

MINE: Mutual Information Neural Estimation [1]

ICML 2018

Jaehyun Ko

March 7, 2023

- 1 Introduction
- 2 Backgrounds
 - Information Theory
 - Donsker-Varadhan Variational Formula
- 3 MINE
- 4 Theoretical Properties
- 5 Experiments
- 6 References

Introduction

Motivation

Estimate mutual information between two random variables using neural networks.

Entropy

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

Definition 1.

Entropy For any probability density function p , entropy is defined 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 [2].

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

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

Mutual Information(cont.)

we can rewrite the mutual information as follows.

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

MI between X and Z can be understood as the decrease of the uncertainty in X given Z. and it also represented 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

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

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 $T^ : X \rightarrow \mathbb{R}$ 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

$$D_{\text{KL}}(P|Q) = \mathbb{E}_P[T^*(X)] - \log(\mathbb{E}_Q[e^{T^*(X)}]) \quad (2)$$

$$= \mathbb{E}_P[\log \frac{P(X)}{Q(X)}] - \log(\mathbb{E}_Q[e^{\log \frac{P(X)}{Q(X)}}]) \quad (3)$$

$$= D_{\text{KL}}(P|Q) - \log(\mathbb{E}_Q[\frac{P(X)}{Q(X)}]) \quad (4)$$

$$= D_{\text{KL}}(P|Q) - \log(\int_{\mathcal{X}} Q(x) \frac{P(x)}{Q(x)} dx) \quad (5)$$

$$= D_{\text{KL}}(P|Q) - \log(\int_{\mathcal{X}} P(x) dx) \quad (6)$$

$$= D_{\text{KL}}(P|Q) - \log(1) \quad (7)$$

$$= D_{\text{KL}}(P|Q) \quad (8)$$

Donsker-Varadhan Representation(cont.)

Lower bound for the Kullback Liebler Divergence

Lemma 7.

For any function $T : X \rightarrow \mathbb{R}$ the following inequality holds:

$$D_{KL}(P|Q) \geq \sup_{T: X \rightarrow \mathbb{R}} \mathbb{E}_P[T(X)] - \log \mathbb{E}_Q[e^{T(X)}]$$

suppose new probability density function G is defined as follows:

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

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

Donsker-Varadhan Representation(cont.)

Lower bound for the Kullback Liebler Divergence

$$D_{\text{KL}}(P|Q) = \sup_{T:\mathcal{X}\rightarrow\mathbb{R}} \mathbb{E}_P[T(X)] + \log \mathbb{E}_Q[e^{T(X)}] \quad (11)$$

$$= \mathbb{E}_P[\log \frac{P(X)}{Q(X)} - T(X)] + \log(\mathbb{E}_Q[e^{T(X)}]) \quad (12)$$

$$= \mathbb{E}_P[\log \frac{P(X)}{Q(X)e^{T(X)}}] - \log(\mathbb{E}_Q[e^{T(X)}]) \quad (13)$$

$$= \mathbb{E}_P[\log \frac{P(X)\mathbb{E}_Q[e^{T(X)}]}{Q(X)e^{T(X)}}] \quad (14)$$

$$= \mathbb{E}_P[\log \frac{P(X)}{G(X)}] \quad (15)$$

$$= D_{\text{KL}}(P|G) \geq 0 \quad (16)$$

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:

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

MINE(cont.)

Estimation of each term of MINE

In order to estimate each term of MINE (17),

- 1 we need to the full knowlage of the joint and marginal distribution of X and Y .
- 2 we need to find maximization of $T(X, Y)$ over all possible functions, $T \in \mathcal{F}$.

MINE(cont.)

Estimation of each term of MINE

In assume that we have a sample of n independent and identically distributed samples from P_{XY} .

then, we can use the law of large numbers to obtain following approximation of the first expectation:

$$\mathbb{E}_{P_{XY}}[T(X, Y)] \approx \frac{1}{n} \sum_{i=1}^n T(X_i, Y_i) \quad (18)$$

MINE(cont.)

Estimation of each term of MINE

To obtain an approximation of the second expectation, we cannot use the law of large numbers, because the expectation is over the product of two random variables.

We can use some tricks: suffling. artificially construct a tuple set (X_i, \tilde{Y}_i) , where \tilde{Y}_i is a randomly sampled from $(Y_i)_{i=1}^n$.

Then, we can use the law of large numbers to obtain following approximation of the second expectation:

$$\mathbb{E}_{P_X \otimes P_Y} [e^{T(X,Y)}] \approx \frac{1}{n} \sum_{i=1}^n T(X_i, \tilde{Y}_i) \quad (19)$$

MINE(cont.)

