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

Lecture 2 — January 12

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

2.1 Outline

- Entropy, mutual information
- Relative entropy

Reading: CT 2.8, 3.1, 3.2, 4.1

2.2 Entropy

2.2.1 Recap

For a random variable X with probability mass function $p(x)$:

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

Here are some additional facts about entropy:

$$H(X) \geq 0$$
$$H(X) \text{ is label invariant.}$$

Notably, $H(X)$ is itself a random variable, because X is a random variable.

2.3 The Chain Rule for Entropy

Suppose we have two random variables, X and Y. They could be two flips of a coin, for example. If they are independent, the entropy just adds

$$H(X,Y) = H(X) + H(Y) \tag{2.1}$$

However, if X and Y are not independent, observing one might give some information about the other, so simply adding the informations is over counting. In this case, we have

$$\begin{aligned} H(X,Y) &= \mathbb{E}\left[\log \frac{1}{p(x)p(y|x)}\right] \qquad (2.2)\\ &= \mathbb{E}\left[\log \frac{1}{p(x)}\right] + \mathbb{E}\left[\log \frac{1}{p(y|x)}\right] \\ H(X,Y) &= H(X) + H(Y|X) \qquad (2.3) \end{aligned}$$

We define conditional entropy

Definition 1. Conditional Entropy

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

This definition is valid because conditioning Y on X moves us into a different probability space, but all of the normal concepts of probability still apply in this new space.

Another important note here is that, in equation (2.1) p are not all the same. More precisely, the denominator of (2.2) should read $p_X(x)p_{Y|X}(y|x)$. However, since that is a lot to write, we will just leave them all as p, and you can glean from context which exact distribution it refers to. As a sanity check, you can convince yourself that this is correct by recognizing that if Y is independent of X, then $p(y|x) = p(y)$, and equation (2.3) simplifies to equation (2.1).

2.3.1 More than two variables

Equation (2.3) is the chain rule for entropy for two random variables, but it can easily be extended to more variables. For example, for three random variables X, Y, and Z,

$$\begin{aligned} H(X,Y,Z) &= H(X) + H(Y,Z|X) \\ &= H(X) + H(Y|X) + H(Z|X,Y) \end{aligned} \tag{2.4}$$

The first step comes from directly applying the chain rule for two variables, but what about the second step? Let's dig a little deeper to make sure we understand what is going on there. Let's start by looking at $H(Y|X)$.

$$\begin{aligned} H(Y|X) &= \mathbb{E}\left[\log \frac{1}{p(y|x)}\right] \\ &= \sum_{x,y} p(x,y) \log \frac{1}{p(y|x)} \end{aligned} \tag{2.5}$$

Note that the probability distributions for the expectation and in the function itself are not the same! If you are unhappy with that, just remember that for any arbitrary function, $\mathbb{E}\left[f(X,Y)\right] = \sum_{x,y} p(x,y) f(x,y)$, and in this case, that arbitrary function is $\log \frac{1}{p(y|x)}$.

Okay, so let's go back to (2.5)

$$\begin{aligned} H(Y|X) &= \sum_{x,y} p(x,y) \log \frac{1}{p(y|x)} \\ &= \sum_{x} p(x) \sum_{y} p(y|x) \log \frac{1}{p(y|x)} \\ &= \sum_{x} p(x) H(Y|X=x) \end{aligned}$$

This is a measure of, on average, how much extra information you get by observing a second variable Y, given that you have already observed the first variable X.

We did this because we wanted to verify the chain rule extension in (2.4), namely that $H(Y,Z|X) = H(Y|X) + H(Z|X,Y)$. Following the same steps as before, but keeping everything conditioned on X, we can write

$$\begin{aligned} H(Y,Z|X) &= \sum_x p(x) H(Y,Z|X=x) \\ &= \sum_x p(x) H(Y|X=x) + \sum_x p(x) H(Z|Y,X=x) \\ &= H(Y|X) + H(Z|Y,X) \end{aligned}$$

verifying equation (2.4). Note that this is equivalent to the original two variable chain rule given in equation (2.3), $H(Y,Z) = H(Y) + H(Z|Y)$, except that everything is now conditioned on X. As we touched on before, conditioning on an event creates a new probability space where all the same concepts of probability apply. We simply added the same conditioning event to all three terms!

Now let's go back to equation (2.4), the chain rule for entropy.

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

You can interpret these terms as how much additional information you gain from observing the variable, given that you may have already observed some other variables. For the first term, since you haven't observed anything yet, all the information is new. The next term is the additional information you gained by observing Y, given that you already observed X, and so on. This makes some intuitive sense, and suggests a flow of sorts: we observe a series of events, and each of them tells us a little bit more about the world. Chain rules like this are important because we often encounter long chains of random variables, not just one or two!

2.4 Mutual Information

Given two random variables X and Y, we want to define a measure of the information that Y provides about X when Y is observed, but X is not. We call this measure **mutual information**, which is defined as:

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

Which can be intuitively understood as the information that Y provides about X. At first glance, this expression seems to be asymmetric, but we will show that in fact:

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

Expanding $H(X) - H(X|Y)$, we have:

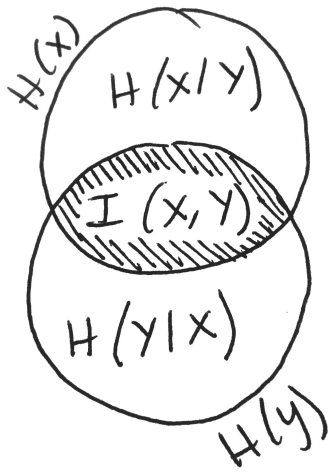
$$\begin{aligned} H(X) - H(X|Y) &= \mathbb{E}\left[\log \frac{1}{p(X)}\right] - \mathbb{E}\left[\log \frac{1}{p(X|Y)}\right] \\ &= \mathbb{E}\left[\log \frac{p(X|Y)}{p(X)}\right] \\ &= \mathbb{E}\left[\log \frac{p(X|Y)\,p(Y)}{p(X)\,p(Y)}\right] \\ &= \mathbb{E}\left[\log \frac{p(X,Y)}{p(X)\,p(Y)}\right]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

It is for this reason that we call this quantity *mutual* information—because $I(X;Y)$ does not "prefer" X or Y. Formally, mutual information is defined as:

Definition 2. Mutual Information

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

The meaure of entropy, relative entropy and mutual information can be visualized in the figure belownora



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

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

We note that $H(X|X) = 0$, because in the expectation, X can only take on one fixed, given value with probability 1. Therefore, $H(X|X) = \log 1 = 0$. Thus:

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

Meaning that X tells us *everything* about itself!

2.4.1 Chain Rule for Mutual Information

Let's say that we have three random variables: X, Y_1, and Y_2. Then the mutual information of these three random variables can be decomposed as:

$$I(X;Y_1,Y_2) = I(X;Y_1) + I(X;Y_2|Y_1) \tag{2.6}$$

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

Note: Equation (2.6) can be derived be expressing mutual information in terms of entropies and then using the chain rule for entropy. This is left as en exercise to the reader.