Information Theory

Jorge Aurelio Menéndez February 11, 2020

Contents

1. Shannon information and entropy

Our goal is to transmit a message about the identity of a discrete signal S. The most straight-forward way of doing this is to use a binary code to encode its identity. A binary message of length L allows us to encode 2^L possible messages. Thus, if the signal can take on M different values, the length L of messages required to faithfully encode it is

$$2^L = M \quad \Leftrightarrow \quad L = \log_2 M$$

But note that we have assumed in this code that all possible signals are encoded with the same length of message. This is a wasteful use of bits when you consider the fact that some values of the signal might be more frequent than others – our messages about the signal could be much shorter on average if we used shorter messages to represent the values that occured the most frequently.

It turns out that the optimal length of a message encoding the signal value[1] s is given by the logarithm of the inverse probability of the signal taking on that value[2]:

$$L(s) = \log_2 \frac{1}{p(s)}$$

Under this code, the average length of our messages will then be:

$$H[p] = -\sum_{s=1}^{M} p(s) \log_2 p(s)$$

This is called the **entropy** of $p(s)$: the minimum achievable average message length about a discrete random variable S with probability distribution $p(s)$. Thus, the entropy is a functional of the probability mass function $p(s)$ over the discrete values the signal can take on. We'll generalize this notion to continuous random variables and probability *density* functions below.

In the limit of no randomness, i.e. $p(s^*) = 1$ and $p(s) = 0$ for any other $s \neq s^*$, one can easily see that the entropy goes to 0: if the identity of the signal can be known *a priori* without even having to sample it, there's no need to send a message about it. In a sense, any message about it becomes completely uninformative. On the other hand, if the signal values are uniformly distributed $p(s) = \frac{1}{M}$, the entropy reaches its maximum:

$$H[p] = \sum_{s=1}^{M} p(s) \log_2 \frac{1}{p(s)} \overset{*}{\leq} \log_2 \sum_{s=1}^{M} \frac{p(s)}{p(s)} = \log_2 M = -\sum_{s=1}^{M} \frac{1}{M} \log_2 \frac{1}{M}$$

where inequality $*$ follows from Jensen's inequality. In this case, messages about the signal become critical to the receiver's knowledge about it: without such information, a random guess is an educated guess. Thus, in this scenario messages about the signal are highly informative. It is in this sense that entropy measures the information in a signal. This framework for quantifying information is referred to as Shannon information, as it was developed by mathematician Claude Shannon in the 1950's.

Because we have been using binary codes, the unit of measurement is bits: this is the unit of measurement for the length of a message in binary code. But note that we need not use a binary code for

[1] We'll use upper-case letters (e.g. S) to denote random variables and lower-case letters (e.g. s) to denote the values they can take on, with $p(s)$ denoting the probability that the random variable S takes on the values s.

[2] For an intuitive demonstration of this, see http://colah.github.io/posts/2015-09-Visual-Information/

any of the above math to work out. We could have used a decimal code, for example, in which case all of the logarithms above would have to be changed to be base 10 instead of base 2 (in which case message length is measured in dits, or hartleys). Indeed, we can choose this base as we please. The convention is to use the natural logarithm of base e, in which case the unit of measurement is called a nat.

We can also generalize all these notions to continuous random variables with probability density functions $p(s)$. In this case, the conventional definition of entropy breaks down, because you need an infinite number of bits to encode a real number. This becomes evident in directly deriving the entropy of a continuous random variable S. One way to do this is to bin the possible values of S into discrete bins s_i of size Δs and then take the limit of $\Delta s \to 0$:

$$\begin{aligned} H[p] &= \lim_{\Delta s \to 0} -\sum_i p(s_i)\Delta s \log[p(s_i)\Delta s] \\ &= \lim_{\Delta s \to 0} -\sum_i p(s_i)\Delta s \log p(s_i) - \lim_{\Delta s \to 0} \log \Delta s \\ &= -\int p(s) \log p(s) \mathrm{d}s + \infty \end{aligned}$$

For continuous random variables, we thus use the **differential entropy**

$$h[p] = -\int p(s) \log p(s) \mathrm{d}s$$

obtained by simply throwing out the term that diverges.

