Bootstrap Estimation of a Non-Parametric Information Divergence Measure

Arizona State University

Abstract

This work details the bootstrap estimation of a non-parametric information divergence measure - the D_p divergence measure - in the context of a binary classification problem. In practice only finite length data sets are available for analysis. To reduce the limitation of finite data size, a bootstrap approach is used to calculate the divergence measure. Monte-Carlo iterations are performed at various sample sizes of the data set, the D_p value is found for each value of sample size, and a power law curve is used to find the asymptotic convergence value of the D_p divergence measure as a function of sample size. The divergence measure can then be used to estimate the Bayes classification error rate. This method is applied to several data sets, and the Bayes error rate is calculated from the D_p divergence.

Introduction

Information divergence measures have a wide variety of applications in machine learning, pattern recognition, statistics, and information theory. A common problem in machine learning is the binary classification problem, in which data $x_i \in \mathbf{R^n}$ is assigned a class label $c_i \in \{0, 1\}$ according to a classification rule, where class labels c_0 and c_1 correspond to respective probability distributions $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$. The Baysian classifier assigns class labels to x_i such that the posterior probability is maximized. The error rate of this classifier, the Bayes error rate, provides an absolute lower bound on the classification error rate. Estimating the best achievable classification error rate makes it possible to quantify the usefulness of a feature set or the performance of a classifier [1]. Given the two conditional distributions, $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$, it is possible to write the Bayes error rate in terms of the prior probabilities p_0 and p_1 as given in [2]:

$$E_{Bayes} = \int_{r_1} p_0 f_0(\mathbf{x})\, dx + \int_{r_0} p_1 f_1(\mathbf{x})\, dx \tag{1}$$

Here, r_1 and r_0 refer to the region where the corresponding posterior is the larger.[5]

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, it is possible to derive bounds for the Bayes error rate.
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 non-parametric divergence measure given in [3], and the Bayes error estimates derived in [2] will be used in this study.
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 used. In Section 4 we will apply the method to several generated datasets and real world datasets. 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 [cite repository here].

Previous Work

Background

The D_p Divergence Measure

Bootstrap Sampling

-lowest sub sample size has to be greater than the dimension of the data

Examples

Uniform Dataset

Table 1: Uniform Dataset for Which D_p Value is Found via Bootstrap Method and Analytically Verified

D_0								
μ_0	0	0	0	0	0	0	0	0
σ_0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

D_1								
μ_0	$\frac{1}{2}$	0	0	0	0	0	0	0
σ_0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

Pima Indians Dataset

Table 2: D_p

Sample Size	Monte-Carlo Iterations	D_p Asymptotic Value	Bayes Error Rate (Lower Bound)
100	50	0.2816	23.47%
100	200	0.2752	23.77%
100	5000	0.3170	21.85%
200	50	0.2829	23.40%
200	200	0.2925	22.96%
200	5000	0.3176	23.40%
300	50	0.2970	21.82%
300	200	0.3030	22.48%
300	5000	0.3033	22.46%

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

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.

Figure 1: Asymptotic Convergence for Pima Indian Data Set, N = 5000 trials

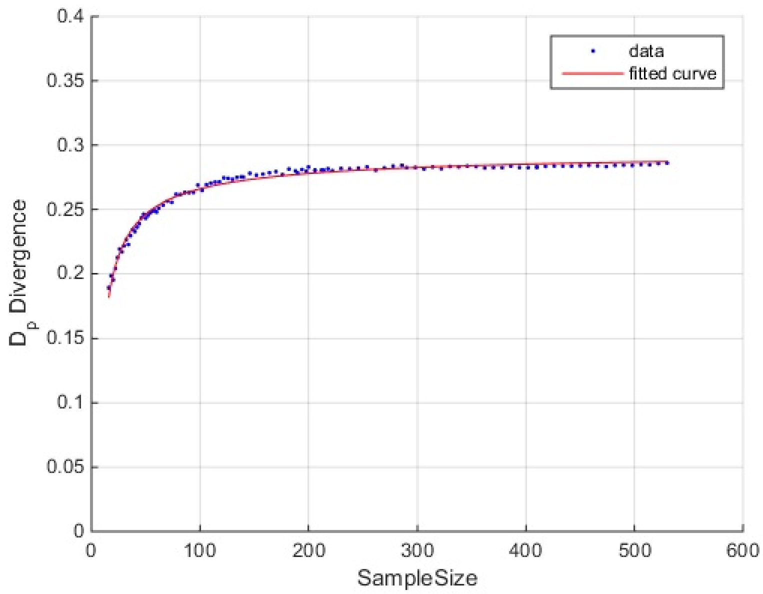

Figure 2: Asymptotic Convergence for Pima Indian Data Set, N = 5000 trials

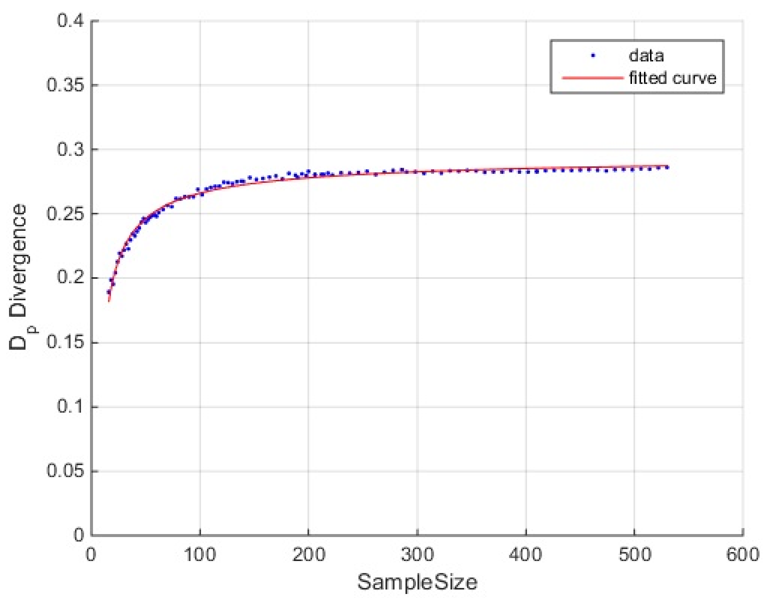

Figure 3: Asymptotic Convergence for Pima Indian Data Set, N = 5000 trials