Information Theory

Outline

1. Entropy
2. Mutual Information
3. Application to communication complexity.

Entropy

Consider a discrete random variable X. Entropy is a measure of the amount of randomness in X. We will see that the following ideas are equivalent and are appropriate definitions of entropy:

1. The amount of randomness in X in bits.
2. Random bits needed to generate a draw from X on average.
3. Bits needed to store a draw from X on average.
4. Bits needed to communicate one draw from X on average.
5. Yes/no questions needed to guess a draw from X on average.

We first give the mathematical definition of entropy, and then explore its relation to the intuitive definitions above through examples.

Definition: Entropy is defined as

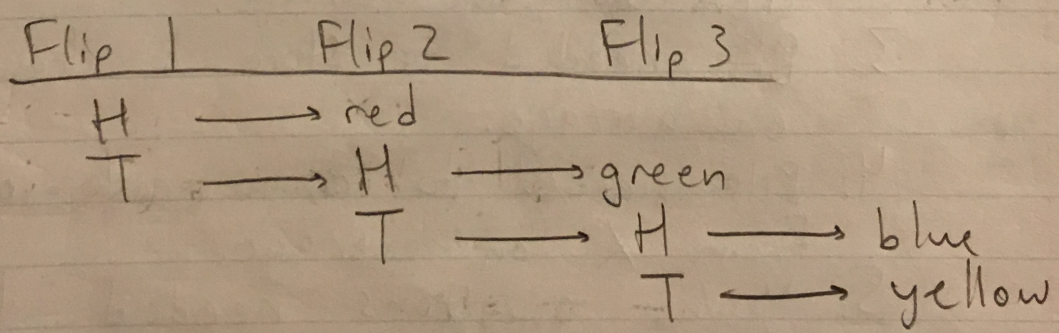
$$H(X) = \sum_{x \in \text{range}(X)} P_X(x) \cdot \log_2 \frac{1}{P_X(x)},$$

where $P_X(x) = \Pr[X = x]$.

Example: Define X as follows:

$$X = \begin{cases} \text{red} & \text{with probability } \frac{1}{2} \\ \text{green} & \text{" } \text{" } \frac{1}{4} \\ \text{blue} & \text{" } \text{" } \frac{1}{8} \\ \text{yellow} & \text{" } \text{" } \frac{1}{8} \end{cases}$$

Consider perspective 2 above. We will give a program to generate a draw of X using random bits, or coin flips. The following program yields the desired probabilities:



This program is efficient: we use only as many flips as are necessary. The average number of flips needed to generate a draw from X is:

$$F(X) = \frac{1}{2}(1 \text{ flip}) + \frac{1}{4}(2 \text{ flips}) + \frac{1}{8}(3 \text{ flips}) + \frac{1}{8}(3 \text{ flips})$$

$$= \frac{7}{4} \text{ flips}$$

$$= H(X).$$

Next, consider perspective 3. We can store the result of a draw from X using the sequence of coin flips that generated the draw:

red $\Leftrightarrow$ H
green $\Leftrightarrow$ TH
blue $\Leftrightarrow$ TTH
yellow $\Leftrightarrow$ TTT

Here, the average number of bits needed store a draw is

$$S(X) = \frac{1}{2}(1 \text{ bit}) + \frac{1}{4}(2 \text{ bits}) + \frac{1}{8}(3 \text{ bits}) + \frac{1}{8}(3 \text{ bits})$$

$$= \frac{7}{4} \text{ bits}$$

$$= H(X).$$

This is also the number of bits needed to communicate a draw from X.

Consider perspective 5. We could ask

1. Is x red?
2. Is x green?
3. Is x blue?

The number of questions needed to determine x is on average

$$Q(X) = \frac{1}{2}(1 \text{ question}) + \frac{1}{4}(2 \text{ questions}) + \frac{1}{8}(3 \text{ questions}) + \frac{1}{8}(3 \text{ questions})$$

$$= \frac{7}{4} \text{ questions}$$

$$= H(X).$$

In each of the above cases, we could have picked a different scheme. For example, the following series of questions also allows us to determine X:

1. Is x blue?
2. Is x green?
3. Is x red?

However, the average number of questions needed is $\frac{5}{2}$. Entropy reflects the most efficient scheme. This is achieved by ordering questions

in a way that maximizes information gain, or reduction in entropy, at each step.