2. KL divergence and mutual information

What if we don't use the optimal code? For example, what if we are mistaken about the true signal distribution $p(s)$ and instead construct our code optimally with respect to some other distribution $q(s)$? The average message length is then given by the **cross-entropy**

$$H_q[p] = -\sum_s p(s) \log q(s)$$

Naturally, this will be larger than the average message length obtained from using the optimal code. We can see this by taking the difference in message lengths under the two codes, called the **Kullback-Leibler (KL) divergence** between the two signal distributions:

$$\begin{aligned} \mathrm{KL}[p(s)||q(s)] &= H_q[p] - H[p] \\ &= -\sum_s p(s) \log q(s) + \sum_s p(s) \log p(s) \\ &= \sum_s p(s) \log \frac{p(s)}{q(s)} \end{aligned}$$

Applying Jensen's inequality, we can see that

$$-\mathrm{KL}[p(s)||q(s)] \leq \log \sum_s p(s) \frac{q(s)}{p(s)} = 0 \quad \Leftrightarrow \quad \mathrm{KL}[p(s)||q(s)] \geq 0 \quad \Leftrightarrow \quad H_q[p] \geq H[p]$$

Note that the KL divergence can be directly applied to continuous random variables, as the terms that diverge when computing the entropy of a continuous random variable cancel each other out when taking the difference between two entropies.

Another thing we can ask is how to update our code in the face of new information about the signal. Consider, for example, that we observe the firing rate R of a neuron that responds to a signal s according to the probability distribution $p(r|s)$. Using Bayes' rule to compute a posterior distribution $p(s|r)$ over signal values given the observed firing rate r, we could then update our code accordingly by using $p(s|r)$ instead of $p(s)$ to derive the optimal message length for each signal value s. Averaging over all possible neural responses r, we obtain the minimum average message length

$$H[S|R] = \sum_r p(r) \left(-\sum_s p(s|r) \log p(s|r) \right) = -\sum_{r,s} p(s,r) \log p(s|r)$$

which is called the **conditional entropy**.

We can now use this to ask: how much does observing the firing rate R improve our code? We can answer this by simply computing the resulting decrease in entropy (i.e. reduction in average message length), called the **mutual information**:

$$\begin{aligned}
\mathrm{I}[S,R] &= H[S] - H[S|R] \\
&= -\sum_s p(s)\log p(s) + \sum_{r,s} p(s,r)\log p(s|r) \\
&= -\sum_{s,r} p(s,r)\log p(s) + \sum_{r,s} p(s,r)\log\frac{p(s,r)}{p(r)} \\
&= \sum_{s,r} p(s,r)\log\frac{p(s,r)}{p(s)p(r)}
\end{aligned}$$

One can easily see that if the random variables R, S are independent (such that $p(r,s) = p(r)p(s)$), their mutual information is 0, since in this case observing R won't tell you anything about S (so $p(s|r) = p(s)$ and therefore $H[S|R] = H[S]$). Note as well that the mutual information is symmeteric (unlike the KL divergence). Lastly, note that the mutual information is always positive: knowing more about the signal S (i.e. conditioning $p(S) \to p(S|R)$) can only decrease your uncertainty about S (i.e. decrease the entropy). One can see this most easily by reexpressing the mutual information as a Kullback-Leibler divergence:

$$\mathrm{I}[S,R] = \mathrm{KL}[p(s,r)||p(s)p(r)] \geq 0 \quad \Leftrightarrow \quad H[S|R] \leq H[S]$$

We can use this fact to prove the **data processing inequality**, which tells us that any sequence of transformations on a signal cannot increase its information content. More formally, consider a sequence of transformations to a signal

$$S \to \tilde{S}_1 \to \tilde{S}_2$$

such that each transformation is applied directly to the previous output, forming a first-order Markov chain: $p(s,\tilde{s}_1,\tilde{s}_2) = p(\tilde{s}_2|\tilde{s}_1)p(\tilde{s}_1|s)p(s)$. We first use the positivity of the mutual information to recall that conditioning can only decrease entropy, such that

$$I[S,\tilde{S}_1|\tilde{S}_2] = H[S|\tilde{S}_2] - H[S|\tilde{S}_1,\tilde{S}_2] \geq 0 \quad \Leftrightarrow \quad H[S|\tilde{S}_2] \geq H[S|\tilde{S}_1,\tilde{S}_2]$$

Due to the Markov structure of the sequence, we additionally have that

$$H[S|\tilde{S}_1,\tilde{S}_2] = H[S|\tilde{S}_1]$$

Putting these two together, we then have that

$$H[S|\tilde{S}_2] \geq H[S|\tilde{S}_1]$$

implying that

$$\mathrm{I}[S,\tilde{S}_1] \geq \mathrm{I}[S,\tilde{S}_2]$$

which is the so-called data processing inequality: the amount of information $\tilde{S}_2$ provides about S is at most as much as that provided by $\tilde{S}_1$.

