

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

Let's talk about representation learning, where our goal is to learn 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, and optimize the MI. We first state our general settings and give a detailed analysis of the basics of information theory, then the estimation of the mutual information.

Representation Learning: $Y \sim p(Y|X) \sim P(Y, X)$ = a representation (or feature, for simplification)
 discriminative generative. for X .

max. $I(Y, X)$. the mutual information.

Then use Y for any downstream tasks.

Entropy: $H(X) = -\sum_x p(x) \log p(x) \leq \log |\Omega_X|$ measures the information contained by X .

entropy is independent of Ω_X the support of X .
 bounded by uniform dist.

Relative Entropy / KL divergence:

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

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

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

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

Pointwise Mutual Information (PMI)

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

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

Mutual Information (MI) is the expected PMI.

$$\begin{aligned} I(X, Y) &= \sum_{xy} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} = \mathbb{E}_{p(X, Y)} [\log \frac{p(x, y)}{p(x)p(y)}] \\ &= \text{KL}(p(X, Y) \| p(X)p(Y)) = H(X) - H(Y|X). \end{aligned}$$

There are many non-trivial things in mutual information.

① Conditional Entropy:

$$\underline{H(Y|X)} = \sum_x \underbrace{p(x)}_{\substack{\uparrow \\ \text{independent of } X.}} \sum_y \underline{p(y|x)} \neq \sum_y \underline{p(y|x)} \quad \text{need to sum over } x.$$

This is a common caveat.

Some top-venue papers also make such mistakes.

depend on the value of x .

② Similarly. Conditional Mutual Information:

$$\underline{I(YZ|X)} = \sum_x \underbrace{p(x)}_{\substack{\uparrow \\ \text{independent of } X.}} \sum_{yz} \underline{p(y, z|x)} \log \frac{p(y, z|x)}{p(y|x)p(z|x)} \neq \sum_{yz} \underline{p(y, z|x)} \log \frac{p(y, z|x)}{p(y|x)p(z|x)} \quad \text{need to sum over } x.$$

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.

What are the extreme cases of MI? Is it possible that two variables have infinite MI? How does MI behave when things go extreme? What can we ultimately achieve when we optimize MI?

First, the entropy term:

$H(Y) \in \log |\Omega_Y|$ is bounded by the uniform entropy.

In the case of language modeling, it should be:

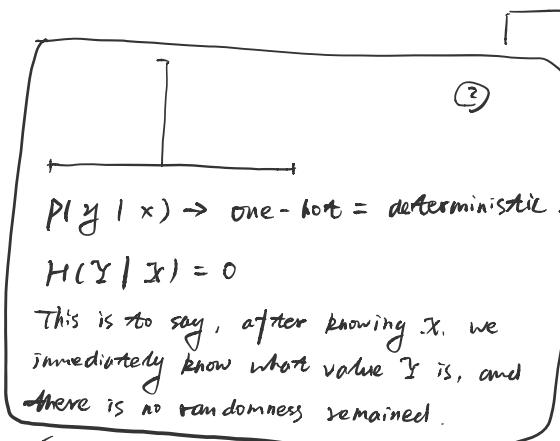
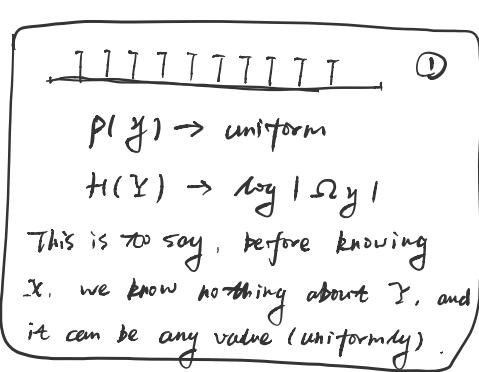
$$\log |V|^n = n \log V \Rightarrow \begin{array}{l} \text{Not a large} \\ \text{number!} \\ \text{V = vocabulary size} \end{array}$$

Second, the conditional entropy:

$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x) \geq 0, "=" \text{ when } p(y|x) = \text{one-hot.}$$

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:

