Estimating Mutual Information

December 15, 2017

One of the our greatest challenges to implementing Tishby's Information Bottleneck Method will be determining consistent, distribution invariant, high dimensional estimators of Mutual Information based on a training dataset. Outlined below are potential methods for estimating $I(X;Y)$ from $\{(x_i, y_i)\}$.

1. **Maximum Likelihood Mutual Information (MLMI):** Proposed by Suzuki et al. (2008) as a kernel based procedure that does not involve density estimation and can be equipped with cross-validation for efficient model selection. Derived from maximizing the likelihood of $p(x,y)$ under certain constraints.

$$\hat{I}(X;Y) = \frac{1}{n}\sum_{i=1}^{n} log\big(\hat{w}(x_i, y_i)\big) \qquad \hat{w}(x_i, y_i) = \alpha^\top \varphi(x,y)$$

   Where α and $\varphi(x,y)$ are chosen through cross validation (see *paper* for more details).

2. **Least Square Mutual Information (LSMI):** Proposed by Suzuki et al. (2008) as a method that uses least-squares for density ratio based mutual information estimation. It is a closely related method to MLMI (see *paper* for more details).

3. **Edgeworth Expansion Method (EDGE):** Proposed by Hulle et al. (2005) as an entropy approximation based on edgeworth expansion. It assumes that the data is from a d-dimensional normal distribution, then uses the entropy estimates to estimate the mutual information.

$$\hat{H} = H_{normal} - \frac{1}{12}\sum_{i=1}^{d} k_{i,i,i}^2 - \frac{1}{4}\sum_{j=1, i\neq j}^{d} k_{i,i,j}^2 - \frac{1}{72}\sum_{i,j,k=1, i<j<z}^{2} k_{i,j,k}^2$$

   Where H_{normal} is the entropy of a normal distribution with covariance matrix equal to the target distributions and $k_{i,j,k}$ is the standardized third cumulant (see *paper* for more details).

4. **K-Nearest Neighbor Method (KNN):** Proposed by Kraskov et al. (2004) as a K-Nearest Neighbor approach to estimating entropies such that the errors do not compound, but lacks a systematic strategy for choosing the value of k. Again the entropy estimates are uses to estimate the mutual information.

$$\hat{I}(X;Y) = \psi(k) = \psi(n) - \frac{1}{k} - \frac{1}{n}\sum_{i=1}^{n}[\psi(n_x(i)) + \psi(n_y(i))]$$

Where $z_i = (x_i, y_i)$, $||z_i|| = \max\{||x_i||, ||y_i||\}$, and $N_k(i)$ is the set of k-nearest neighbor samples of (x_i, y_i) with respect to the norm $||z_i||$. $n_x(i), n_y(i) \subset N_k(i)$ defined as:

$$\begin{aligned} \epsilon_x(i) &= \max\{||x_i - x_{i'}|| \mid (x_{i'}, y_{i'}) \in N_k(i)\} & n_x(i) &= |\{z_{i'} \mid ||x_i - x_{i'}|| \leq \epsilon_x(i)\}| \\ \epsilon_y(i) &= \max\{||y_i - y_{i'}|| \mid (x_{i'}, y_{i'}) \in N_k(i)\} & n_y(i) &= |\{z_{i'} \mid ||y_i - y_{i'}|| \leq \epsilon_y(i)\}| \end{aligned}$$

(see *paper* for more details).

5. **Kernel Density Estimator Method (KDE):** Involves distribution estimation of $p(x, y), p(x), p(y)$ through Kernel Density Estimation, then directly estimates the mutual information with the ratio of these distributions. However, division by estimated densities has been shown to expand and compound the error.

$$\hat{I}(X;Y) = \frac{1}{n}\sum_{i=1}^{n} log(\frac{\hat{p}_{xy}(x_i, y_i)}{\hat{p}_x(x_i)\hat{p}_y(y_i)})$$

Where $\hat{p}_{xy}, \hat{p}_x, \hat{p}_y$ are the KDE estimated distributions (see *paper* for more details).

6. **Variational Method (VAR):** This is a recent method that has had a lot of success in making IB based approaches practical. The basic idea of this method is to approximate the MI by fitting a variational lower bound to it. The advantage is that, the functional form then becomes simple and we can do SGD, which allows us to build objective functions and regularizers inspired by mutual information (see *paper* for more details).
.

7. **Adaptive Partitioning Method (BIN):** Proposed by Celluci et al. (2005) as a slight improvement on a classic binning approach to estimating mutual information. The goal of a bin based method is to estimate $p(x, y), p(x), p(y)$ empirically by constructing histograms from samples, then directly estimating the mutual information from the ratio of these distributions. The challenge is determining the number and size of the bins. As in the KDE method, this approach involves division by estimated densities which leads to compounded error. Additionally, as the dimension of the feature space increases, so does the error of this method (see *paper* for more details).