Information Processing and the Brain

Josh Felmeden

October 8, 2021

Contents

1 Information Theory

Information theory quantifies the amount of information that is potentially available from the communication channel using. So, we look to answer the question: can the receiver decide how informative the channel is likely to be; how much information is in the channel.

1.1 Randomness

The key is that if the information in the channel is predictable, the receiver is not going to learn much from this. Therefore, we will look at randomness (coin flipping, geiger counters, etc).

We will also look and unexpectedness, and consider the relationship between randomness, unexpectedness, and information theory.

Netflix is a good example of this. Their star rating of films helps them to recommend films to you, and use your experiences to recommend films to others. The recommended tab is the *channel* in the information theory. The films are the medium, and Netflix is communicating to the user information about the film. If Netflix recommends a really popular film, the information is predictable, and the user is not learning much. Therefore, we are not gaining much information about the quality of the films.

Another example is the star rating of movies. The star rating is not 'random' enough for the consumer, in that there is not enough information that we as the user can learn from this. A good film with a lot of hype will likely have a good star rating, and therefore we would already know this.

Therefore, we can say that a statement, despite being correct, can be useless if we can predict it.

> *The theory of information starts with an attempt to allow us to quantify the informativeness of information, but not its salience or validity.*

2 Shannon's entropy

For a finite discrete distribution with a random variable X, possible outcomes $\{x_1, , x_2, \ldots, x_n\} \exists \mathcal{X}$ and a probability mass function p_x giving probabilities $p_X(x_i)$, the entropy is:

$$H(X) = -\sum_{x_i \exists \mathcal{X}} p_X(x_i) \log_2 p_X(x_i)$$

This formula *quantifies* some information received. In the example of star ratings on netflix:

1 star	0.016
2 star	0.310
3 star	0.627
4 star	0.057

We obtain the following formula:

$$H(X) = -0.016 \log_2 0.016 - 0.31 \log_2 0.31 - 0.627 \log_2 0.627 - 0.057 \log_2 0.057 \approx 1.28$$

Choosing base two is arbitrary, but because we are dealing with information, it is useful to deal in bits. In other circumstances, it is possible to choose a different base. The result, 1.28 is the entropy for the channel of Netflix film star ratings.

If the star ratings were all equally likely, we would get an entropy of:

$$H(X) = -4 \times 0.25 \log_2 0.25 = 2$$

This is higher value entropy meaning the information is more useful.

In contrast, if all films are rated one star, we end up with entropy 0, meaning that it is useless. We assume $\log 0$ is 0 for shannon's entropy despite this not being the case in reality.

2.1 Nice properties

Shannon's entropy works on any sample space. It's good because you can calculate this on things that would otherwise be impossible, such as items bought from a grocers. This is because it is defined on probability space, while other calculations are calculated in vector space (such as mean, etc).

It's **always positive**, unless it is zero, which is when the distribution *isn't random* (aka *determined*).

If the distribution is *uniform*:

$$p_X(x_i) = \frac{1}{n}$$

for all x_i where (# is equivalent to `len`)

$$n = \#\mathcal{X}$$

then, since $-\log_2(1/n) = \log_2 n$

$$H(X) = \log_2 n$$

We won't prove it here, but is not difficult to prove that:

$$0 \le H(X) \le \log_2 n$$

with $H(X) = 0$ *only* if one probability is one and the rest zero, and $H(X) = \log_2 n$.

2.1.1 Case of n = 2

With two outcomes, a and b with $p(a) = p$ and $p(b) = 1 - p$ then

$$H = -p\log_2 p - (1-p)\log_2(1-p)$$

The resulting graph of entropy will be symmetric; rising towards 1.

2.1.2 Source coding

the main reason to believe that Shannon's entropy is a good quantity for calculating entropy is its relationship to so called source coding.

Consider storing a long sequence of letters A, B, C, D as binary (AABACBDA for example). If we wanted to digitise this, we might use a very simple binary representation of these characters (A = 00, B = 01, ...). Using this representation means the average bits per letter is 2. Now, if we had different distribution of letters, then we could possibly come up with a smaller, more compact version of the code. For example, if we had the following distribution:

A	B	C	D
0.5	0.25	0.125	0.125

We could use the following encoding:

A	B	C	D
0	10	110	111

Don't forget that this resulting code should be *prefix free*.