Example: Define the random variable X as follows:

$$X = \begin{cases} a & \text{with probability } \frac{1}{2} \\ b & " \quad " \quad \frac{1}{3} \\ c & " \quad " \quad \frac{1}{6}. \end{cases}$$

By the mathematical definition of entropy:

$$H(X) = \frac{1}{2}\log_2 2 + \frac{1}{3}\log_2 3 + \frac{1}{6}\log_2 6 \approx 1.46$$

However, $\log_2 3$ and $\log_2 6$ are not integers. If we came up with a scheme for storing a draw from X as in the previous example —

$a \Leftrightarrow H$
$b \Leftrightarrow TH$
$c \Leftrightarrow TT$

we would use $1.5 > H(X)$ bits. (You may ask, why not use the encoding

$a \Leftrightarrow H$
$b \Leftrightarrow T$
$c \Leftrightarrow HT$.

Such an encoding is ambiguous when storing multiple draws. Under the first scheme,

HTHTT

corresponds uniquely to "abc". Under the second, the string could correspond either to "abcb" or "ababb". To disambiguate, we would need a delimiter bit after one of the encodings, thus reintroducing the bits we tried to save.)

In general, while we may not be able to generate, store, or communicate a draw from X using exactly $H(X)$ bits on average, we can do it using the following number of bits:

$$\sum_{x \in \text{range}(X)} P_X(x) \cdot \left\lceil \log_2 \frac{1}{P_X(x)} \right\rceil \leq \sum_{x \in \text{range}(X)} P_X(x) \cdot \left(\log_2 \frac{1}{P_X(x)} + 1 \right)$$

$$= H(X) + \sum_{x \in \text{range}(X)} P_X(x)$$

$$= H(X) + 1.$$

The optimal encoding uses between $H(X)$ and $H(X)+1$ bits on average.

Joint Entropy

Consider two random variables, X and Y, concatenated together. We write the concatenation as XY for simplicity. The range of XY is $\text{range}(X) \times \text{range}(Y)$.

Theorem: If X and Y are independent,

$$H(XY) = H(X) + H(Y).$$

Proof:

$$H(XY) = \sum_{x,y \in \text{range}(XY)} P_{XY}(x,y) \log_2\left(\frac{1}{P_{XY}(x,y)}\right)$$

$$= \sum_x \sum_y P_X(x) P_Y(y) \log_2\left(\frac{1}{P_X(x) P_Y(y)}\right)$$

(by independence of X and Y)

$$= \sum_x \sum_y P_X(x) P_Y(y) \left[\log_2\left(\frac{1}{P_X(x)}\right) + \log\left(\frac{1}{P_Y(y)}\right)\right]$$

$$= \sum_x P_X(x) \log_2\left(\frac{1}{P_X(x)}\right) \sum_y P_Y(y)$$

$$+ \sum_y P_Y(y) \log_2\left(\frac{1}{P_Y(y)}\right) \sum_x P_X(x)$$

$$= H(X) \cdot 1 + H(Y) \cdot 1$$

$$= H(X) + H(Y). \qquad \square$$

Recall that we can encode a draw $x \in X$ using at most $H(X)+1$ bits on average. Consider the scenario where Alice sends n draws from X to Bob. If she communicates the draws individually, she will use

$$n(H(X)+1) = nH(X)+n \text{ bits.}$$

However, if we treat the n copies of X as one random variable, we should be able to transmit n draws from X using

$$H(\underbrace{XX\ldots X}_{n \text{ copies}})+1 = nH(X)+1 \text{ bits}$$

by the previous theorem. This discrepancy can be seen in the following example.

Example: Let X be the uniform distribution over $\{0,1,2\}$.

The most efficient encoding of one draw from X is as follows

$$0 \Leftrightarrow 0$$
$$1 \Leftrightarrow 01$$
$$2 \Leftrightarrow 11$$

This encoding uses on average

$$S(X) = \frac{1}{3}(1 \text{ bit}) + \frac{1}{3}(2 \text{ bits}) + \frac{1}{3}(2 \text{ bits})$$

$$= \frac{5}{3} \text{ bits}$$
$$\approx 1.67 \text{ bits}.$$

However,

$$H(X) = 3\left(\frac{1}{3}\log_2(3)\right)$$

$$= \log_2(3)$$
$$\approx 1.58.$$

We can interpret X as a ternary bit, and multiple draws from X as a ternary number. If we encode $Z = \underbrace{XX\ldots X}_{n \text{ copies}}$ using binary, we use at most

$$\lceil \log_2 3^n \rceil \text{ bits} \leq n \log_2 3 + 1 \text{ bits}.$$

$$= H(\underbrace{XX\ldots X}_{n}) + 1 \text{ bits}.$$

For large n, this encoding is more efficient than sending each draw from X individually using our naive encoding.

