Differential entropy

Differential entropy is the name given to Shannon's entropy for continuous probability distributions; the definition is obvious:

$$h(X) = -\int dx p(x) \log_2 p(x) \tag{1}$$

In other words the only change is to replace the sum over the discrete variable with an integral over the continuous one. We will see later however that the relationship between the two definitions, the one we saw before for discrete distributions and this new definition for continuous distributions, isn't as straight forward as you might hope and so it makes sense to use a different symbol, in this case a small h for the differential entropy.

Lets consider the uniform distribution:

$$p(x) = \begin{cases} 1/a & x \in [0, a] \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

It is easy to calculate the entropy to find

$$h(X) = \log a \tag{3}$$

This immediately demonstrates a difference between differential entropy and the entropy we looked at before for the discrete variable: if $a < 1$ then $h(X) < 0$ and so the differential entropy isn't always positive.

In fact, our previous notion of entropy, based on the source coding theorem, that it quantifies the amount of information in a signal, does not work here since a real number, if written to infinite precision, contains infinite information. Of course, in practice, if a real value's signal is used to communicate information there will be imprecision in both the encode and decoding, limiting the amount of information that can be carried by signal from the person encoding the data to the person decoding it and so it is still useful to study continuous signals using the tools of information theory.

After the uniform distribution the first continuous distribution you think of is probably the Gaußian distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2}\right) \tag{4}$$

Working out $h(X)$ is straightforward once you substitute and use integration by parts you find

$$h(X) = \frac{1}{2} \log 2\pi e\sigma^2 \tag{5}$$

where the e is just the exponential $\exp(1)$. As with the uniform distribution, this formula can give a positive or negative number depending on the size of σ. Interestingly it can be proved that for fixed variance the Gaußian has the highest entropy.

Relationship between the continuous and discrete entropy

We have mentioned, even laboured, the idea that the differential entropy is not the same as the entropy; in this section we will explicitly work out the difference. Consider a probability density $p(x)$; we can use this to define a discrete random variable by discretizing the support; let

$$y_n = [x_{n-1}, x_n) \tag{6}$$

be a subinterval of width $\delta t = x_n - x_{n-1}$ and assign to y_n the probability

$$q_n = \int_{x_{n-1}}^{x_n} p(x)dx \tag{7}$$

so the random variable $X^{\delta x}$ whose outcomes are the subintervals y_n has entropy

$$H(X^{\delta x}) = -\sum_n q_n \log_2 q_n \tag{8}$$

Now we want to replace this with the integral and to do this we need to do something with the q_n; morally

$$q_n = \int_{x_{n-1}}^{x_n} p(x)dx \approx p(\bar{x}_n)\delta x \tag{9}$$

where $\bar{x}_n$ is any point in $[x_{n-1}, x_n)$; as the subinterval gets smaller and smaller this should become a better and better approximation because $p(x)$ gets closer and closer to being constant in the narrower and narrower interval. In fact, although that's how we understand this process, this intuition is not a basis for proving theorems, so we do something more elegant.

Assuming $p(x)$ is continuous we know we can always pick a value of $\bar{x}_n \in [x_{n-1}, x_n]$ such that

$$p(\bar{x}_n)\delta x = \int_{x_{n-1}}^{x_n} p(x)dx \tag{10}$$

exactly. This is an appliation of the Mean Value Theorem which says that for a continous function over an interval, there is point in the interval so that the value of the function is equal to its average. Roughly speaking: in general the average is neither the lowest nor the highest value that function takes over the interval, so taking $f(x)$ as an example function with average value $\bar{f}$ we can pick a point, say a such that $f(a) < \bar{f}$ and another point b such that $f(b) > \bar{f}$. Since the function is continuous as we move from a to b we must at some point pass through a point where $f(x) = \bar{f}$. There may be more than one such point, but the idea here is that there is at least one, and in our application $\bar{x}_n$ is any such point.

If we substitute this into the formula for $H(X^{\delta x})$ we had back in Eq. 9 we get

$$H(X^{\delta x}) = -\sum_n p(\bar{x}_n)\delta x \log_2 p(\bar{x}_n)\delta x \tag{11}$$

Using the law of logs we have

$$H(X^{\delta x}) = -\sum_n \left[p(\bar{x}_n)\log_2 p(\bar{x}_n)\right]\delta x - \sum_n p(\bar{x}_n)\delta x \log \delta x \tag{12}$$

Now, because

$$\int p(x)dx = 1 \tag{13}$$

and

$$p(\bar{x}_n)\delta x = \int_{x_{n-1}}^{x_n} p(x)dx \tag{14}$$

we know

$$\sum_n p(\bar{x}_n)\delta x = 1 \tag{15}$$

Finally we are accustomed to the Riemann approximation

$$\sum_n f(\tilde{x}_n)\delta x \to \int f(x)dx \tag{16}$$

as $\delta x \to 0$ where $\tilde{x}_n \in [x_{n-1}, x_n)$ so we can see that

$$H(X^{\delta x}) + \log_2 \delta x \to h(X) \tag{17}$$

as $\delta x \to 0$.

