EE 376A/Stats 376A: Information theory **Winter 2017**

Lecture 2 — January 12

Lecturer: David Tse *Scribe: Huy Pham, Connor Brinton, Nora Brackbill, Vivek B.*

2.1 Outline

- Entropy and Chain rule
- Conditional entropy
- Mutual information

Reading: CT 2.8, 3.1, 3.2, 4.1

2.2 Entropy

2.2.1 Recap

Consider a discrete random variable X with a pmf - $p(x)$. The entropy of X is defined as

Definition 1. Entropy

$$H(X) \triangleq \mathbb{E}\left[\log \frac{1}{p(X)}\right] = \sum_x p(x) \log \frac{1}{p(x)}$$

Salient features:

1. $H(X) \geq 0$
2. $H(X)$ is label-invariant.

2.3 The Chain Rule for Entropy

Consider two discrete random variables, X and Y (eg - Flips of two coins). If X and Y are not independent, observing one might give some information about the other. Hence, naively adding the individual entropies will result in over-counting the total entropy of X and Y.

By the definition 1, the entropy for a pair of discrete random variables (X, Y) is:

$$\begin{aligned} H(X,Y) &= \mathbb{E}\left[\log \frac{1}{p(X)p(Y|X)}\right] \qquad (2.1) \\ &= \mathbb{E}\left[\log \frac{1}{p(X)}\right] + \mathbb{E}\left[\log \frac{1}{p(Y|X)}\right] \\ H(X,Y) &= H(X) + H(Y|X) \qquad (2.2) \end{aligned}$$

where $H(Y|X)$ is the *conditional entropy*, defined as

Definition 2. Conditional Entropy

$$H(Y|X) \triangleq \mathbb{E}\left[\log \frac{1}{p(Y|X)}\right]$$

If X, Y are independent, $p(y|x) = p(y)$, we have $H(Y|X) = H(Y)$, and thus we recover

$$H(X, Y) = H(X) + H(Y),$$

the relationship derived in the last lecture. As expected, the entropy of two **independent** random variables X, Y is equal to the sum of their individual entropies.

Notation clarification - In the equation (2.1), all p's do not denote the same function. More precisely, the denominator of (2.1) should be written as $p_X(X)p_{Y|X}(Y|X)$. However, to avoid cumbersome notations, we will drop the subscript when the distribution being referred to is clear from the context.

Let's dig a little deeper into the definition of $H(Y|X)$

$$\begin{aligned} H(Y|X) &= \mathbb{E}\left[\log \frac{1}{p(Y|X)}\right] && (2.3)\\ &= \sum_{x,y} p(x,y) \log \frac{1}{p(y|x)} && (2.4)\\ &= \sum_x p(x) \sum_y p(y|x) \log \frac{1}{p(y|x)} \\ &\triangleq \sum_x p(x) H(Y|X=x) \end{aligned}$$

Here $H(Y|X = x)$ is the entropy of the random variable Y, conditional on the event $\{X = x\}$, and $H(Y|X)$ is the expected amount of extra information you get by observing Y, given that you have already observed X.

Chain rule for three random variables

Equation (2.2) is the chain rule for entropy for two random variables, but it can easily be extended to the case of multiple random variables. For three random variables X, Y, and Z, we have

$$\begin{aligned} H(X, Y, Z) &= H(X) + H(Y, Z|X) && (2.5)\\ &= H(X) + H(Y|X) + H(Z|X, Y) && (2.6) \end{aligned}$$

The first equality (2.5) is obtained by applying the chain rule for two variables (by considering (Y, Z) as a single random variable). The second equality can be derived as follows

$$\begin{aligned} H(Y, Z|X) &= \sum_x p(x) H(Y, Z|X=x) \\ \text{(chain rule condition on event } X=x) &= \sum_x p(x) H(Y|X=x) + \sum_x p(x) H(Z|Y, X=x) \\ &= H(Y|X) + H(Z|Y, X) \end{aligned}$$

□

The chain rule for entropy for n random variables is obtained by generalizing the above idea

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

2.4 Mutual Information

Given two random variables X and Y, we want to define a measure which quantifies the amount of information that observing Y provides about X without observing X. We call this measure **mutual information**, defined as

Definition 3. Mutual Information

$$I(X;Y) \triangleq H(X) - H(X|Y)$$

Alternate interpretation: Mutual information can also be interpreted as reduction in the entropy of X, when Y is observed.

At the first glance, expression (3) seems to be asymmetric in X, Y, however, expanding $H(X) - H(X|Y)$, we obtain

$$\begin{aligned} H(X) - H(X|Y) &= \mathbb{E}\left[\log \frac{1}{p(X)}\right] - \mathbb{E}\left[\log \frac{1}{p(X|Y)}\right] \\ &= \mathbb{E}\left[\log \frac{p(X|Y)}{p(X)}\right] \\ &= \mathbb{E}\left[\log \frac{p(X,Y)}{p(X)\,p(Y)}\right] \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{2.7}$$

From the expression (2.7), we see that the mutual information, $I(X;Y)$, is symmetrical with respect to X, Y and hence it is also equal to $H(Y) - H(Y|X)$. Thus we have

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

Now let's ask an interesting question: How much does X tell us about itself? In other words, what is $I(X;X)$? Using the definition (3), we have:

$$I(X;X) = H(X) - H(X|X)$$

Conditioned on X, X takes a fixed value. Therefore, $H(X|X) = \log 1 = 0$. Thus we have

$$I(X;X) = H(X)$$

This implies that X contains all the information about itself, which makes intuitive sense.

The relations among the entropy, relative entropy and mutual information can be visualized in the figure 2.1

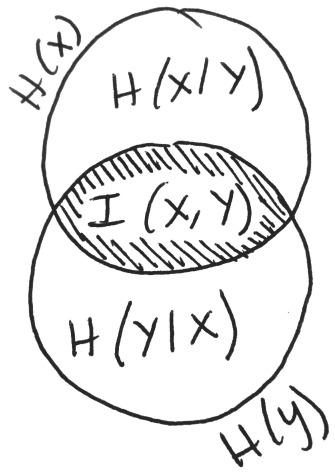

Figure 2.1

2.4.1 Chain Rule for Mutual Information

Consider three random variables: X, Y_1, and Y_2. Then the mutual information of these three random variables can be decomposed as:

$$I\left(X; Y_1, Y_2\right) = I\left(X; Y_1\right) + I\left(X; Y_2 | Y_1\right), \tag{2.8}$$

where $I\left(X; Y_1, Y_2\right)$ represents the amount of information Y_1 and Y_2 *together* give us about X, and $I\left(X; Y_2|Y_1\right)$ represents how much *more* information Y_2 gives us about X given that we already know Y_1.

Equation (2.8) can be derived by expressing mutual information in terms of entropies and then using the chain rule for entropy. This is left as an exercise to the reader. The chain rule (2.8) can be generalized for n random variables as follows

$$I(X; Y_1, Y_2, \cdots, Y_n) = I(X; Y_1) + I(X : Y_2|Y_1) + \cdots + I(X; Y_n|Y_1, Y_2, \cdots, Y_{n-1})$$