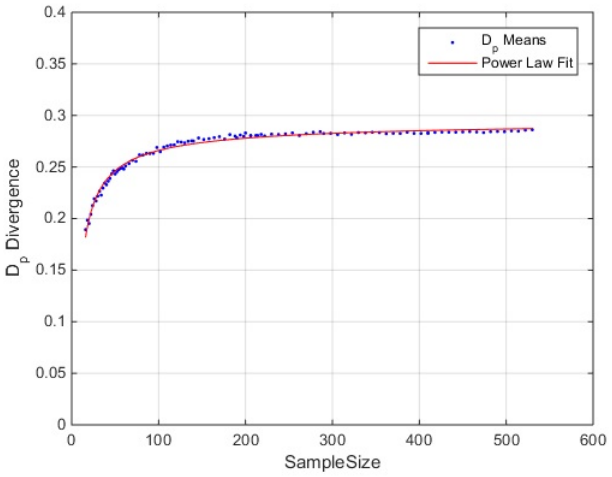


Table 3: Bayes Error Rates in Literature for Pima Indians Data Set [4]

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 4: 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 5: 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 (%), (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

[!h] [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
Contains the pima indian dataset BERs in table format

[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.May20
**Create Citation

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

[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