In general, for n symbols, the amortized cost per symbol on average is at most

$$\frac{1}{n}(nH(X)+1) = H(X)+\frac{1}{n}.$$

As $n \to \infty$,

$$\lim_{n\to\infty} H(X)+\frac{1}{n} = H(X).$$

Mutual Information

Consider random variables X and Y. Here we allow X and Y to be dependent rather than requiring independence. For dependent random variables, we would expect $H(XY)$ to be less than $H(X)+H(Y)$, since X and Y share information. We have the following definition:

Definition: The mutual information between X and Y is

$$I(X;Y) = H(X) + H(Y) - H(XY).$$

We can use the definition of entropy to obtain an equation for $I(X;Y)$ in terms of the probability distributions of X, Y, and XY:

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(XY) \\ &= \sum_{x,y} p_{xy}(x,y) \left(\log_2\left(\frac{1}{p_x(x)}\right) + \log_2\left(\frac{1}{p_y(y)}\right) - \log_2\left(\frac{1}{p_{xy}(x,y)}\right) \right) \\ &= \sum_{x,y} p_{xy}(x,y) \log_2\left(\frac{p_{xy}(x,y)}{p_x(x)\,p_y(y)}\right). \end{aligned}$$

If X and Y are independent, then

$p_{XY}(x,y) = p_X(x)\,p_Y(y)$, so the log factor vanishes, and $I(X;Y) = 0$, as expected.

Example: Consider the following joint distribution:

x	y	$p_{XY}(x,y)$
0	0	1/2
1	1	1/16
1	2	1/16
1	3	1/16
1	4	1/16
2	1	1/16
2	2	1/16
2	3	1/16
2	4	1/16

Then

$$H(X) = \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{4}\cdot 2$$
$$= \frac{3}{2}$$

$$H(Y) = \frac{1}{2}\cdot 1 + \sum_{i=1}^{4} \frac{1}{8}\cdot 3$$
$$= 2.$$

So $H(X) + H(Y) = \frac{7}{2}$. But

$$H(XY) = \frac{1}{2}\cdot 1 + \sum_{i=1}^{8} \frac{1}{16}\cdot 4$$
$$= \frac{5}{2}.$$

Thus,

$$I(X;Y) = \frac{7}{2} - \frac{5}{2} = 1.$$

This means that we save one bit when generating X and Y together. Or, on average, we learn one bit's worth of information about Y by learning X, and vice versa.

In general, we have the following bounds on mutual information:

$$0 \leq I(X;Y) \leq H(X) + H(Y).$$

The lower bound is an equality when X and Y are independent. The upper-bound is an equality when XY is constant (so $H(X,Y) = 0$).

There is actually a better upper bound:

$$I(X;Y) \leq \min\{H(X), H(Y)\}.$$

The intuition behind this bound is that the most information you can save when generating XY is the information in either variable. If X uniquely determines Y, then $I(X;Y) = H(Y)$, for example.

Example: Let X be n independent bits and Y be the parity of X.

Then X uniquely determines Y. If we generate X, we can generate Y with no additional random bits.

$$H(X) = n$$
$$H(Y) = 1$$
$$H(XY) = n$$

$$I(X;Y) = H(X) + H(Y) - H(XY)$$
$$= 1.$$

Conditional entropy

Definition: The entropy of Y conditioned on X is

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

The conditional entropy is the remaining information in Y if we have already mutual information

