CIE6020/MAT3350
Selected Topics in Information Theory

Lecture 2: Mutual Information and Divergence

---

11 Janurary 2019

The Chinese University of Hong Kong, Shenzhen

Conditional Entropy and Mutual Information

Conditional Entropy

- For random variables X and Y, the *conditional entropy* $H(Y|X)$ is defined as

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

- Denote

$$H(Y|X=x) = H(p_{Y|X}(\cdot|x)) = -\sum_{y} p(y|x) \log p(y|x).$$

- We can write

$$H(Y|X) = \sum_{x} p(x) H(Y|X=x).$$

- In other words, the conditional entropy is the expectation of the entropy of the conditional distribution of Y given $X = x$.

Basic Properties

- $H(Y|X) \geq 0$ with equality iff Y is a function of X (over the support of X).
- (Chain rule) $H(X, Y) = H(X) + H(Y|X)$.
- $H(Y|X) \leq H(Y)$ with equality iff X and Y are independent. In other words, conditioning reduces entropy.

Mutual Information

Definition
The *mutual information* between random variables X and Y is defined as

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

Remark

1. $I(X;Y)$ is symmetrical in X and Y.
2. $I(X;X) = H(X)$: observing X can get all the information of X.
3. $I(X;Y) \geq 0$ (Log-sum inequality).
4. $I(X;Y)$ only depends on the joint distribution $p_{X,Y}$, so we also write $I(X;Y) = I(p_{X,Y})$.

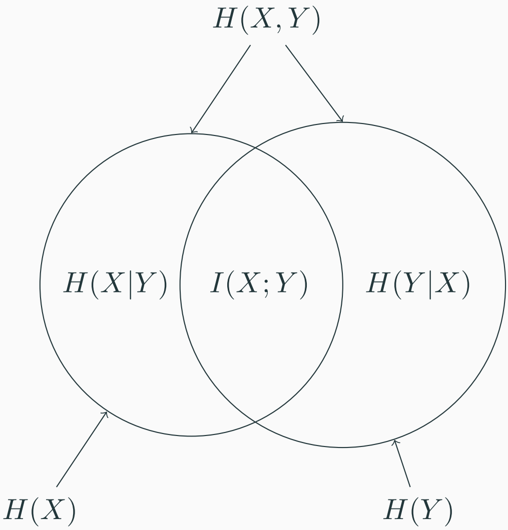
Relations

- We have the following equalities:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y). \end{aligned}$$

- If the alphabets are not finite, the above equalities hold provided that all the entropies and conditional entropies are finite.

Information Diagram of Two Random Variables



Example
Let X and Y have the following joint distribution:

$Y \backslash X$	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

Chain Rules

Exercise

Theorem

$$H(X, Y|Z) = H(X|Z) + H(Y|X, Z).$$

Theorem (Chain rule for entropy)

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, \dots, X_{i-1}).$$

Proof.

$$H(X_1, X_2, \dots, X_n) = H(X_1, X_2, \dots, X_{n-1}) + H(X_n | X_1, X_2, \dots, X_{n-1}).$$

$\square$

Theorem (Independence bound on entropy)

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

with equality iff X_i are independent.

Proof.
Chain rule for entropy and conditioning reduces entropy. $\square$

Chain Rule for Conditional Entropy

Theorem

$$H(X_1, X_2, \ldots, X_n | Y) = \sum_{i=1}^{n} H(X_i | X_1, \ldots, X_{i-1}, Y).$$

Conditional Mutual Information

Definition
The *conditional mutual information* of random variables X and Y given Z is defined by

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}. \end{aligned}$$

Conditional Mutual Information

Definition
The *conditional mutual information* of random variables X and Y given Z is defined by

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}. \end{aligned}$$

Analogous to conditional, write

$$I(X;Y|Z) = \sum_{z} p(z) I(X;Y|Z=z)$$

where

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

Theorem (Chain rule for mutual information)

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1).$$

Proof.
Apply chain rules for entropy. □

Theorem (Chain rule for conditional mutual information)

$$I(X_1, X_2, \ldots, X_n; Y|Z) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \ldots, X_1, Z).$$

Information Divergence

Relative Entropy

Definition
The *relative entropy* (*information divergence* or *Kullback-Leibler distance*) between two probability mass function $p(x)$ and $q(x)$ is defined as

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \log \frac{p(X)}{q(X)},$$

where we adopt the convention that $0 \log \frac{0}{0} = 0$, $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$.

Remark

- $I(X;Y) = D(p(x,y)\|p(x)p(y))$.
- $D(p\|q) \geq 0$ with equality iff $p = q$.
- $D(p\|q)$ is not symmetric, i.e., we do not have $D(p\|q) = D(q\|p)$ in general.

Example

Consider two binary distributions p and q on $\{0, 1\}$. Let $p(1) = r$ and $q(1) = s$. Calculate $D(p||q)$ and $D(q||p)$. When they are the same?

Convexity

- $D(p\|q)$ is convex in the pair (p, q), which implies
- $H(p)$ is a concave function of p, and
- $I(X; Y)$ is 1) a concave function of $p(x)$ for fixed $p(y|x)$ and is 2) a convex function of $p(y|x)$ for fixed $p(x)$.

Theorem (Chain rule for relative entropy)

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)).$$

Proof.

$$\begin{aligned}
D(p(x,y)||q(x,y)) &= \mathbb{E}_p \log \frac{p(X,Y)}{q(X,Y)} \\
&= \mathbb{E}_p \log \frac{p(X)p(Y|X)}{q(X)q(Y|X)} \\
&= \mathbb{E}_p \left[\log \frac{p(X)}{q(X)} + \log \frac{p(Y|X)}{q(Y|X)}\right] \\
&= \mathbb{E}_p \log \frac{p(X)}{q(X)} + \mathbb{E}_p \log \frac{p(Y|X)}{q(Y|X)}
\end{aligned}$$

□