Ensembles of Nonparametric Entropy Estimators

Vy Nguyen, Majeed Thaika
andrewID: vyn, mthaika
36/10-702: Progress Report

1 Introduction

Information entropy is the average amount of information produced by a random variable. Shannon (1949) defines the differential entropy H of a random variable X and probability mass function $p(x)$ as

$$H(X) = -\int p(x) \log p(x) dx$$

Since the true probability is not known, it is not possible to calculate $H(X)$ directly.

Entropy estimation has many important applications. For example, it can be used to estimate the mutual information of two random variables and provide insights about their relationship. Furthermore, information entropy has other applications in encoding data, data compression, clustering, and a criterion for feature-splitting in decision trees.

Entropy estimation is difficult because it requires estimating the non-smooth function $f(x) = -x \log(x)$, that is not differentiable at $x = 0$. One approach is to use the naive plugin estimator from the empirical distribution and get

$$\hat{H}(X) = -\sum_{i=1}^{n} \hat{p}(x_i) \text{log} \hat{p}(x_i)$$

where $\hat{p}(x_i) = \dfrac{h_i}{n}$ is the MLE of each probability $p(x_i)$ and $h_i = \sum_{k=1}^{n} \mathbb{I}(X_k = i)$ is the histogram over the outcomes.

However, Basharin (1959) and Harris (1975) have shown that the naive plugin estimator always underestimates the true entropy. Another result from Paninski (2003) proves that there exists no unbiased estimator for entropy. Furthermore, many of the existing estimators suffer from the curse of dimensionality and converge slowly at the rate of $O(T^{-\gamma/d})$ where T is the number of samples and γ is a positive rate parameter.

In this progress report, we will summarize the results of some well-known entropy estimators as well as a weighted ensemble method that, under the right conditions, can remove the dependency on the dimension and ensure a convergence rate of $O(T^{-1})$.

2 Entropy Estimation Overview

Beirlant et al. (2001) gives an overview of several methods in use for the nonparametric estimation of entropy. Besides the MLE plug-in estimator, there are other plug-in estimators that utilize techniques such as kernel density estimation, splitting data, or cross-validation.

Consider $X_1, ...X_n$ and an estimator of the form $H_n = -\dfrac{1}{n}\sum_{i=1}^{n} \log f_n(X_i)$ where f_n is a kernel density estimator. Joe (1989) shows that the MSE converges at the rate $O(n^{-1}) + O(n^{-2}h^{8-d}) + O(n^{-2}h^{-d}) + O(n^{-1}h^{8-d})+)(n^{-2}h^{4-2d}) + O(h^8)$ where n is the sample size, h is the binwidth, and d is the dimension. The analysis shows that the sample size needed for good estimators increases rapidly with the dimension of the multivariate density.

Another plug-in estimator is the splitting-data estimator. Start by decomposing the sample into two sets: $X_1, ..., X_i$ and $X_{i+1*}, ..., X_{n*}$. Using the first set of data, construct a density estimate f_i and then, using f_i

and the second set, construct $H_n = -\frac{1}{n-(i+1)} \sum_{j=i+1}^{n} \mathbb{I}_{X_{j*}} \log f_j(X_j*)$. For f_j being the histogram density estimate, kernel density estimate, and any L_1-consistent density estimator, Gyorfi and van der Meulen (1987, 1989, 1990) show that the estimator H_n converges almost surely to $H(f)$ under some mild tail and smoothness condition.

The final plug-in estimator is based on cross-validation. Let $f_{n,i}$ denotes the kernel density estimator based on $X_1, ..., X_n$ leaving X_i out, then the corresponding density estimator is $H_n = -\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i} \log f_{n,i}(X_i)$.

Another interesting technique to estimate entropy is by using the nearest neighbor distances. Let $\rho_{n,i}$ be the nearest neighbor distance of X_i and the other $X_j : \rho_{n,i} = \min_{j \neq i, j \leq n} ||X_i - X_j||$. The nearest neighbor estimate is $H_n = \frac{1}{n} \sum_{i=1}^{n} \ln(n\rho_{n,i}) + \ln(2) + C_E$ where the Euler constant $C_E = -\int_0^\infty e^{-t} \ln(t) dt$.