Thus, if you consider a continuous random variable as a limit of discrete distributions, the entropy will go to infinity because the number of values you are summing over becomes infinite; however, the definition of the differential entropy has an extra term, the $\log \delta x$ which cancels this effect. The approximate formula is also instructive:

$$h(X) \approx H(X^{\delta x}) + \log_2 \delta x \tag{18}$$

If the support is finite, say of length L, with $N = L/\delta x$ the number of subintervals, then, approximately, the differential entropy has two components: the entropy of the N-bit encoding of the result and another term corresponding to what remains after the N-bit encoding.

Changes of variable

The probability density is a density: what this means is that it changes under a change of variable. This is easiest to see if go back to the definition:

$$P(x \in [x_0, x_1)) = \int_{x_0}^{x_1} p_X(x) dx \tag{19}$$

Now what happens if we do a change of variable to $y(x)$; lets ignore any issues with multivaluedness of the inverse or whatever and assume y is a strictly monotonic function of x so

$$dx = \left| \frac{dx}{dy} \right| dy \tag{20}$$

so if $y_0 = y(x_0)$ and $y_1 = y(x_1)$ we have

$$P(y \in [y_0, y_1)) = \int_{y_0}^{y_1} p_X(x(y)) \frac{dy}{|dx/dy|} \tag{21}$$

where we've inverted the function $y(x)$ to work out which value of p_X to use for a given y. Now changing variables shouldn't change the actual probability, we should have

$$P(y \in [y_0, y_1)) = P(x \in [x_0, x_1)) \tag{22}$$

and, of course if $p_Y(y)$ is the probability density for y we should have

$$P(y \in [y_0, y_1)) = \int_{y_0}^{y_1} p_Y(y) dy \tag{23}$$

Since these things should be true for all intervals we have

$$p_Y(y) = \frac{p_X(x(y))}{|dx/dy|} \tag{24}$$

Thus, the probability density function changes when we change variables; in fact this behaviour is the definition of a density.

This means that the differential entropy is not invariant under a change of variable; substituting in the formula above we can show

$$h(Y) = h(X) + \int p_X(x) \log_2 \left| \frac{dx}{dy} \right| dx \tag{25}$$

In fact this is a general statement of something we have already seen a specific example of; using $y = ax$ in this formula shows us that

$$h(aX) = h(X) + \log |a| \tag{26}$$

so the differential entropy is not invariant under scaling; this is something we observed when we calculated the differential entropy for the uniform distribution and the Gaußian.

The differential mutual information

The differential mutual information is

$$I(X, Y) = \int p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} \tag{27}$$

and satisfies the same identities as the mutual information

$$I(X, Y) = h(X) + h(Y) - h(X, Y) \tag{28}$$

It has some advantages over the differential entropy; in particular it is invariant under changes of variable in X and Y. We won't prove that here because we have only discussed the one dimensional differential entropy; in

higher dimensions the role of the derivative factor is played by the Jacobian. You can, however, see how the invariance works: the extra Jacobian factor that appears because of the change of variables cancels, or put another way, the fraction inside the log is not a density, it is a ratio of densities and the Jacobian factors cancel to give an ordinary function.

In fact, the differential mutual information is the same as the mutual information in the sense that, using the notation above

$$I(X^{\delta x}, Y^{\delta y}) \to I(X, Y) \tag{29}$$

as δx and δy approach zero. Furthermore

$$I(X, Y) \geq 0 \tag{30}$$

Applications of differential entropy

One of the glories of the entropy was its relationship to communication through the source coding theorem. At first it might appear that there is no analogous set of ideas for differential entropy since, in a sense, the amount of information encoded in the outcome of a continuous variable must be infinite, if read to infinite precision. However, of course, in real world communication nothing is read to infinite precision and there is a theory of communication using continuous signals which includes the signal noise and imprecision of signal transmission. We won't look at this here, interesting though it is and will consider the example of Infomax