3. Maximum entropy distributions

As discussed in section 1, the entropy of a random variable in some sense measures our uncertainty about it: the more uncertain the receiver is about the signal, the longer the average message length we need to communicate it. For example, when the random variable has no stochasticity, the receiver can know with complete certainty what its value will be without ever even having to read the message, so the minimum average message length required is 0. On the other hand, if the random variable is uniformly distributed, the receiver is completely uncertain about the value and the minimum average message length required is large.

The entropy of a probability distribution thus provides a useful criterion for deriving probability distributions that are in some sense as "uncertain" as possible. This can be useful when we want to estimate the probability distribution of a random variable when we only know a few of its properties, e.g. its mean and variance. In this case, we'd like to assign it the most "uncertain" probability distribution

with those properties. We can do this by finding the **maximum entropy distribution** that satisfies them. Here we consider the case of continuous random variables, in which case the criterion we use is the differential entropy.

First consider the case in which we know nothing about a random variable X. In this case, we have no idea what values are more likely than others, so a natural choice for estimating its distribution is a uniform probability distribution. As we showed in section 1, this is indeed the probability distribution with highest entropy when the random variable is discrete. In the continuous case, we can verify this intuition still holds by using the calculus of variations to derive the probability density function $p^*(x)$ with highest differential entropy. We do this by solving a constrained variational optimization problem: find the function that that maximizes the differential entropy subject to the constraint that it integrate to 1. By incorporating this constraint, we ensure that the solution will be a probability density function. We thus proceed by writing down the Lagrangian functional

$$p^*(x) = \arg\max_{p(\cdot)} \left(-\int p(x) \log p(x) \mathrm{d}x \right) + \lambda_0 \left(\int p(x) \mathrm{d}x - 1 \right)$$

where λ_0 is a Lagrange multiplier. Setting its functional derivative to 0 and solving, we obtain

$$0 = -\log p^*(x) - 1 + \lambda_0 \quad \Rightarrow \quad p^*(x) = e^{\lambda_0 - 1}$$

Evidently, the probability density function $p^*(x)$ with maximal differential entropy is constant with respect to x, meaning that it is indeed that of a uniform distribution.

Now consider the case in which we know only the mean μ of the random variable X. What is the maximum entropy distribution $p^*(x)$ with this mean? Again, we can answer this question by expressing it as a constrained variational optimization problem. But we now have two constraints, which we incorporate into the Lagrangian functional with two Lagrange multipliers λ_0, λ_1:

$$p^*(x) = \arg\max_{p(\cdot)} \left(-\int p(x) \log p(x) \mathrm{d}x \right) + \lambda_0 \left(\int p(x) \mathrm{d}x - 1 \right) + \lambda_1 \left(\int x p(x) \mathrm{d}x - \mu \right)$$

Setting the functional derivative to 0 and solving, we obtain

$$0 = -\log p^*(x) - 1 + \lambda_0 + \lambda_1 x \quad \Rightarrow \quad p^*(x) \propto e^{\lambda_1 x}$$

i.e. the maximum entropy distribution with mean μ is the exponential distribution with that mean (the mean falls out of the derivation after solving for λ_1, λ_2). Applying this result to the distribution of intervals between events in a point process with a constant mean rate, we get that the **maximum entropy point process** with a constant mean rate is one with exponentialy distributed inter-event intervals – that is, a homogenous Poisson process.

