MINE: Mutual Information Neural Estimation

ICML 2018

Jaehyun Ko

March 2, 2023

Introduction

Motivation

Estimate mutual information between two random variables using neural networks.

Toy data

Two distribution

$$X \sim \text{sgn}(\mathcal{N}(0,1))$$
$$Y \sim X + \mathcal{N}(0,0.2)$$

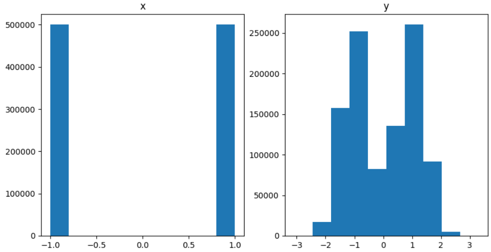

Figure: Toy data

Entropy

Entropy is a measure of the uncertainty of a random variable.

Definition 1.

Entropy For any probability density fuction p, entropy is defied as

$$H(x) = \mathbb{E}_p[-\log p(x)] - \int p(x) \log p(x) dx$$

- **Entropy** is a measure of the uncertainty of a random variable.
- average bit-length to representate RV [1].

Cross Entropy

The cross-entropy between two probability distributions p and q over the same underlying set of events measures the average number of bits needed to identify an event drawn from the set if a coding scheme used for the set is optimized for an estimated probability distribution q, rather than the true distribution p.

Definition 2.

Cross Entropy(CE) is defined as

$$H(p, q) = \mathbb{E}_p[-\log q(x)] = \int p(x) \log q(x) dx$$

Kullback-Leibler Divergence

Definition 3.

Kullback-Leibler Divergence (KLD) For two probability densities p(x), q(x) is defined as

$$D(p(x)||q(x)) = \int p(x) \log \frac{p(x)}{q(x)} dx,$$

it can be interpreted as difference of two entropy.

$$\begin{aligned} D(p(x)||q(x)) &= \int p(x)(-\log q(x))dx - \int p(x)(-\log p(x))dx \\ &= H(p,q) - H(p) \end{aligned}$$

Mutual Information

MI is a measure of the dependence between two random variables.

Definition 4.

Mutual Information (MI) Let X and Y be two random variables with a joint distribution $P(x,y)$ and P_x, P_y are marginal probability distribution each. The Mutual Information $I(X;Y)$ is defined as

$$I(X;Y) = \mathbb{E}_{P_{xy}}[\log \frac{P_{xy}}{P_x P_y}]$$

Mutual Information(cont.)

we can rewrite the mutual information as follows.

$$I(X;Z) = \mathbb{E}_P[-\log P_x] - \mathbb{E}_P[-\log \frac{P_y}{P_{xy}}]$$
$$= H(X) - H(X|Z)$$

MI between X and Z can be understood as the decrease of the uncertainty in X given Z. and it also representated as KLD between joint distribution and product of marginal distribution.

$$I(X;Z) = D(P_{xy}||P_x \otimes P_y) \tag{1}$$

Donsker-Varadhan Representation

Theorem 5.

Donsker-Varadhan Representation (DV) *Let X be a random variable with domain $\mathcal{X}$, let P, Q be two probability density functions and T be a function on $\mathcal{X}$, Then, for any $x \in \mathcal{X}$, the KLD admits the following dual Representation*

$$D(P||Q) = \sup_{T:\mathcal{X}\to\mathbb{R}} \{\mathbb{E}_P[T] - \log \mathbb{E}_Q[e^T]\}$$

the proof of theorem consists of two steps.

- **Step 1** : Existence of supremum in Donsker-Varadhan variational representation
- **Step 2** : Lower bound for the Kullback Liebler Divergence

Donsker-Varadhan Representation(cont.)

Existence of supremum in Donsker-Varadhan variational representation

Lemma 6.

There exists a function $T^* : X \to \mathbb{R}$ such that satisfies the condition of equality.

choise $T^* = \log \frac{P}{Q}$, then prove in the following page.

Donsker-Varadhan Representation(cont.)