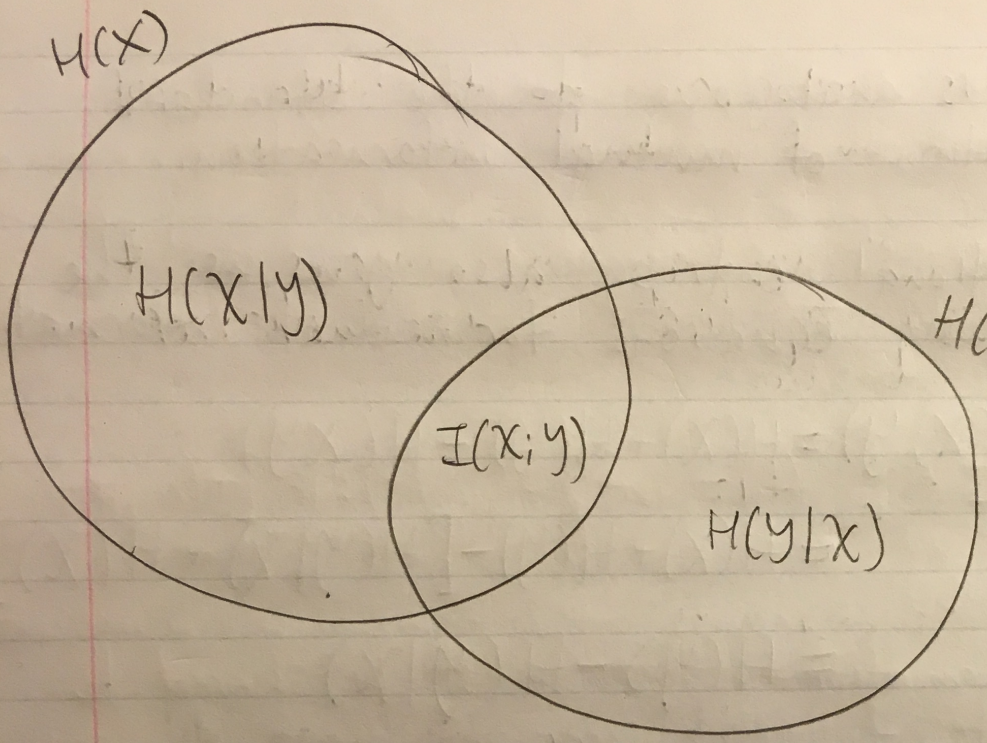


Figure: Conditional Entropy.

The definition of conditional entropy gives us

$$H(Y|X) = H(XY) - H(X)$$

or

$$H(XY) = H(Y|X) + H(X)$$

We can similarly define conditional mutual information. We wish to know the mutual information between X and Y, given that we have information about another random variable Z:

Definition: The conditional mutual information of X, Y given Z is

$$I(X,Y|Z) = H(X|Z) + H(Y|Z) - H(XY|Z)$$

This is analogous to the standard definition of mutual information.

Conditional entropy also gives us the following equations for mutual information:

$$I(X;Y) = H(X) + H(Y) - H(XY)$$
$$= H(X) + H(Y) - [H(Y|X) + H(X)]$$
$$= H(Y) - H(Y|X)$$

Similarly,

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

Application in communication complexity

We can use information theory to prove bounds on communication complexity. For example, for EQ, the equality problem, if Alice and Bob have independently chosen strings from a distribution X, then Alice must send at least $H(X)$ bits to Bob for Bob to verify equality.

We can think of distributional communication complexity as computing a function $f(x,y)$ over the distributions X and Y.

Definition: Let π be a deterministic protocol. The information cost is

$$IC_{\mu_{XY}}(\pi) = I(\pi; X|Y) + I(\pi; Y|X),$$

where μ_{XY} is the joint distribution of XY.

Intuitively, the information cost is just the information that must be exchanged by Alice and Bob under protocol π.

Alice or Bob should then be able to use the exchanged information to compute

$f(x,y)$ with high probability. To determine the minimum information needed to compute $f(x,y)$ with high probability, we take the infimum over all protocols π that satisfy our desired error bound:

$$IC^{\varepsilon}_{\mu_{xy}}(f) = \inf\left\{IC_{\mu_{xy}}(\pi) : \Pr_{(x,y)\sim\mu_{xy}}[\pi(x,y)\neq f(x,y)] \leq \varepsilon\right\}.$$

Since this represents the minimum amount of information that must be shared to compute f with error at most ε and inputs drawn from μ_{xy}, we have that

$$D^{\varepsilon}_{\mu}(f) \geq IC^{\varepsilon}_{\mu}(f).$$