Bootstrap Estimation of a Non-Parametric Information Divergence Measure

Prad Kadambi and Visar Berisha
Arizona State University
School 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. Parametric divergence measures are functions of an unknown parameter θ, and describe the information contained in the data about θ [18]. Nonparametric divergence measures, notably including f-divergences such as the Kullback-Leibler divergence, measure the difference between two distributions P and Q.

Normally, in 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_f, is of the form:

$$D_f = \int_\Omega \phi(\frac{dP}{dQ})dP \tag{1}$$

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

A naive method to calculate the divergence between the i.i.d. data is to first find the densities for X_i and Y_i, and then calculate the divergence from the computed density estimates.

A common problem in machine learning is binary classification, in which data $x_i \in \mathbf{R^n}$ are assigned a class label $c_i \in \{0, 1\}$, with the aim of minimizing errors in class assignment. Given class labels c_0 and c_1 corresponding to respective probability distributions $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$, prior probabilities p and q, the Bayesian classifier assigns class labels to x_i such that the posterior probability is maximized. 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, $p(\mathbf{x})$ and $q(\mathbf{x})$, it is possible to write the Bayes error rate in terms of the prior probabilities p and q as given in [2]:

$$E_{Bayes} = \int_{r_1} pf_0(\mathbf{x})\, d\mathbf{x} + \int_{r_0} qf_1(\mathbf{x})\, d\mathbf{x} \tag{2}$$

Here, r_1 and r_0 refer to the regions where the respective posterior probabilities p_1 and p_0 are larger.

Direct evaluation of this integral can be quite involved and impractical, as it is challenging to create an exact model for the posterior 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 from f-divergences [5]. However, in estimating f-divergences, many of the same problems with estimating BER are encountered. Namely, $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ are unknown,

Equally important to estimating the divergence measures, is obtaining a metric of estimator quality, such as variance or 95% confidence interval. Before the estimator is applied in data analysis, knowledge of its approximate sampling distribution is critical in quantifying its usefulness. We analyze a nonparametric, asymptotically consistent divergence measure and apply it to the binary classification task.

We arrive at an estimate of the Bayes error rate by using expressions that give bounds on the classification error in terms of information divergence measures. However, common methods of estimating the Bayes error rate via divergence measure still require information about the conditional distributions corresponding to both class labels. Therefore the nonparametric divergence measure given in [3] will be used in conjunction with the Bayes error estimates derived for this divergence measure in [2] to conduct the analysis.

The work is organized as follows: the remainder of Section 1 is devoted to previous work, Section 2 provides a description of the divergence measure used, and its relation to the Bayes error rate, and Section 3 introduces the bootstrap sampling method, and the power law used to estimate D_p. In Section 4 we will apply the method to several generated datasets and real world datasets to show that the power law method can successfully be applied to several distributions. In 4.1 we will consider the generated example datasets, and in 4.2 we will perform our analysis on the Pima Indians dataset and the Banknote dataset found in the University of California, Irvine machine learning repository [6].

Background

1.1 The D_p Divergence Measure

The divergence measure introduced in [2]

1.2 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.

Because bounds on the Bayes Error Rate can be calculated directly from D_p, 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.

Bootstrap resampling, first introduced by Efron in [10], is a powerful method to find the sampling distribution of an estimator. Given a data set of size N, bootstrap straIn order to characterize the

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

3 Results

Uniform Dataset

Table 1: Uniform Dataset for Bootstrap Analysis of D_p

	D_0								
	μ_0	0	0	0	0	0	0	0	0
	σ_0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
e32									
	D_1								
	μ_1	$\frac{1}{2}$	0	0	0	0	0	0	0
	σ_1	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

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

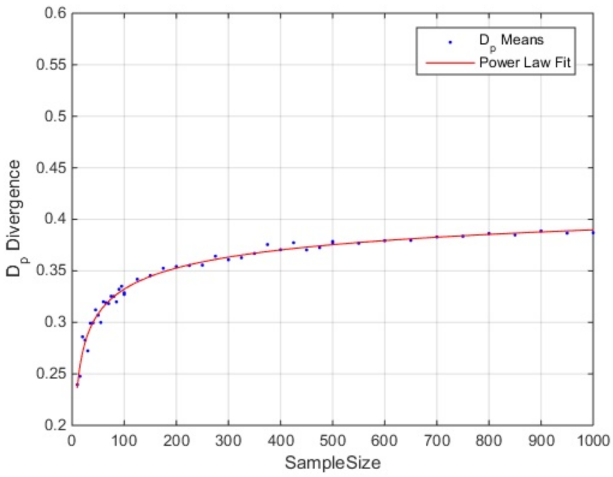
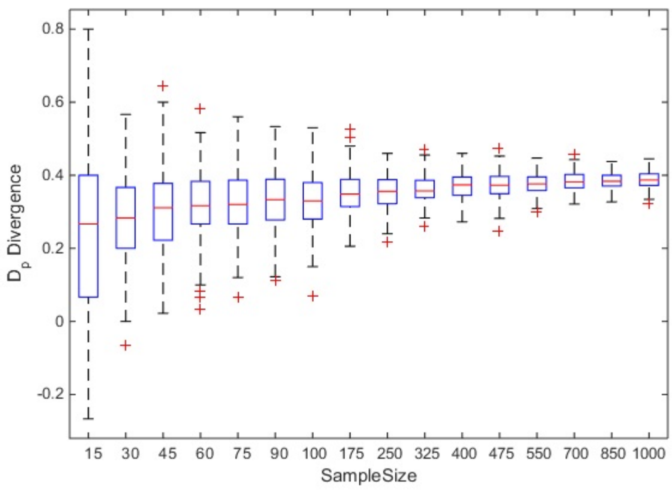