Now, we can prove that this code is shorter on average:

$$L = 0.5 \times 1 + 0.25 \times 2 + 0.125 \times 3 + 0.125 \times 3 = 1.75 < 2$$

This is also the same as the Shannon's entropy of the code. Each of the logs are giving us the length of the corresponding code word.

The source coding theorem states that, for the most efficient code, we get:

$$H(X) \le L < H(X) + 1$$

This means that if you have a sequence of objects that have some regularity and you want to code them into binary, the channel's entropy is a lower bound on the average length of a message and you can get a *prefix free* code within one of Shannon's entropy.

2.2 Joint and Conditional Entropy

Typically, we want to use information theory to study the relationship between two random variables. So far, we have only looked at single variables.

> **Joint Entropy**
>
> Given two random variables X and Y, the probability of getting the pair (x_i, y_j) is given by the **joint probability** $p(X, Y)(x_i, y_j)$. THe joint entropy is just the entropy of the join distribution:
>
> $$H(X, Y) = -\sum_{i,j} p_{X,Y}(x_i, y_j) \log_2 p_{X,Y}(x_i, y_j)$$
>
> This is essentially the probability that x will produce x_i and that y will produce y_j

Here's an example:

	x_0	x_1
y_0	$1/4$	$1/4$
y_1	$1/2$	0

$$H(X, Y) = -\frac{1}{2} \log_2 \frac{1}{4} - \frac{1}{2} \log_2 \frac{1}{2} = \frac{3}{2}$$

2.2.1 Conditional Entropy

$p_{X|Y}(x_i|x_j)$ is the **conditional probability** of x_i given y_j. If we know that $Y = y_j$, it gives us the probability that the pair is (x_i, y_j).

$$p_{X|Y}(x_i|y_j) = \frac{p_{(X,Y)}(x_i, y_j)}{p_Y(y_j)}$$

The **marginal probability** for x is: $p_X(x_i) = \sum_j p_{(X,Y)}(x_i, y_j)$. This is the way of getting the probabilities for one of the two random variables from the joint distribution.

Let's now substitute the conditional probability into the formula for entropy:

$$H(X|Y = y_j) = -\sum_i p_{(X|Y)}(x_i|y_j) \log_2 p_{X|Y}(x_i, y_j)$$

This is the entropy of X as we know $Y = y_j$. We call this the **conditioned entropy**.

The conditional entropy is the average amount of information still in X when we know Y. It also has some pretty nice properties:

If X, Y are *independent*, then:

$$p_{X,Y}(x_i, y_j) = p_X(x_i)p_Y(y_j)$$

For all i, j and

$$p_{X|Y}(x_i|y_j) = p_X(x_i)$$

so

$$H(X|Y) = -\sum_{i,j} p_{X,Y}(x_i, y_j) \log_2 p_{X|Y}(x_i|y_j) = H(X)$$

Which is what we want, since if Y tells us nothing about X, then the conditional entropy should just be the same as the entropy of X.

Conversely, if X is *determined* by Y, then $H(X|Y) = 0$. This could happen if the only x_j, y_i pairs that actually occur are (x_i, y_i).

2.3 Mutual Information

Given two (overlapping) variables, there will be some information that is shared by these two entropies, and this is called **mutual information**, and is defined as:

$$I(X, Y) = H(X) + H(Y) - H(X, Y)$$

Or:

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

By substituting in the formulas, we end up with:

$$I(X, Y) = \sum_{i,j} p_{X,Y}(x_i, y_j) \log_2 \frac{p_{X,Y}x_i y_j}{p_X(x_j)p_Y(y_j)}$$

Going back to the previous example:

	x_0	x_1
y_0	1/4	1/4
y_1	1/2	0

This has $H(X, Y) = 3/2, H(X) \approx 0.81$ and $H(Y) = 1$ so $I(X, Y) \approx 0.31$

If X and Y are independent, we just end up with $I(X, Y) = 0$. In fact, $I(X, Y) \geq 0$

2.4 Correlation

The correlation is defined as:

$$C(X, Y) = \frac{\langle (X - \mu_X)(Y - \mu_Y) \rangle}{\sigma_X \sigma_Y}$$

Where μ_X is the average of X and σ_X is the standard deviation.