Finally, consider the case in which we know the mean μ and variance σ^2 of X. In this case, our Lagrangian becomes

$$p^*(x) = \arg\max_{p(\cdot)} \left(-\int p(x) \log p(x) \mathrm{d}x \right) + \lambda_0 \left(\int p(x) \mathrm{d}x - 1 \right) + \lambda_1 \left(\int (x-\mu)^2 p(x) \mathrm{d}x - \sigma^2 \right)$$

Setting the functional derivative to 0, we arrive at

$$0 = -\log p^*(x) - 1 + \lambda_0 + \lambda_1 (x-\mu)^2 \quad \Rightarrow \quad p^*(x) \propto e^{\lambda_1 (x-\mu)^2}$$

i.e. the maximum entropy distribution of a random variable with given mean and variance is Gaussian. We can also derive this result by considering an arbitrary probability density $q(x)$ with the same mean μ and variance σ^2 and calculating its KL divergence with the mean- and variance- matched Gaussian $p(x)$:

$$\begin{aligned}
\mathrm{KL}[q||p] &= \int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x \\
&= -h[q] - \int q(x) \log p(x) \mathrm{d}x \\
&= -h[q] + \frac{1}{2} \log 2\pi\sigma^2 + \frac{1}{2\sigma^2} \int q(x)(x-\mu)^2 \mathrm{d}x \\
&= -h[q] + \frac{1}{2} \log 2\pi\sigma^2 + \frac{1}{2} \\
&= -h[q] + \frac{1}{2} \log 2\pi\sigma^2 e
\end{aligned}$$

Noting that

$$h[p] = \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\int p(x)(x-\mu)^2 \mathrm{d}x = \frac{1}{2}\log 2\pi\sigma^2 e$$

the positivity of the KL divergence gives us that

$$\mathrm{KL}[q||p] = -h[q] + h[p] \geq 0 \quad \Leftrightarrow \quad h[p] \geq h[q]$$

In other words, any probability distribution with the same mean and variance will have less (differential) entropy than the Gaussian with that mean and variance. The maximum entropy distribution with that mean and variance is thus the Gaussian.

4. Information channels

Shannon information can be useful for analyzing **channels**, in which information about a signal S is transmitted by transforming it into an output R with conditional probability distribution $p(r|s)$. For example, a simple channel would be one that outputs the signal corrupted with some 0-mean noise η,

$$r = \tilde{s} + \eta$$

which might be a good model of a cable transmitting a signal that gets corrupted by e.g. thermal noise. The distribution $p(r|s)$ thus reflects not only any transformation of the signal but also the stochasticity involved in the transmission. An important question we might then ask is how good a given channel is at transmitting information about the signal, i.e. how informative its outputs R are about the signal S.

A natural way to quantify this is to use the mutual information $\mathrm{I}[S, R]$. But note that the mutual information depends on the signal distribution $p(s)$. To really characterize the intrinsic properties of the channel itself, we'd like to make general statements about its information transmission regardless of what the signal distribution $p(s)$ is. We thus define the **channel capacity** to be the maximal mutual information over all possible signal distributions, expressed mathematically through a supremum

$$C_{R|S} = \sup_{p(S)} \mathrm{I}[S, R]$$

The channel capacity is thus the theoretical limit on the amount of information that can be transmitted through the channel. To use this channel optimally, then, you should ensure that the distribution of the signal being passed through it saturates this upper bound. This distribution can be found using a simple iterative EM-like algorithm called the Blahut-Arimoto algorithm.

However, typically the distribution of the signal S we want to transmit information about is not under our control. Given a fixed channel, we might then find a transformation $\tilde{S} = f(S)$ of the signal such that the distribution $p(\tilde{s})$ of the transformed signal saturates the channel capacity, i.e. maximizes the mutual information $I[\tilde{S}|R]$ with the channel output. Or, we might be in a situation in which we are designing the channel itself, and would like to make sure that it is tuned to the signal distribution in such a way that its capacity is maximized while ensuring that it is saturated by the signal distribution.

