ISSN: 2320-1363

A Novel Nonparametric Plugin-Entropy Estimator

R.Divakara Sarma*[1], Dr. T.Bhaskara Reddy*[2]

Research Scholar, Dept. of Computer Science & Technology, S.K. University, Anantapur
Associate Professor, Dept. of Computer Science & Technology, S.K. University, Anantapur

Abstract

The problem of estimation of entropy functional of probability densities were on lime light in the machine learning, information theory and statistics communities. Kernel density plug-in estimators are functions that are simple in design, easy to implement and widely used for estimation of entropy. This paper proposes a plugin-entropy estimator which is based on renyi entropy and discusses the merits.

Key words: Entropy, Non-parametric, Shanon Entropy, Renyi Entropy, Parzen, Kernel Density Estimation, Plugin Entropy.

Introduction

The analysis of distributions is fundamental in machine learning and statistics. Algorithms in these fields rely on information theoretic approaches like entropy, mutual information and Kullback–Leibler divergence.

Non-linear entropy functionals arise in applications like machine learning, mathematical statistics, and statistical communication theories. Important examples include Shannon and Renyi entropies. Entropy based applications include image registration and texture classification, ICA, anomaly detection, data and image compression, testing of statistical models and parameter estimation.

In these applications, the functional of interest must be estimated empirically from sample realizations of the densities underlying. Several estimators of entropy measures have been proposed for general multivariate densities. The list include consistent estimators based on histograms, kernel density plug-in estimators, entropic graphs, gap estimators and nearest neighbor distances. Kernel density plug-in estimators are simple, easy to implement, computationally fast and therefore widely used for estimation of entropy. However, these estimators suffer from mean squared error (MSE) rates which typically grow with feature dimension d as $O(T^{-}/d)$, where T is the number of samples and γ is a positive rate parameter.

In non-parametric statistics, a **kernel** is a weighting function used in non-parametric estimation techniques. A generalization of the individual data-point feature mapping done in classical kernel

ISSN: 2320-1363

methods.

Kernels are used in kernel density estimation to estimate random variables density functions, or in kernel regression to estimate the conditional expectation of a random variable. Kernels are also used in time-series, in the use of the periodogram to estimate the spectral density where they are known as window functions. An additional use is in the estimation of a time-varying intensity for a point process where window functions (kernels) are convolved with time-series data.

In functional analysis (a branch of mathematics), a **reproducing kernel Hilbert space** (RKHS) is a Hilbert space associated with a kernel that reproduces every function in the space or every evaluation functional is bounded in RKHS. The RKHS reproducing kernel was first introduced in the 1907 work of Stanisław Zaremba concerning boundary value problems for harmonic and biharmonic functions. James Mercer also at the same time examined functions that satisfy the reproducing property in the theory of integral equations. The idea of the reproducing kernel remained unexplored for nearly twenty years The subject was eventually systematically developed in the early 1950s by Nachman Aronszajn and Stefan Bergman.

The kernel of a reproducing kernel Hilbert space is used in the suite of techniques known as kernel methods to perform tasks such as statistical classification, regression analysis, and cluster analysis on data in an implicit space. This usage is particularly common in machine learning.

These spaces have wide applications, including complex analysis, harmonic analysis, and quantum mechanics. Reproducing kernel Hilbert spaces are particularly important in the field of statistical learning theory because of the celebrated Representer theorem which states that every function in an RKHS can be written as a linear combination of the kernel function evaluated at the training points. This is a practically useful result as it effectively simplifies the empirical risk minimization problem from an infinite dimensional to a finite dimensional optimization problem.

Density estimators

The plug-in estimators defined above can be obtained by using any density estimator. commonly, kernel widths must also be specified when running a non-parametric estimation. Some popular examples are the following:

1. Kernel Density Estimator (KDE)

Kernel density estimation takes the approach of estimating density at a given point using a kernel K with bandwidth parameter h to form a weighted average using other points from the sample. Intuitively, the points that are closer to the point whose density is being estimated will have a higher contribution to the density than points that are further away. The selection of the kernel and the bandwidth parameter adjust the characteristics of this relationship.