Estimation of each term of MINE

Now, we can estimate the mutual information by maximizing the following cost function:

$$\hat{I}(X; Y) = \sup_{T: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}} \frac{1}{n} \sum_{i=1}^n T(X_i, Y_i) - \log \frac{1}{n} \sum_{i=1}^n T(X_i, \tilde{Y}_i) \quad (20)$$

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:

$$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n) \sim P_{XY}$$

 Draw n samples from the marginal distribution: $\tilde{Y}_1, \tilde{Y}_2, \dots, \tilde{Y}_n \sim P_Y$

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

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

until *convergence*

return An estimate of the mutual information $I_\theta(X; Y)$

Strong Consistency

Proof of Approximation of MINE

Definition 8.

The estimator $\hat{I}_\theta(X; Y)$ is said to be strongly consistent if:
for all $\epsilon > 0$, there exists a positive integer N and a choice of statistics network such that:

$$\forall n \geq N, |I(X; Y) - \widehat{I_\theta(X; Y)}_n| \leq \epsilon \quad (21)$$

where the probability is over a set of samples.

Experiment 1

Toy data

Two distribution

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

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

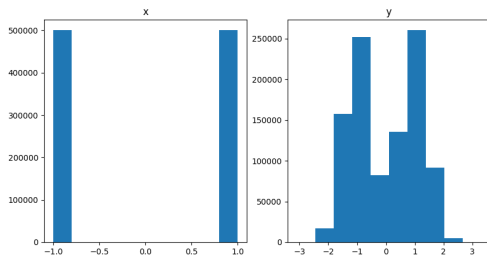


Figure: Toy data

Experiment 1

Toy data

compute mutual information of two distribution directly

$$\begin{aligned} I(X; Y) &\approx \frac{1}{n} \sum_{i=1}^n \frac{\log P_{X,Y}(x_i, y_i)}{\log P_X(x_i) P_Y(y_i)} \\ &= \frac{1}{n} \sum_{i=1}^n \log \frac{P_{Y|X}(y_i|x_i)}{P_{Y|X}(y_i|X) P_X(X)} \\ &= \frac{1}{n} \sum_{i=1}^n \log \frac{P_{Y|X}(y_i|x_i)}{0.5 P_{Y|X}(y_i|X = -1) + 0.5 P_{Y|X}(y_i|X = 1)} \\ &\approx 0.65865 \end{aligned}$$

Experiment 1

Toy data

```
1 def forward(self, x, y):
2     batch_size = x.size(0)
3     tiled_x = torch.cat([x, x], dim=0)
4     idx = torch.randperm(batch_size)
5
6     shuffled_y = y[idx]
7     concat_y = torch.cat([y, shuffled_y], dim=0)
8
9     inputs = torch.cat([tiled_x, concat_y], dim=1)
10    logits = self.layers(inputs)
11
12    pred_xy = logits[:batch_size]
13    pred_x_y = logits[batch_size:]
14    loss = -(torch.mean(pred_xy)
15             - torch.log(torch.mean(torch.exp(pred_x_y))))
16
17    return loss
```

Figure: MINE implementation

Experiment 1

Toy data

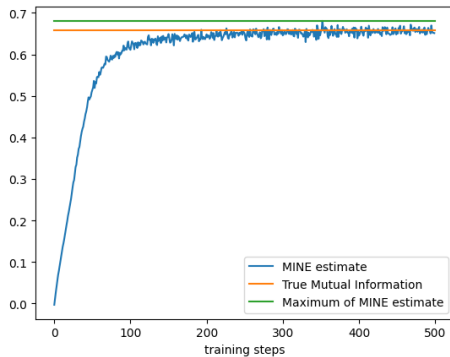


Figure: MINE result of toy data

Experiment 2

Two Gaussian with different correlation

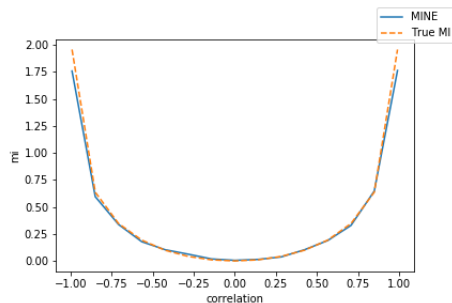




Figure: MINE result of gaussian

-  Ishmael Belghazi, Sai Rajeswar, Aristide Baratin, R. Devon Hjelm, and Aaron C. Courville.
MINE: mutual information neural estimation.
CoRR, abs/1801.04062, 2018.
-  Claude Elwood Shannon.
A mathematical theory of communication.
The Bell System Technical Journal, 27(3):379–423, 1948.