Analysis of Nonparametric Entropy Estimators

Vy Nguyen, Majeed Thaika
andrewID: vyn, mthaika
36/10-702: Project Proposal

1 Problem Description

Information entropy is the average amount of information produced by a discrete random variable. Shannon (1949) defines the entropy H of a discrete random variable X and probability mass function $p(X)$ as $H(X) = \sum_{i=1}^{n} p(x_i) log p(x_i)$. 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 even as a criterion for decision tree feature splitting.

Entropy estimation is hard because it requires estimating the non-smooth function $f(x) = -x ln x$, that is not differentiable at $x = 0$. One approach is to use the naive plugin estimator from the empirical distribution and get $\hat{H}(X) = \sum_{i=1}^{n} \hat{p}(x_i) log \hat{p}(x_i)$ where $\hat{p}(x_i) = \frac{h_i}{n}$ 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. Under this context, the objective of this project is to further explore the problem of entropy estimation and study the different types of entropy estimators.

2 Scope of Work

By the progress report, we aim to summarize some key theoretical results in entropy estimation, including the minimax rate for estimating entropy, which is $\mathcal{O}((\frac{1}{n})^{min\{\frac{8\beta}{4\beta+d},1\}})$ as shown by Birge and Massart (1995). Also, despite there existing no unbiased entropy estimators, we will analyze selected entropy estimators that have provably low bias and/or low variance bounds, under certain assumptions, that work well in practice and are based on some algorithms we've covered previously/are familiar with: including histograms, k-NN graphs, and minimum spanning trees (we plan to finish this by the next progress report as well). Finally, we aim to evaluate the different algorithms ourselves using real-life and synthetic data.

3 Reading List

[1] Paninski, Liam. "Estimation of Entropy and Mutual Information." Neural Computation, vol. 15, no. 6 (2003) pp. 1191–1253. https://www.stat.berkeley.edu/ binyu/summer08/L2P2.pdf

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

[3] 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

[4] Kandasamy, Kirthevasan, Krishnamurthy, Akshay, Poczos, Barnabas, Wasserman, Larry, et al. Nonparametric von mises estimators for entropies, divergences and mutual informations (2015). https://pdfs.semanticscholar.org/69b5/f4b09cc9d399f6ccf41de104e214c6850d65.pdf

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