ISSN: 2320-1363

2. KNN Estimator

KNN density estimation at a point x follows by obtaining n samples from a distribution and computing the volume V needed to encapsulate k nearest points to x and then taking the ratio

The mean square error of the Kernel Density Estimator decomposes into a Bias and a Variance term:

The bandwidth parameter may be selected to balance the bias-variance tradeoff . One strategy would be to set the derivative of the sum equal to 0 and solve.

Density Estimation:

The embedding of distributions into infinite-dimensional feature spaces can preserve all of the statistical features of arbitrary distributions, while allowing one to compare and manipulate distributions using operations such as inner products, distances, projections, linear transformations, and spectral analysis. This learning framework is very general and can be applied to distributions over any space on which a sensible kernel function (measuring similarity between elements) may be defined. However, to estimate these quantities, one must first either perform density estimation, or employ sophisticated space-partitioning/bias-correction strategies which are typically infeasible for high-dimensional data.

Commonly, methods for modeling complex distributions rely on parametric assumptions that may be unfounded or computationally challenging (e.g. Gaussian mixture models), while nonparametric methods like kernel density estimation or characteristic function representation break down in high-dimensional settings.

Advantages:

Methods based on the kernel embedding of distributions sidestep these problems and also possess the following advantages:

- Data may be modeled without restrictive assumptions about the form of the distributions and relationships between variables
- Intermediate density estimation is not needed
- Practitioners may specify the properties of a distribution most relevant for their problem (incorporating prior knowledge via choice of the kernel)
- If a *characteristic* kernel is used, then the embedding can uniquely preserve all information about a distribution, while thanks to the kernel trick, computations on the potentially infinite-dimensional RKHS can be implemented in practice as simple Gram matrix operations
- Dimensionality-independent rates of convergence for the empirical kernel mean (estimated using samples from the distribution) to the kernel embedding of the true underlying distribution can be proven.
- Learning algorithms based on this framework exhibit good generalization ability and finite sample convergence, while often being simpler and more effective

ISSN: 2320-1363

than information theoretic methods

Learning via the kernel embedding of distributions offers a principled drop-in replacement for information theoretic approaches and is a framework which not only subsumes many popular methods in machine learning and statistics as special cases, but also can lead to entirely new learning algorithms.

Entropy :

Entropy is a measure of *unpredictability* or *information content*. Entropies quantify the diversity, uncertainty, or randomness of a system. Ross Quinlan developed an algorithm based on Hunt's theory called **ID3 (Interactive Dichotomizer 3)**, in which he used **Shannon's entropy** as a criterion for selecting the most significant / discriminatory feature:

The Shannon Entropy is given by

$$Entropy\ (S) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

The **Rényi entropy** generalizes the Shannon entropy. The Rényi entropy is named after Alfred Renyi and is defined as

The Rényi entropy of order α, where $\alpha \geq 0$and $\alpha \neq 1$, is defined as

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{n} p_i^\alpha \right)$$

Here, $X$is a discrete random variable with possible outcomes $1, 2, ..., n$ and corresponding probabilities $p_i \doteq \Pr(X = i)$for $i = 1, \ldots, n$, and the logarithm is base 2. If the probabilities are $p_i = 1/n$for all $i = 1, \ldots, n$, then all the Rényi entropies of the distribution are equal: $H_\alpha(X) = \log n$. In general, for all discrete random variables X, $H_\alpha(X)$is a non-increasing function in α.

Characteristics:

As α approaches zero, the Rényi entropy increasingly weighs all possible events more equally, regardless of their probabilities. In the limit for, the Rényi entropy is just the logarithm of the size of the support of X. The limit for equals the Shannon entropy, which has special properties. As approaches infinity, the Rényi entropy is increasingly determined by the events of highest probability.