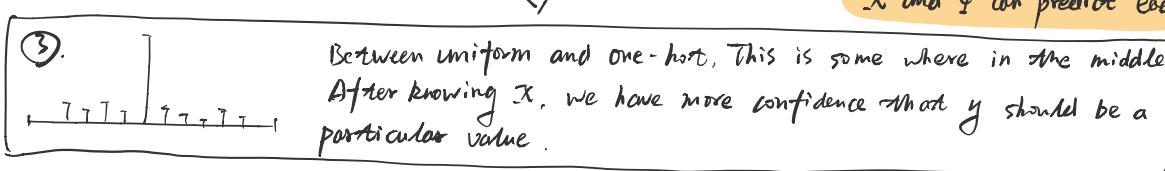


This also measures:
[To what extent Y can be predicted by X].

Recall the symmetry:
 $I = H(Y) + H(Y|X)$
 $= H(X) + H(X|Y)$.

so we also get:
[To what extent X can be pred. by Y]

Combining above we get: to what extent X and Y can predict each other. which is a



classical interpretation of MI.

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. Here we leave out the asymptotic behavior of continuous MI.

⑤. Joint v.s. Product of Marginal.

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

Consider two sampling problems:

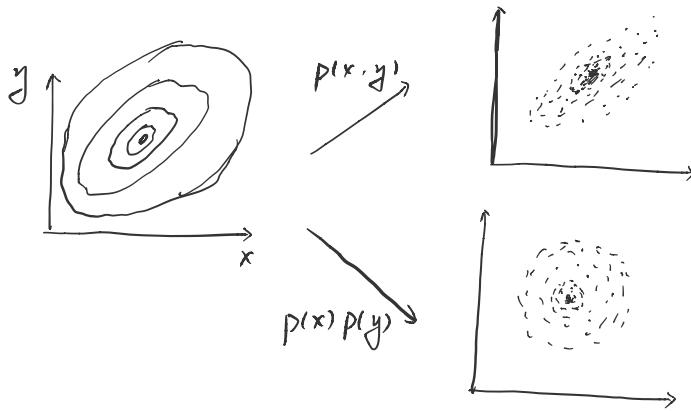
(1). sample $x'' y'' \sim p(x,y)$.

(2) sample $x'' \sim p(x) = \sum_y p(x,y)$, $y'' \sim p(y) = \sum_x p(x,y)$.

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

A: x'' and y'' are correlated with each other.

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



In other words, $p(x)p(y)$ constructs an independent dist $g(x,y) = p(x)p(y)$ out of 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 $p(x) \cdot p(y)$

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

Then: (1) $x \sim \text{empirical}$, $y \sim p(y|x)$, so we can do condition (1)

(2) $p(x,y) = p(y|x)p(x)$, we can eval. $p(y|x)$, but $p(x)$ undefined.

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

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

Then: (1) $x \sim \text{empirical}$, $y \sim p(y|x)$

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

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 - Varadhan representation.

$$\begin{aligned}\hat{I}_{\text{DV}}(X, Y) &= \mathbb{E}_{p(x, y)} \left[\log \frac{p(x, y)}{p(x)p(y)} \right] \\ &\geq \sup_{T_\eta} \mathbb{E}_{p(x, y)} [T_\eta(x, y)] - \log \mathbb{E}_{p(x)p(y)} [\exp(T_\eta(x, y))]\end{aligned}$$

1.2. The Jensen - Shannon representation (more stable)

$$\begin{aligned}\hat{I}_{\text{JSD}}(X, Y) &= \mathbb{E}_{p(x, y)} [-\text{sp}(-T(x, y))] - \mathbb{E}_{p(x)p(y)} [\text{sp}(T(x, y))] \\ \text{sp}(x) &= \log(1 + e^x).\end{aligned}$$

1.3. The Noise - Contrastive Estimation:

$$\hat{I}_{\text{NCE}}(X, Y) = \mathbb{E}_{p(x, y)} [T(x, y) - \mathbb{E}_{x \sim p(x), y \sim p(y)} [\log \frac{1}{x} e^{T(x, y)}]]$$

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.