Entropy and Mutual Information

Erik G. Learned-Miller
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA 01003

September 16, 2013

Abstract

This document is an introduction to entropy and mutual information for discrete random variables. It gives their definitions in terms of probabilities, and a few simple examples.

1 Entropy

The *entropy* of a random variable is a function which attempts to characterize the "unpredictability" of a random variable. Consider a random variable X representing the number that comes up on a roulette wheel and a random variable Y representing the number that comes up on a fair 6-sided die. The entropy of X is greater than the entropy of Y. In addition to the numbers 1 through 6, the values on the roulette wheel can take on the values 7 through 36. In some sense, it is less predictable.

But entropy is not just about the number of possible outcomes. It is also about their frequency. For example, let Z be the outcome of a weighted six-sided die that comes up 90% of the time as a "2". Z has lower entropy than Y representing a fair 6-sided die. The weighted die is less unpredictable, in some sense.

But entropy is not a vague concept. It has a precise mathematical definition. In particular, if a random variable X takes on values in a set $\mathcal{X} = \{x_1, x_2, ..., x_n\}$, and is defined by a probability distribution $P(X)$, then we will write the entropy of the random variable as

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \log P(x). \tag{1}$$

We may also write this as

$$H(P(x)) \equiv H(P) \equiv H(X).$$

If the log in the above equation is taken to be to the base 2, then the entropy is expressed in *bits*. If the log is taken to be the natural log, then the entropy is expressed in *nats*. More commonly, entropy is expressed in bits, and unless otherwise noted, we will assume a logarithm with base 2.

Example 1. To compute the entropy of a fair coin, we first define its distribution:

$$P(X = heads) = \frac{1}{2} \qquad P(X = tails) = \frac{1}{2}.$$

Using Equation (1), we have:

$$\begin{aligned} H(P) &= -\sum_{x \in \{heads, tails\}} P(x) \log P(x) && (2)\\ &= -\left[\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right] && (3)\\ &= -\left[-\frac{1}{2} + -\frac{1}{2}\right] && (4)\\ &= 1. && (5) \end{aligned}$$

Example 2. Let X be an unfair 6-sided die with probability distribution defined by $P(X = 1) = \frac{1}{2}$, $P(X = 2) = \frac{1}{4}$, $P(X = 3) = 0$, $P(X = 4) = 0$,

$P(X=5)=\frac{1}{8}$, and $P(X=6)=\frac{1}{8}$. The entropy is

$$\begin{aligned} H(P) &= -\sum_{x\in\{1,2,3,4,5,6\}} P(x)\log P(x) && (6)\\ &= -\left[\frac{1}{2}\log\frac{1}{2}+\frac{1}{4}\log\frac{1}{4}+0\log 0+0\log 0+\frac{1}{8}\log\frac{1}{8}+\frac{1}{8}\log\frac{1}{8}\right] && (7)\\ &= -\left[-\frac{1}{2}+-\frac{1}{2}+0+0+-\frac{3}{8}+-\frac{3}{8}\right] && (8)\\ &= 1.75. && (9)\end{aligned}$$

Notice that we have used $0\log 0=0$. The justification for this is that the limit of $x\log x$ as x becomes small is 0.

2 Joint Entropy

Joint entropy is the entropy of a joint probability distribution, or a multi-valued random variable. For example, one might wish to the know the joint entropy of a distribution of people defined by hair color C and eye color E, where C can take on 4 different values from a set $\mathcal{C}$ and E can take on 3 values from a set $\mathcal{E}$. If $P(E,C)$ defines the joint probability distribution of hair color and eye color, then we write that their joint entropy is:

$$H(E,C)\equiv H(P(E,C))=-\sum_{e\in\mathcal{E}}\sum_{c\in\mathcal{C}}P(e,c)\log P(e,c). \tag{10}$$

In other words, joint entropy is really no different than regular entropy. We merely have to compute Equation (1) over all possible pairs of the two random variables.

Example 3. Let X represent whether it is sunny or rainy in a particular town on a given day. Let Y represent whether it is above 70 degrees or below seventy degrees. Compute the entropy of the joint distribution $P(X,Y)$ given by

$$P(sunny,hot)=\frac{1}{2} \tag{11}$$
$$P(sunny,cool)=\frac{1}{4} \tag{12}$$
$$P(rainy,hot)=\frac{1}{4} \tag{13}$$
$$P(rainy,cool)=0. \tag{14}$$

Using Equation (10), or Equation (1), we obtain

$$H(X,Y) = -\left[\frac{1}{2}\log\frac{1}{2}+\frac{1}{4}\log\frac{1}{4}+\frac{1}{4}\log\frac{1}{4}+0\log 0\right] \tag{15}$$

$$= -\left[-\frac{1}{2} + -\frac{1}{2} + -\frac{1}{2} + 0\right] \tag{16}$$

$$= \frac{3}{2}. \tag{17}$$

3 Mutual Information

Mutual information is a quantity that measures a relationship between two random variables that are sampled simultaneously. In particular, it measures how much information is communicated, on average, in one random variable about another. Intuitively, one might ask, how much does one random variable tell me about another?

For example, suppose X represents the roll of a fair 6-sided die, and Y represents whether the roll is even (0 if even, 1 if odd). Clearly, the value of Y tells us something about the value of X and vice versa. That is, these variables share *mutual information*.

On the other hand, if X represents the roll of one fair die, and Z represents the roll of another fair die, then X and Z share no mutual information. The roll of one die does not contain any information about the outcome of the other die. An important theorem from information theory says that the mutual information between two variables is 0 if and only if the two variables are *statistically independent*.

The formal definition of the mutual information of two random variables X and Y, whose joint distribution is defined by $P(X, Y)$ is given by

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}.$$

In this definition, $P(X)$ and $P(Y)$ are the *marginal distributions* of X and Y obtained through the marginalization process described in the Probability Review document.