Non parametric Entropy Estimators

The primary variable parameters of the entropy estimator that can be user defined or are a function of the problem at hand are the following

1. *N*: The number of data samples or exemplars

2. *σ*: The kernel size for the Parzen PDF estimate

3. *M*: The dimension of the dataset

4. *d*: A measure of the extent (or variance) of the data

The kernel(Parzen) estimate of the

ISSN: 2320-1363

Probability Density Function using an arbitrary kernel function $\kappa\sigma(.)$ is given by

$$\hat{p}(x) = \frac{1}{N\sigma} \sum_{i=1}^{N} k \left(\frac{x - x_i}{\sigma}\right)$$

Where σ is the kernel size or bandwidth parameter

Plug-In Estimates for Entropy

The plug-in entropy estimates are obtained by simply inserting a consistent density estimator of the data in place of the actual PDF in the entropy expression

Four types of approaches could be followed when using a plug-in estimate.

- ***integral estimates***, evaluates exactly or approximately the infinite integral existing in the entropy definition.
- ***resubstitution estimates***, further includes the approximation of the expectation operator in the entropy definition with the sample mean.
- ***splitting data estimate***, and is similar to the resubstitution estimate, except that now the sample set is divided into two parts and one is used for density estimation and the other part is used for the sample mean.
- ***cross-validation estimate,*** uses a leave–one–out principle in the resubstitution estimate. The entropy estimate is obtained by averaging the leave-one-out resubstitution estimates of the dataset.

The key difference between the Re-substitution estimate and the Splitting Data Estimate is that the splitting estimate sums over different samples than the ones used for estimating the density

Non parametric Estimator for α-Renyi's Entropy was proposed by J.C. Principe as

$$\hat{H}_{\alpha}(X) = \frac{1}{1-\alpha} \log \frac{1}{N} \sum_{j=1}^{N} \left(\frac{1}{N} \sum_{i=1}^{N} k_{\sigma}(x_j - x_i) \right)^{\alpha - 1}$$

Where K_{σ} is the kernel function.

This kernel function has to obey the following properties

1. $k(x) \geq 0$
2. $\int_R k(x)dx \lim_{x \to \infty} |xk(x)| = 0$
3.

motivated by the above aspects, we designed a kernel which satisfies the conditions and will be able to provide better classification for some standard datasets. We present the performance of our kernel against the regular entropy based classifier (ID3). The UCI Machine Learning Repository Standard Datasets are taken for evaluation.

The Percentage of Correct Classification is as follows:

Table -1 Percentage Correct

Data Set	**NRC**	**ID3**
Monks-Prob -1- Test	49.07	5.09
Monks-Prob -1- Train	54.84	29.84
Monks-Prob -2- Test	49.07	10.88
Monks-Prob -2- Train	47.34	35.50
Monks-Prob -3- Test	49.07	4.86

ISSN: 2320-1363

Monks-Prob -3- Train	51.64	31.97
Nursery	97.58	98.19
Solar-Flare -1	97.83	95.36
Solar-Flare -2	98.97	98.69
Spect-Test	69.52	57.22
Spect-Train	63.75	52.50

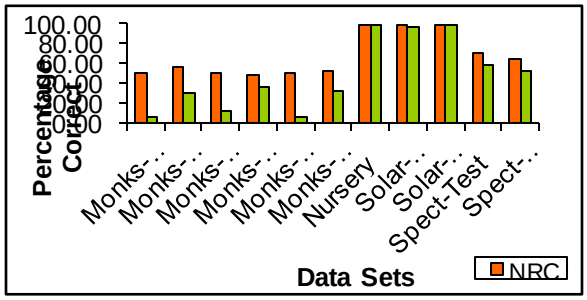

Fig 1. Percentage Correct Proposed Classifier-1 (NRC) vs ID3

Table- 2 Mean Absolute Error

Data Set	NRC	ID3
Monks-Prob -1- Test	0.5000	0.9431
Monks-Prob -1- Train	0.4955	0.6705
Monks-Prob -2- Test	0.5000	0.8456
Monks-Prob -2- Train	0.5002	0.5884
Monks-Prob -3- Test	0.5000	0.9484
Monks-Prob -3- Train	0.4996	0.6455
Nursery	0.0057	0.0018
Solar-Flare -1	0.0424	0.0405
Solar-Flare -2	0.0056	0.0050
Spect-Test	0.3777	0.4228
Spect-Train	0.4355	0.5047

