MINE: Mutual Information Neural Estimation

Jaehyun Ko

February 22, 2023

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.

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

Donsker-Varadhan Representation

Theorem 5.

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

Existence of supremum in Donsker-Varadhan variational representation

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