Notes on mutual information

Yao Fu, Columbia University
yao.fu@columbia.edu

Let's talk about representation learning, where our goal is to learning effective, interpretable representations for downstream tasks. We use mutual information as a measure for the non-linear dependency between the learned representations and the data. We first discuss some basics of information theory, then the estimation of the mutual information.

Entropy: $H(X) = -\sum_x p(x) \log p(x) \le \log |\Omega_x|$ measures the information contained by X.
- entropy is independent of Ω_x the support of X.
- bounded by uniform dist.

Relative Entropy / KL divergence:

$$R(Y|X) = \sum_{xy} p(y) \log \frac{p(y)}{p(x)}$$

measures the "distance" between $P(Y)$ and $P(X)$.
- also independent of Ω_x and Ω_y.
- Not symmetric. JSD is.

Note if $Y = f(X)$. f is a deterministic function. Then

$$H(Y) \le H(X).$$

"=" sufficient when f is a one-to-one mapping.

Also related to the data-processing inequality
$X \to Y \to Z$. $I(X, Y) \ge I(X, Z)$
when $Y = X$ and $Z = f(Y)$.

Pointwise Mutual Information (PMI)

$$PMI(x, y) = \log \frac{p(x, y)}{p(x)\, p(y)}.$$

Mutual Information (MI) is the expected PMI:

$$I(X, Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)}$$

$$= \mathbb{E}_{p(X, Y)}\left[\log \frac{p(x, y)}{p(x)\, p(y)} \right]$$

$$= KL(\, p(X, Y) \,\|\, p(X)\, p(Y)\,).$$

$$= H(Y) - H(Y|X).$$

There are many non-trivial things in mutual information:

① Conditional Entropy:

$$\underbrace{H(Y|X)}_{\text{independent of } x.} = \underbrace{\sum_x p(x)}_{\text{need to sum over } x.} \sum_y p(y|x) \neq \underbrace{\sum_y p(y|x)}_{\text{depend on the value of } x.}$$

This is a common caveat. Some top-venue papers also make such mistakes.

② Similarly. Conditional Mutual Information:

$$\underbrace{I(Y\, Z | X)}_{\text{Independent of } x.} = \underbrace{\sum_x p(x)}_{\text{need to sum over } x.} \sum_{y,z} p(y, z|x) \log \frac{p(y, z|x)}{p(y|x)\, p(z|x)} \neq \underbrace{\sum_{y,z} p(y, z|x) \log \frac{p(y, z|x)}{p(y|x)\, p(z|x)}}_{\text{depend on the value of } x.}$$

So this is ironic: The conditional MI / Ent is independent of the condition.
This is NOT conditional probability $p(y|x)$. which depends on x.

③ Bounds / Asymptotic behaviors.

Actually, MI is not bounded, which might be problematic in optimization as its gradient may overwhelm the grad. of other loss terms. But we do have some discussions on its asymptotic behaviors:

First, the entropy term:

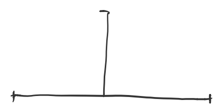
$$H(Y) \le \log|\Omega_y| \quad \text{is bounded by the uniform entropy.}$$

Second, the conditional entropy:

$$H(Y|X) = -\sum_{x,y} p(x,y)\log p(y|x) \ge 0.$$

is lower-bounded by the one-hot distribution. When a conditional dist. $p(y|x)$ is one-hot, there is a deterministic mapping $f(x)=y$

Overall, optimizing MI will push the prior to uniform, and the posterior to one-hot. As is shown:

 $\rightarrow$ 

$p(y) \to$ uniform

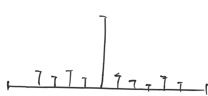
$H(Y) \to \log|\Omega_y|$

This is to say, before knowing X, we know nothing about Y, and it can be any value (uniformly).

$p(y|x) \to$ one-hot = deterministic.

$H(Y|X) = 0$

This is to say, after knowing X, we immediately know what value Y is, and there is no randomness remained.



This is somewhere in the middle. After knowing X, we have more confidence that y should be a particular value.

Also note that these bounds relies on the assumption that X and Y are discrete. Things will be different when they are continuous.

④. Differential Entropy.

In above discussion, we assume all variables are discrete. When they are continuous, it is possible that:

$$H(Y) = \int p(y)\log\frac{1}{p(y)}\,dy < 0.$$

An example is a uniform dist over $[0, \frac{1}{2}]$. Also the entropy can be arbitrarily negative, making the mutual information unbounded, leading to potential optimization issues. See the MINE paper.