Figure 2: Distribution of D_p Values for 8-Dimensional Uniform Data Set, N = 200 trials



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

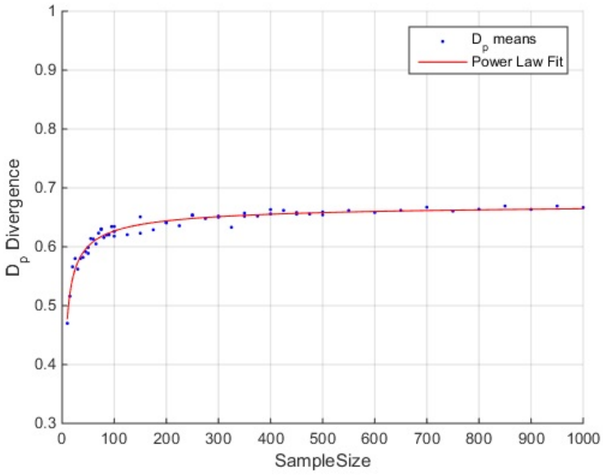
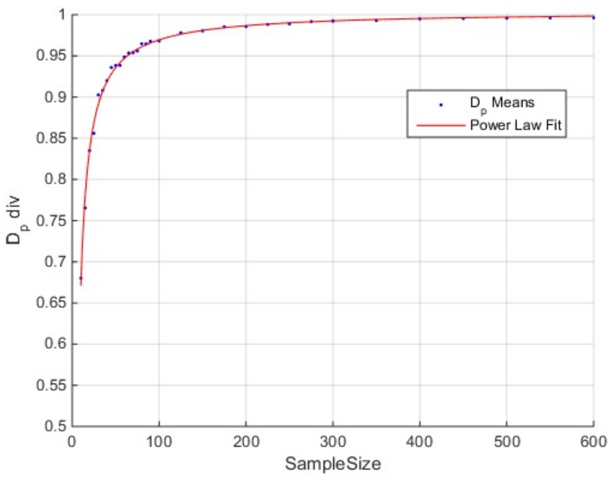


Figure 4: Convergence of D_p for Banknote Authentication Data Set, N = 50 trials



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.

Pima Indians Dataset

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

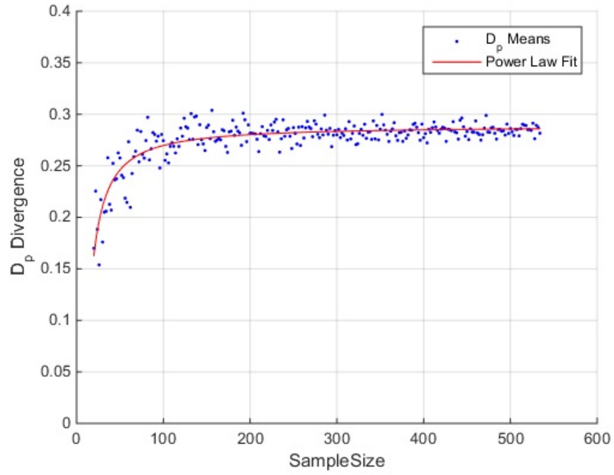


Figure 6: Asymptotic Convergence for Pima Indian Data Set, N = 200 trials

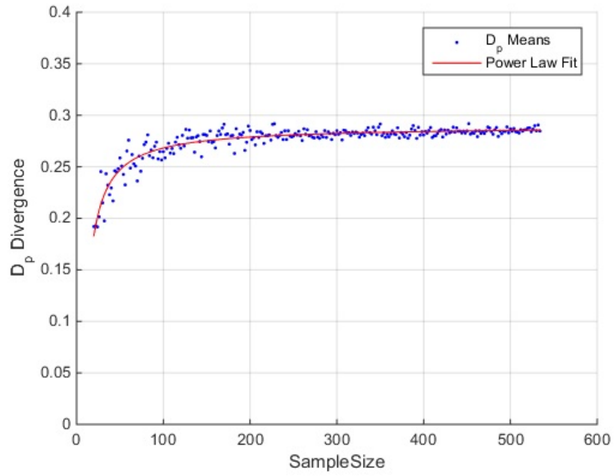


Figure 7: Asymptotic Convergence for Pima Indian Data Set, N = 5000 trials

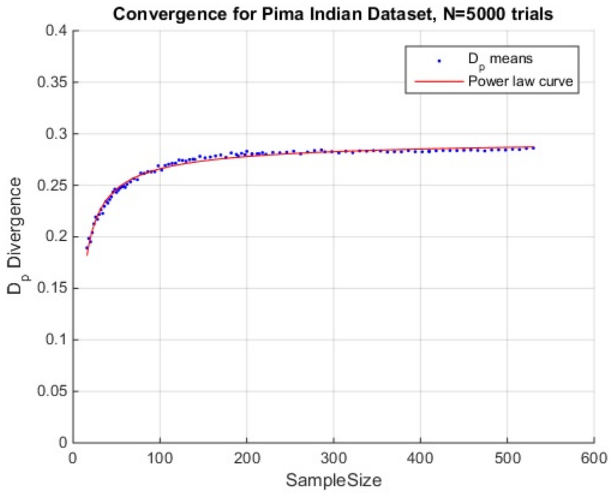


Table 3: Bayes Error Rates in Literature for Pima Indians Data Set [4]

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)