One simple scenario in which we can do this is the noiseless case, in which the channel simply implements a deterministic transformation

$$R = f(S)$$

In this case, the conditional entropy $H[R|S]$ is 0, and the mutual information can be expressed as

$$\mathrm{I}[S, R] = H[R] - H[R|S] = H[R]$$

Maximizing the capacity of this channel can therefore be achieved by setting the encoding function $f(\cdot)$ so as to maximize the entropy of the channel output distribution $p(r)$. Given a restricted range of possible outputs Δ, the maximum entropy output distribution is the uniform distribution,

$$p(r) = \frac{1}{\Delta}$$

where Δ may reflect, for example, hardware constraints on the outputs of the channel. To solve for the encoding function $f(\cdot)$ that achieves this maximum, we first express the probability density function $p(r)$

in terms of the encoding function and the signal distribution. Assuming $f(\cdot)$ to be invertible, we can write

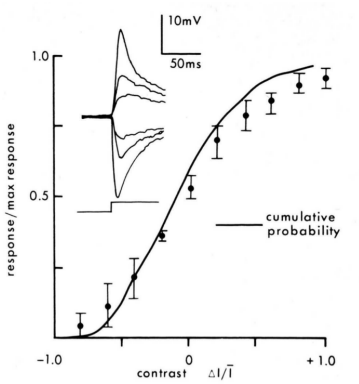
$$\begin{aligned} p(r)\mathrm{d}r &= p(s)\mathrm{d}s, \quad r = f(s) \\ \Leftrightarrow p(r) &= p(s)\frac{\mathrm{d}s}{\mathrm{d}r} \\ &= p(s)\left.\frac{\mathrm{d}f^{-1}}{\mathrm{d}r}\right|_{r=f(s)} \\ &= p(s)\frac{1}{f'(s)} \end{aligned}$$

where in the last equality we used the inverse function theorem[3], and $f'(s) = \left.\frac{\mathrm{d}f}{\mathrm{d}s}\right|_s$ denotes the derivative of $f(\cdot)$ at s. Plugging in the desired uniform distribution for $p(r)$, we arrive at the following optimal encoding distribution:

$$f'(s) = \Delta p(s) \quad \Leftrightarrow \quad f(s) = r_{min} + \Delta \int_{-\infty}^{s} p(s')\mathrm{d}s'$$

where r_{min} denotes the minimum channel output. The result is simple and intuitive: the optimal encoding function is proportional to the cumulative probability distribution of the signal. The first of these equations provides intuition for why this is the most informative encoding: it demands that the channel output be most sensitive to signal values the highest probability density. Thus, large portions of the range of possible outputs are dedicated to ranges of signal values with highest probability density, where the cumulative density changes most. Conversely, ranges of signal values with low probability density are regions of s-space over which the cumulative density is flat and changes little, such that only a small portion of the output range is allocated to them.

In the image processing community, this technique is referred to as **histogram equalization**, because the re-encoding R of the input S has a uniform distribution, i.e. a flat histogram. Such encoding has been proposed to be at play in the contrast sensitivity of large monopolar cells in the insect compound eye. [Laughlin, 1981] investigated whether this was the case by estimating the probability distribution of contrast in patches of natural images, and then comparing the fly large monopolar cell contrast sensitivity function to the cumulative probability distribution. He found that these matched remarkably well (cf. figure below), suggesting that the circuitry generating such responses may have evolved to maximize the capacity of this channel.



Recorded fly large monopolar cell responses to different contrasts, overlaid on the empirically measured cumulative probability distribution of contrasts in natural images patches. Copied from [Laughlin, 1981].

[3] Easily proved for invertible $f(\cdot)$ by noting that

$$f^{-1}\left(f(s)\right) = s \quad \Leftrightarrow \quad \frac{\mathrm{d}}{\mathrm{d}s}\left[f^{-1}\left(f(s)\right)\right]\Big|_s = 1$$

and then using the chain rule to re-write the second equality as

$$\left.\frac{\mathrm{d}f^{-1}}{\mathrm{d}r}\right|_{r=f(s)} \left.\frac{\mathrm{d}f}{\mathrm{d}s}\right|_s = 1 \quad \Leftrightarrow \quad \left.\frac{\mathrm{d}f^{-1}}{\mathrm{d}r}\right|_{r=f(s)} = \left(\left.\frac{\mathrm{d}f}{\mathrm{d}s}\right|_s\right)^{-1}$$

Plugging in $f'(s) = \left.\frac{\mathrm{d}f}{\mathrm{d}s}\right|_s$ gives us the equality used above.

Index

References

[Laughlin, 1981] Laughlin, S. (1981). A Simple Coding Procedure Enhances a Neuron's Information Capacity. *Zeitschrift für Naturforschung C*, 36(9-10):910–912.