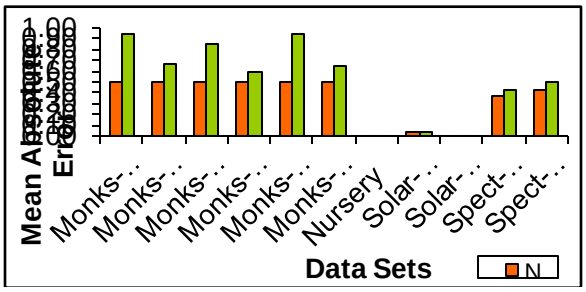

Fig 2. Mean Absolute Error: Proposed Classifier-1 (NRC) vs ID3

Conclusion:

For all data sets the Decision tree constructed with our proposed Non-parametric kernel based entropy performs much better than regular Decision tree classifiers based on shanon entropy. Out of the 11 datasets we tested in 10 data sets our classifier was worked better than the ID3 in percentage of correct classification. The Mean Absolute Error was also very low compared to ID3 in 8 cases out of 11 datasets.

References:

1. Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81-106
2. Rényi, Alfréd (1961). "On measures of information and entropy" . Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability 1960. pp. 547–561.

ISSN: 2320-1363

3. "Exploratory Data Analysis: Kernel Density Estimation " https://chemicalstatistician.wordpress.com/2013/06/09/exploratory-data-analysis-kernel-density-estimation-in-r-on-ozone-pollution-data-in-new-york-and-ozonopolis/
4. An introduction to information theory and entropy - Tom Carter astarte.csustan.edu/~tom/SFI-CSSS/info-theory/info-lec.pdf
5. Parzen-Window Density Estimation https://www.cs.utah.edu/~suyash/Dissertation_html/node11.html
6. "Plugin entropy estimators " http://rpackages.ianhowson.com/cran/entropy/man/entropy.plugin.html
7. N. Leonenko, L. Prozanto, and V. Savani. A class of Renyi information estimators for multidimensional densities. Annals of Statistics, 36:2153–2182, 2008.
8. E. Gin´e and D.M. Mason. Uniform in bandwidth estimation of integral functionals of the density function. candinavian Journal of Statistics, 35:739761, 2008.
9. M. Goria, N. Leonenko, V. Mergel, and P. L. Novi Inverardi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. Nonparametric Statistics, 2004.
10. V. Vapnik. Statistical Learning Theory. Wiley-Interscience, 1998.
11. Classification Algorithms for Codes and Designs, Petteri Kaski, Patric R.J. Östergård, Algorithms and Computation in Mathematics Volume 15, Springer, 2006.
12. Data Mining: Concepts, Models, Methods, and Algorithms by Mehmed Kantardzic , John Wiley & Sons,2003

Authors

Dr.T.Bhaskara Reddy is an Associate Professor in the department of Computer Science and Technology at S.K.University, Anantapur A.P. He currently holds the post of Head of the Department of Computer Science Technolgy. He was served the university in different cadres as Deputy Director of Distance Education, CSE Co-ordinator of Engineering at S.K.University College of Engineering & Technology. He holds M.Sc and Ph.D in computer science. He has acquired M.Tech from Acharya Nagarjuna University. He has been continuously imparting his knowledge to several students from the last 18 years. He has published 62 National and International publications. He has completed one major research project (UGC). Seven Ph.D and Three M.Phil have been awarded under his guidance. His research interests are in the field of image Processing, computer networks, data mining and data warehouse, robotics. Email:bhaskarreddy_sku@yahoo.co.in

Mr.R.Divakara Sarma is an research scholor in the department of Computer Science and Technology at S.K University, Anantapur A.P. He has completed M.Sc. in Mathematics from S.K. University and M,Tech in Computer Science from A.N. University, Guntur. M.Phil in Computer Science and Technology from S.K. University.