Existence of supremum in Donsker-Varadhan variational representation

$$
\begin{aligned}
D_{\mathsf{KL}}(P|Q) &= \mathbb{E}_P[T^*(X)] - \log(\mathbb{E}_Q[e^{T^*(X)}]) && (2)\\
&= \mathbb{E}_P[\log \frac{P(X)}{Q(X)}] - \log(\mathbb{E}_Q[e^{\log \frac{P(X)}{Q(X)}}]) && (3)\\
&= D_{\mathsf{KL}}(P|Q) - \log(\mathbb{E}_Q[\frac{P(X)}{Q(X)}]) && (4)\\
&= D_{\mathsf{KL}}(P|Q) - \log(\int_{\mathcal{X}} Q(x)\frac{P(x)}{Q(x)}dx) && (5)\\
&= D_{\mathsf{KL}}(P|Q) - \log(\int_{\mathcal{X}} P(x)dx) && (6)\\
&= D_{\mathsf{KL}}(P|Q) - \log(1) && (7)\\
&= D_{\mathsf{KL}}(P|Q) && (8)
\end{aligned}
$$

Donsker-Varadhan Representation(cont.)

Lower bound for the Kullback Liebler Divergence

Lemma 7.

For any function $T : X \to \mathbb{R}$ *the following inequality holds:*
$D_{KL}(P|Q) \geq \sup_{T:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[T(X)] - \log \mathbb{E}_Q[e^{T(X)}]$

suppose new probability density function G is defined as follows:

$$G(x) = \frac{Q(x)e^T}{\mathbb{E}_Q[e^{T(X)}]} \tag{9}$$

$$\int_{\mathcal{X}} G(x)dx = \frac{\int_{\mathcal{X}} Q(x)e^T}{\mathbb{E}_Q[e^{T(X)}]} = \frac{\mathbb{E}_Q[e^{T(X)}]}{\mathbb{E}_Q[e^{T(X)}]} = 1 \tag{10}$$

Donsker-Varadhan Representation(cont.)

Lower bound for the Kullback Liebler Divergence

$$
\begin{aligned}
D_{\mathsf{KL}}(P|Q) - \sup_{T:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[T(X)] + \log \mathbb{E}_Q[e^{T(X)}] && (11)\\
= \mathbb{E}_P[\log \frac{P(X)}{Q(X)} - T(X)] + \log(\mathbb{E}_Q[e^{T(X)}]) && (12)\\
= \mathbb{E}_P[\log \frac{P(X)}{Q(X)e^{T(X)}}] - \log(\mathbb{E}_Q[e^{T(X)}]) && (13)\\
= \mathbb{E}_P[\log \frac{P(X)\mathbb{E}_Q[e^{T(X)}]}{Q(X)e^{T(X)}}] && (14)\\
= \mathbb{E}_P[\log \frac{P(X)}{G(X)}] && (15)\\
= D_{\mathsf{KL}}(P|G) \geq 0 && (16)
\end{aligned}
$$

MINE

Mutual Information Neural Estimation

in this section, we will Donsker-Varadhan variational formulation in order to estimate mutual information, via approximating T using neural network. according to discussion so, we can estimate the mutual information by maximizing the following cost function:

$$I(X;Y) = \sup_{T:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}} \mathbb{E}_{P_{XY}}[T(X,Y)] - \log \mathbb{E}_{P_X\otimes P_Y}[e^{T(X,Y)}] \tag{17}$$

Algorithm 1: Mutual Information Neural Estimation (MINE)

Input: Joint distribution P_{XY} and neural network architecture

Output: An estimate of the mutual information $I(X;Y)$

Initialize network parameters θ **repeat**

Draw mini-batch of samples: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_m, Y_m) \sim P_{XY}$;

Draw n samples from the marginal distribution: $Y_1, Y_2, \ldots, Y_n \sim P_Y$;

Evaluate: $\hat{I}_\theta(X;Y) \to \frac{1}{n}\sum_{i=1}^n T_\theta(X_i, Y_i) - \log(\frac{1}{n}\sum_{i=1}^n e^{T_\theta(X_i, \tilde{Y}_i)})$;

Update network parameters: $\theta \to \theta + \nabla_\theta \hat{I}_\theta(X;Y)$;

until *convergence*;

return An estimate of the mutual information $I(X;Y)$

Claude Elwood Shannon.
A mathematical theory of communication.
The Bell System Technical Journal, 27(3):379–423, 1948.