Information and Coding Theory **Autumn 2014**

Lecture 4: October 9, 2014

Lecturer: Madhur Tulsiani Scribe: Hai Wang

1 Mutual Information

The mutual information between two random variables X and Y is defined by the formula

$$I(X;Y) = H(X) - H(X|Y) \tag{1}$$

where $H()$ denotes the entropy. Mutual information measures how much information in the X about Y (vice versa), and mutual information is not symmetric. Using the *Chain Rule* for entropy $H(X,Y) = H(X) + H(Y|X)$, we have:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) \tag{2}$$

Example 1.1 *Consider the random variable (X,Y) with $X \vee Y = 1$, $X \in \{0,1\}$ and $Y \in \{0,1\}$ such that:*

$$(X,Y) = \begin{cases} 10 & w.p\ 1/3 \\ 01 & w.p\ 1/3 \\ 11 & w.p\ 1/3 \end{cases}$$

Then, we can calculate the entropy as following:

$$H(X) = \frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log 3 \tag{3}$$

$$H(Y) = \frac{1}{3}\log 3 + \frac{2}{3}\log\frac{3}{2} \tag{4}$$

$$H(X,Y) = 3 \times \frac{1}{3}\log 3 = \log 3 \tag{5}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = \log 3 - \frac{4}{3}\log 2 \tag{6}$$

Let's consider the mutual information $I(X;Y|Z)$ which can be defined as:

$$I(X;Y|Z) = \mathbb{E}_Z[I(X|Z=z;Y|Z=z)] = H(X|Z) - H(X|Y,Z) = \mathbb{E}_Z[H(X|Z=z) - H(X|Y,Z=z)] \tag{7}$$

We know that in entropy, we have $H(X|Y) \leq H(X)$, then in mutual information, can we have the similar conclusion that $I(X;Y|Z) \leq I(X;Y)$? The answer is no, let's take a look at the following example

Example 1.2 *Consider the random variable (X, Y, Z), $X \in \{0, 1\}$, $Y \in \{0, 1\}$ and $Z = X \oplus Y$ such that:*

$$(X, Y, Z) = \begin{cases} 000 & w.p\ 1/4 \\ 011 & w.p\ 1/4 \\ 101 & w.p\ 1/4 \\ 110 & w.p\ 1/4 \end{cases}$$

We know in this case, X, Y are independent and thus $I(X; Y) = 0$, but

$$\begin{aligned} I(X : Y|Z) &= \mathbb{E}_Z[I(X|Z = z; Y|Z = z)] \\ &= \frac{1}{2}I(X|Z = 0; Y|Z = 0) + \frac{1}{2}I(X|Z = 1; Y|Z = 1) \\ &= \frac{1}{2}\log 2 + \frac{1}{2}\log 2 = \log 2 \end{aligned}$$

Recall that in entropy, we have the following *chain rule*:

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \ldots + H(X_n|X_1, \ldots, X_{n-1}) \tag{8}$$

Similarly, in mutual information, we have:

Lemma 1.3 $I((X_1, \ldots, X_n); Y) = \sum_{i=1}^{n} I(X_i; Y|X_1, \ldots, X_{i-1})$

Proof:

$$\begin{aligned} I((X_1, \ldots, X_n); Y) &= H(X_1, \ldots, X_n) - H(X_1, \ldots, X_n|Y) \\ &= \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}) - \sum_{i=1}^{n} H(X_i|Y, X_1, \ldots, X_{i-1}) \\ &= \sum_{i=1}^{n}(H(X_i|X_1, \ldots, X_{i-1}) - H(X_i|Y, X_1, \ldots, X_{i-1})) \\ &= \sum_{i=1}^{n} I(X_i; Y|X_1, \ldots, X_{i-1}) \end{aligned}$$

■

Lemma 1.4 (Special case of data processing inequality) *Let g be a function of Y, then*

$$I(X; Y) \geq I(X; g(Y)).$$

Proof: we first know that $H(X|Y, g(Y)) = H(X|Y)$, since $g(Y)$ is totally determined by Y, then

$$I(X; Y) = H(X) - H(X|Y) = H(X) - H(X|Y, g(Y)) \geq H(X) - H(X|g(Y)) = I(X; g(Y)) \tag{9}$$

■

Definition 1.5 *Suppose $I(X; Y) = I(X; g(Y))$, then $g(Y)$ is called sufficient statistic to X.*

Example 1.6

$$X = \begin{cases} \frac{1}{2} & w.p\ 1/2 \\ \frac{1}{3} & w.p\ 1/2 \end{cases}$$

Let Y be a sequence of n tosses of a coin with probability of heads given by X. Let $g(Y)$ be the number of heads in Y.