In other words, Discrete MI / Ent. is bounded while continuous MI / Ent is NOT.

⑤. Joint v.s. Product of Marginal.

The KL version of MI = KL ($p(x, y) \| p(x) p(y)$) involves the joint $p(x, y)$ and the product of marginal $p(x) \cdot p(y)$. How do they different w. each other ?

Consider the sampling problems:

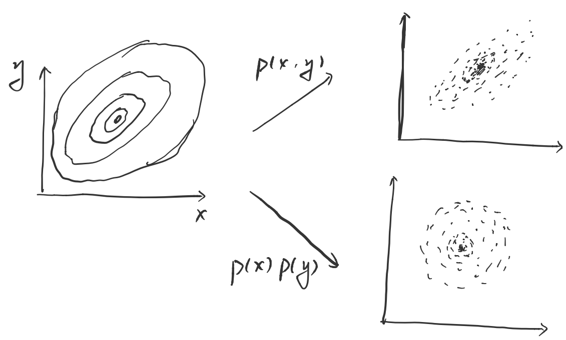
(1). sample $x^{(1)}, y^{(1)} \sim p(x, y)$.

(2) sample $x^{(2)} \sim p(x) = \sum_y p(x, y)$. $y^{(2)} \sim p(y) = \sum_x p(x, y)$.

Q: How does ($x^{(1)}$, $y^{(1)}$) compared to ($x^{(2)}$, $y^{(2)}$) ?

A: $x^{(1)}$ and $y^{(1)}$ are correlated with each other.

$x^{(2)}$ and $y^{(2)}$ are independent of each other.



In other words. $p(x)p(y)$ makes an indepent dist $q(x, y) = p(x)p(y)$. from a correlated $p(x, y)$.

The more MI, the higher divergence $p(x, y)$ from $p(x)p(y)$. the more dependence. The less MI. the less dependence.

⑥. Why MI hard to estimate ?

Suppose X is observed. Y latent. To estimate MI, we need:

(1) samples from $p(x, y)$.

(2) evaluate the joint $p(x, y)$.

(3) evaluate the marginal $p(x)$. $p(y)$

When we have a discriminative model $p(y \mid x)$.

Then: (1) $x \sim$ emprical. $y \sim p(y \mid x)$. so we can do condition (1)

(2) $p(x, y) = p(y \mid x) p(x)$. we can eval. $p(y \mid x)$. but $p(x)$ undifined.

(3) $p(y)$. is also undefined.

When we have a factorized generative model: $p(x, y) = p(y \mid x) p(x)$.

Then: (1) $x \sim$ empirical. $y \sim p(y \mid x)$

(2) $\dfrac{p(x, y)}{p(y) p(x)} = \dfrac{p(y \mid x)}{p(y)}$ $p(y)$ still undefined. and cannot do $p(y) = \sum_x p(y \mid x) p(x)$

When we have a full generative model $p(x, y)$.

Then still cannot evaluate $p(x) p(y)$.

This is why we need approximate inference methods.

In the above discussion, we show many basics and discussions about what is mutual information and what we are actually optimizing over. We have also shown why it is hard to estimate. Now we show the recent estimators for MI.

1. MINE. mutual information neural estimation.

1.1. Original Donsker - Varadam representation.

$$\hat{I}_{DV}(X,Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)\,p(y)}\right].$$

$$\geq \sup_{\eta} \mathbb{E}_{p(x,y)}[T_\eta(x,y)] - \log \mathbb{E}_{p(x)p(y)}[\exp(T_\eta(x,y))]$$

1.2. The Jensen - Shanon representation (more stable).

$$\hat{I}_{JSD}(X,Y) = \mathbb{E}_{p(x,y)}[-sp(-T(x,y))] - \mathbb{E}_{p(x)p(y)}[sp(T(x,y))]$$

$$sp(x) = \log(1+e^x).$$

1.3. The Noise - Contrastive Estimation:

$$\hat{I}_{NCE}(X,Y) = \mathbb{E}_{p(x,y)}\left[T(x,y) - \mathbb{E}_{x'\sim p(x)\, y'\sim p(y)}\left[\log \sum_{x'} e^{T(x',y')}\right]\right].$$

Details TBC.

2. The Variational Estimators.

TBC.

References
Hjelm 19, Learning deep representations by mutual information estimation and maximization
Belghazi 18, Mutual information neural estimation
Cover and Tomas 91, Elements of information theory. Chapter 2.