3 Weighted Ensemble Estimator

Let T be the total sample size. Sricharan et al. (2013) propose a weighted density functionals estimator that takes the weighted affine combination of an ensemble of slow functional estimators with MSE decay rate of order $O(T^{-\gamma/d})$ and converges at a much faster dimension-invariant rate of $O(1/T)$. They apply this weighted estimator, with an ensemble of truncated uniform kernel density estimators, to the problem of Shannon entropy estimation and show that it performs better than many well-known plugin estimators, especially in higher dimensions. By the final report, we aim to replicate their findings and also experiment with using different ensemble of entropy estimators.

We define our ensemble estimator as $\hat{\mathbf{E}}_w = \sum_{l \in \bar{l}} w(l) \hat{\mathbf{E}}_l$, where $\bar{l}$ denotes the set of parameters and $\hat{\mathbf{E}}_l$ are the L parameterized, unbiased estimators. We also add the weights constraint that $\sum_{l \in \bar{l}} w(l) = 1$, which guarantees that the weighted estimator will be asymptotically unbiased. For each $\hat{\mathbf{E}}_l$, we assume that bias and variance is given by:

$$\mathbb{B}(\hat{\mathbf{E}}_l) = \sum_{i \in [I]} c_i \psi_i(l) T^{-i/2d} + O\left(\frac{1}{\sqrt{T}}\right)$$

$$\mathbb{V}(\hat{\mathbf{E}}_t) = c_v \left(\frac{1}{T}\right) + o\left(\frac{1}{T}\right)$$

where c_i and c_v are constants, I is a finite index set with cardinality $I < L$, and $\psi_i(l)$ are basis functions which only depend on l. With above bias and variance conditions, a weight vector w_0 can be found by solving a convex optimization problem such that $\mathbb{E}[(\hat{\mathbf{E}}_{w_o} - E)^2] = O(1/T)$. The optimization problem is defined as:

$$\begin{aligned} \min_w \quad & ||w||_2 \\ \text{subject to } \sum_{l \in \bar{l}} w(l) \quad & = \quad 1, \\ \gamma_w(i) \quad & = \quad \sum_{l \in \bar{l}} w_l \psi_i(l) = 0, i \in [I], \end{aligned}$$

From this setting, the bias and variance of the weighted estimator can be calculated as:

$$\begin{aligned} \mathbb{B}(\hat{\mathbf{E}}_w) \quad & = \quad \sum_{i \in I} c_i \gamma_w(i) T^{-i/2d} + O\left(\frac{||w||_1}{\sqrt{T}}\right) \\ & = \quad \sum_{i \in I} c_i \gamma_w(i) T^{-i/2d} + O\left(\frac{\sqrt{L}||w||_2}{\sqrt{T}}\right) \end{aligned}$$

Denote the covariance matrix $\{\hat{\mathbf{E}}_l; l \in \bar{l}\}$ by Σ_L. Let $\bar{\Sigma}_L = \Sigma_L T$,

$$\mathbb{V}(\hat{\mathbf{E}}_w) = \mathbb{V}\left(\sum_{l\in\bar{l}} w_l \hat{\mathbf{E}}_l\right) = w^T \Sigma_L w = \frac{w^T \bar{\Sigma}_L w}{T} \leq \frac{L||w||_2^2}{T}$$

We can also rewrite the earlier convex optimization constraints in matrix form as $\min_w ||w||_2^2$ subject to $A_0 w = b$, where b is vector of zeros with a leading one and A_0 is the basis projection matrix—where first row is a vector of ones and the i^{th} row $(A_0)_i = [\psi_i(l_1), ..., \psi_i(l_L)]$. Also, let A_1 be the A_0 matrix without the first row of ones, then solving for the minimum square norm $\eta_L(d) := ||w_0||_2^2$ is given by $\eta_L(d) = \dfrac{\det(A_1 A_1')}{\det(A_0 A_0')}$ where $\eta_L(d)$ is independent of T given fixed number of estimators L and fixed dimension d, i.e. $\eta_L(d) = \Theta(1)$. The bias and variance of the weighted estimator is therefore:

$$\mathbb{B}[\hat{\mathbf{E}}_{w_0}] = O\left(\sqrt{L\eta_L(d)/T}\right) = O\left(1/\sqrt{T}\right)$$

$$\mathbb{V}[\hat{\mathbf{E}}_{w_0}] = O(L\eta_L(d)/T) = O\left(1/T\right)$$

Thus, the overall MSE rate converges in dimension invariant rate of $O\left(1/T\right)$.

4 Truncated Uniform Kernel Density Estimator

We can now estimate any general d-dimensional, non-linear density functionals of the form $G(f) = \int g(f(x), x) f(x) dx$.

