Popcast™ Digital Pop

Informational Entropy

motionworks

A Brief Discussion of Informational Entropy

Entropy is an abstract principle that is often thought of as a measure of "randomness" or "chaos". While this is a helpful way of interpreting entropy, it is not the only interpretation. Here, entropy will be discussed in terms of information theory, which includes the interpretation of entropy as measuring the "amount of information" present. For completeness, entropy is also interpreted in information theory as a measure of "surprise" or "uncertainty," but this discussion will heuristically focus on the "amount of information" interpretation.

First, a technical definition of entropy is given, which is then unpacked using two simple, demonstrative examples. Next, the principle of maximum entropy is introduced in order to understand how entropy maximization algorithms use relative entropy to build a digital population.

What is Entropy?

Consider a system with distribution X. The entropy of X is defined in words as a measure of the amount of information in X; however, the mathematical definition is much more useful. In mathematics, the entropy $H(X)$ of the distribution X is defined as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \ln P(x_i)$$

with the probability of event x_i occurring $P(x_i)$; the natural log operator ln; and the sigma operator $\sum_{i=1}^{n}$, which denotes taking

the sum of all n possible states. To see how the mathematical definition is useful, consider the entropy of a fair-coin toss and a fair-die roll.

A fair coin is one that when flipped has equal probability of landing heads up as it does heads down. In terms of the mathematical definition, the two possible events with corresponding probabilities are given in the table below.

Event	Probability of Event
$x_1 \equiv$ heads up	$P(x_1) = \frac{1}{2}$
$x_2 \equiv$ tails up	$P(x_2) = \frac{1}{2}$

The entropy of a fair-coin toss X is therefore

$$H(X) = -\left[\left[\frac{1}{2}\ln\frac{1}{2}\right] + \left[\frac{1}{2}\ln\frac{1}{2}\right]\right] = \ln 2 \approx 0.693$$

.

A fair die is one that when rolled has an equal probability of landing on any of the sides. For this example, the canonical six-sided, fair die is used with events given below.

Event	Probability of Event
$x_1 \equiv$ roll a 1	$P(x_1) = \frac{1}{6}$
$x_2 \equiv$ roll a 2	$P(x_2) = \frac{1}{6}$
$x_3 \equiv$ roll a 3	$P(x_3) = \frac{1}{6}$
$x_4 \equiv$ roll a 4	$P(x_4) = \frac{1}{6}$
$x_5 \equiv$ roll a 5	$P(x_5) = \frac{1}{6}$
$x_6 \equiv$ roll a 6	$P(x_6) = \frac{1}{6}$

Following the same procedure as before, the corresponding entropy of a fair-die roll Y is

$$H(Y) = \ln 6 \approx 1.792$$

What does it mean, though, for a fair-coin toss to have an entropy of $H(X) \approx 0.693$ and a fair-die roll to have an entropy of $H(Y) \approx 1.792$? Furthermore, why does the coin toss have

less entropy, and what does that mean? These questions are best answered by comparing the two calculated entropies.

First, since $H(Y) > H(X)$, the state of a fair-die roll holds more information than the state of a fair-coin toss. Put differently, if a coin is tossed heads up, it is immediately known that the coin is also tails down—"everything" is known about the coin. If a fair-die rolls a one and is oriented so that each of the sides face one of the cardinal directions, it is clear that the face-down number is six; however, it is not clear which number is facing North. In other words, more information is needed to know "everything" about the die roll. (Specifically, to fully describe the state of the system, the number facing North is also required—then "everything" is known about the die roll.) Therefore, describing the state of a fair-die roll requires more information than a fair-coin toss.

Note, this does not imply that one is necessarily "better" than the other. To make such a judgment would require more information regarding why the entropy is being calculated in the first place. For instance, if the smallest unit of information is sought, the fair-coin would be judged "better" than the fair-die. The following section discusses another way entropy can also be used to make judgment calls when invoking the principle of maximum entropy.

Principle of Maximum Entropy

To understand when it is appropriate to use entropy as a tool for choosing the better of multiple distributions, consider now the case of two candidate distributions A and B that both describe the same target distribution C with corresponding entropies $H(A)$, $H(B)$, and $H(C)$. The principle of maximum entropy suggests that if $H(A) > H(B)$, A has more information about the target distribution C and is therefore the better of the two candidates. Therefore, if a large set of candidate distributions can be produced, the best candidate from the set can be chosen. This is the foundational concept that entropy maximization algorithms (EMA) exploit; however, many EMAs actually use relative entropy as an objective function and attempt to minimize the objective function.

Relative Entropy

Consider again the candidate distributions A and B and target distribution C from above. The relative entropy $D(X||C)$ between candidate distribution X and target distribution C is given by

$$D(X||C) = H(C) - H(X)$$

If $D(A||C) < D(B||C)$, distribution A is the ideal candidate. Note, $D(A||C) < D(B||C)$ and $H(A) > H(B)$ are equivalent statements, because

$$\begin{aligned} D(A||C) < D(B||C) \rightarrow & H(C) - H(A) < H(C) - H(B) \\ \rightarrow & -H(A) < -H(B) \\ \rightarrow & H(A) > H(B). \end{aligned}$$

If $D(A||C) = 0$, then $A = C$.