Correlation only works if X, Y take their values in vector space (meaningfully be able to multiply the two together).

Consider:

	-1	0	1
1	1/4	0	1/4
0	0	1/2	0

Then:

$$C(X, Y) = 0$$

Whereas $I(X, Y) = 1$.

2.5 The data processing inequality

There is something called **conditional independence**. Imagine playing a game of snakes and ladders, and X, Y, Z are the resulting position after moves 1,2,3 respectively. Knowing X will change the probability of Z, but only if we don't know Y, since if we know Y, it doesn't really matter what X was. This means that X and Z are conditionally independent (conditioned on Y).

The data processing inequality states that if:

$$X \to Y \to Z$$

then:

$$I(X, Y) \geq I(X, Z)$$

With equality if and only if $X \to Z \to Y$

3 Information Theory in the Brain

3.1 Information in the Brain

Spike trains are when neurons spike with some level of voltage when they are activated. This spike passes from neuron to neuron, which is how information is passed along an axon.

We can use Shannon's entropy to look at the similarities in these spikes to see the shared information.

An experiment was performed on a fly where it was shown a series of bars on a TV screen and the spike trains from the neurons at the back of the brain (the ones that observe horizontal movement) was recorded. More specifically, the timing of the spikes was recorded. This information was then *discretised*, and evaluated. The information recorded are the action spikes. *Discretisation* meant that the information was looked at in 3ms, and then return a 1 if there is a spike, or a 0 if not (think time bins). This binary sequence was then split up into words. Each word corresponds to the amount of time that an entire piece of information has been encoded. The calculations performed revealed that the neurons that were being looked at only remember things in 30ms, so there was a natural 'word length' for splitting up the binary sequence, thus 30ms was taken.

These words can now be converted into a set of objects. The random variable that is taken as the communication channel is a chunk of spike train for 30ms, represented by one of the words. Now, we want to look at the probability of different words:

$$p(w_0) \approx \frac{\#(\text{occurance of } w_0)}{\#(\text{trials})}$$

Now, two different tests were performed. The first is where the stimulus is random. The second is where a fixed five minute segment of the stimulus is shown ($H(W)$), and the responses are considered ($H(W|S)$).

Now, we can get the mutual information about the words given:

$$I(W, S) = H(W) - H(W|S)$$

Or, the information about S is the total information in W subtract the noise.

3.1.1 Discretisation size

One important question is how to decide how small to make hte discretisation time and how long to make the words. In our example, given that $\delta t = 3ms$ gives 78 ± 5 bits per second or 1.8 ± 0.1 bits per spike.

Unfortunately, if we have a 30ms word with 3ms letters, this gives us 10 letters per word giving $2^{10} = 1024$. If six seconds of data are used for the repeating stimulus, that is 100 different stimuli, then even for a three hour recording, there are 1800 trails for each stimulus. This is not a big amount for estimating 1024 probabilities.

3.2 Differential Entropy

Differential Entropy is the name given to Shannon's entropy for continuous probability distributions where the sample space is $\mathcal{X} \subseteq \mathbf{R}^d$. In the examples we will be looking at, $d = 1$. The idea is that: $h(X) = \int dx p(x) \log_2 p(x)$. This is a good guess as to what the continuous distribution would look like of Shannon's entropy.

If we consider a uniform distribution where:

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

So

$$h(X) = \int_{-\infty}^{\infty} dx p(x) \log_2 p(x) = -\frac{1}{a} \int_0^a dx \log_2 \frac{1}{a}$$

And so:

$$h(X) = \log_2 a$$

This, unfortunately, does not guarantee that the result will be positive, which is one of the useful properties of Shannon's entropy.

3.2.1 Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}}$$

If we substitute and integrate by parts, we get:

$$h(x) = \frac{1}{2} \log_2 2\pi e \sigma^2$$

As you expand the disrtibutions, we can prove that for fixed variance, Gaussian has the highest entropy.

Densities are not probabilities. The discrete case $p(x)$ is the probability that $X = x$. For the continous case:

$$\int_{x_0}^{x_1} dx p(x)$$

is the probability that $x_0 \leq x \leq x_1$. The usual sums and probabilities go to integrals and densities doesn't work because there is a p in the log.

We also need to model the sensitivity of the receiver as well as the behaviour of the source.