Let $T = N + M$ and represent the i.i.d observations from f as $\{X_1, ..., X_N, X_{N+1}, ..., X_{N+M}\}$. We have shown an ensemble estimator that will converge much faster than their slower estimator components, if they match the bias and variance conditions. We present one such plugin estimator called the truncated uniform kernel density estimator:

$$\hat{f}_k(X) = \frac{\sum_{i=1}^{M} \mathbb{I}_{\{X_i \in S_k(X)\}}}{M V_k(X)}$$

where $k \leq M$ is a real number, $S_k(X)$ is the truncated kernel region for each X in finite support $[a, b]^d$, and V_k is the volume of the truncated uniform kernel under $S_k(X)$.

Given that we form the estimate $\hat{f}_k$ at N points using M observations, the plug-in estimator is given by:

$$\hat{G}_k = \frac{1}{N} \sum_{i=1}^{N} g(\hat{f}_k(X_i), X_i)$$

where $\hat{G}_k$ is identical to the standard kernel density estimator G_k' except for the volume—which in KDE is set to the untruncated value $V_k(X) = k/M$. It can be shown that the biases of the plug-in estimators $\hat{G}_k, G_k'$ are given by

$$\mathbb{B}(\hat{G}_k) = \sum_{i=1}^{d} c_{1,i} \left(\frac{k}{M}\right)^{i/d} + \frac{c_2}{k} + o\left(\frac{1}{k}\right) + \frac{k}{M}$$

$$\mathbb{B}(G_k') = c_1 \left(\frac{k}{M}\right)^{1/d} + \frac{c_2}{k} + o\left(\frac{1}{k} + \frac{k}{M}\right)$$

The variances of the plug-in estimators $\hat{G}_k, G_k'$ are identical up to leading terms and are given by

$$\mathbb{V}(\hat{G}_k) = \mathbb{V}(G_k') = c_4 \left(\frac{1}{N}\right) + c_5 \left(\frac{1}{M}\right) + o\left(\frac{1}{M} + \frac{1}{N}\right)$$

where c_4 and c_5 are constants that depend on g and f. We also need the extra assumptions that $k \to \infty$ and $k/M \to 0$ to ensure that estimators are unbiased. We can see that as $N \to \infty$ and $M \to \infty$, variance converges to 0.

To minimize the bias in both plugin estimators, the optimal choice of $k = \Theta\left(M^{\frac{1}{1+d}}\right)$, which gives us a bias of $\Theta\left(M^{\frac{-1}{1+d}}\right)$. Since $N, M \leq T$, both KDE and truncated KDE have the assumed bias and variance rate necessary for the weighted estimator to achieve parametric MSE rate of convergence.

5 Work Division

We both read the papers mentioned above and discussed them during our meetings. The composition of the progress report was also divided equally; Vy wrote the first half of the report and Majeed finished the remaining half. For our future work, we anticipate to maintain the same work dynamic. Specifically, we will each finish the remaining reading list and then meet to discuss the results. For the simulation, we will consult each other to write the weighted ensemble program and apply it to other estimators besides the uniform kernel density estimator. Lastly, we also expect to divide the writing assignment equally for the final report.

6 Remaining Work

We will look into the derivation of the minimax rate for estimating entropy, which is $O\left(n^{-min\{\frac{8\beta}{4\beta+d},1\}}\right)$ as shown by Birge and Massart (1995). We will also analyze more plugin entropy estimators, such as the k-NN and intrinsic dimension estimators, and see whether an ensemble of these can beat the performance of the weighted estimator that uses only truncated KDEs.

7 References

[1] J Beirlant, E J Dudewicz, L Gyorfi, and E C Van Der Meulen. Nonparametric entropy estimation: An overview (2001) http://jimbeck.caltech.edu/summerlectures/references/Entropy

[2] K Sricharan, D Wei, A Hero III. Ensemble estimators for multivariate entropy estimation (2013). https://arxiv.org/pdf/1203.5829.pdf

[3] Singh, Shashank; Poczos, Barnabas. Analysis of k-Nearest Neighbor Distances with Application to Entropy Estimation (2016). https://arxiv.org/pdf/1603.08578.pdf

[4] Birge, Lucien; Massart, Pascal. Estimation of Integral Functionals of a Density. Ann. Statist. 23 (1995), no. 1, 11–29. doi:10.1214/aos/1176324452. https://projecteuclid.org/euclid.aos/1176324452