Exercise 1.7 *Prove $I(X;Y) = I(X;g(Y))$ in the above example.*

2 KL-divergence

Let P and Q be two distributions on a universe U, then the KL-divergence between P and Q can be defined as:

$$D(P||Q) = \sum_{x \in U} P(x) \log \frac{P(x)}{Q(x)} \tag{10}$$

It's easy to check that $D(P||Q)$ and $D(Q||P)$ are not equal.

Example 2.1 *Suppose $U = \{a, b, c\}$, and $P(a) = \frac{1}{3}$, $P(b) = \frac{1}{3}$, $P(c) = \frac{1}{3}$ and $Q(a) = \frac{1}{2}$, $Q(b) = \frac{1}{2}$, $Q(c) = 0$. Then*

$$D(P||Q) = \frac{2}{3} \log \frac{2}{3} + \infty = \infty\,.$$

$$D(Q||P) = \log \frac{3}{2} + 0 = \log \frac{3}{2}\,.$$

Even though the KL-divergence is not symmteric, it is often used as a measure of "dissimilarity" between two distribution. Towards this, we first prove that it is non-negative and is 0 if and only if $P = Q$.

Lemma 2.2 *Let P and Q be distributions on a finite universe U. Then $D(P||Q) \geq 0$ with equality if and only if $P = Q$.*

Proof: Let $\mathrm{Supp}(P) = \{x : P(x) > 0\}$. Then, we must have $\mathrm{Supp}(P) \subseteq \mathrm{Supp}(Q)$ if $D(P,Q) < \infty$. We can then assume without loss of generality that $\mathrm{Supp}(Q) = U$. Using the fact the log is a concave function, with Jensen inequality, we have:

$$\begin{aligned} D(P||Q) = \sum_{x \in U} P(x) \log \frac{P(x)}{Q(x)} &= \sum_{x \in \mathrm{Supp}(P)} P(x) \log \frac{P(x)}{Q(x)} \\ &= -\sum_{x \in \mathrm{Supp}(P)} P(x) \log \frac{Q(x)}{P(x)} \\ &\geq -\log \left(\sum_{x \in \mathrm{Supp}(P)} P(x) \cdot \frac{Q(x)}{P(x)} \right) \\ &= -\log \left(\sum_{x \in \mathrm{Supp}(P)} Q(x) \right) \\ &\geq -\log 1 = 0\,. \end{aligned}$$

For the case when $D(P||Q) = 0$, we note that this implies $P(x) = Q(x)\ \forall x \in \text{Supp}(P)$, which in turn gives that $P(x) = Q(x)\ \forall x \in U$. ■

We note that KL-divergence also has an interesting interpretation in terms of source coding. Writing

$$D(P||Q) \;=\; \sum P(x)\log\frac{P(x)}{Q(x)} \;=\; \sum P(x)\log\frac{1}{Q(x)} - \sum P(x)\log\frac{1}{P(x)}\,,$$

we can view this as the number of extra bits we use (on average) if we designed a code accoriding to the distribution P, but used it to communicate outcomes of a random variable X distributed according to Q.

We now relate KL-divergence to some other notions of distance between two probability distributions.

Definition 2.3 *Let P and Q be two distributions on a finite universe U. Then the* total-variation distance *between P and Q is defined as*

$$\delta_{TV}(P,Q) \;=\; \frac{1}{2}\cdot\|P-Q\|_1 \;=\; \frac{1}{2}\cdot\sum_{x\in U}|P(x)-Q(x)|\;.$$

The quantity $\|P-Q\|_1$ is referred to as the ℓ_1-distance *between P and Q.*

In many applications, we want to actually bound the ℓ_1-distance between P and Q but it's easier to analyze the KL-divergence. The following inequality helps relate the two.

Lemma 2.4 (Pinsker's inequality) *Let P and Q be two distributions defined on a universe U. Then*

$$D(P||Q) \;\geq\; \frac{1}{2\ln 2}\cdot\|P-Q\|_1^2\;.$$

We will see the proof of this in the next lecture.