Information theory

The theory of information is a theory of communication. It is the mathematical langauge which allows us to calculate the amount of information that can be communicated along a channel, based only on our measurement of the statistics of the data travelling along that channel. It is not a theory of salience, it doesn't tell us if the information is important, or what is happening to the information as it is communicated from one part of the network to another.

The theory of information is part of the language of the study of computation; it forms part of the framework we use to think about how data is processed in networks, both artificial networks like deep learning nets and organic networks, such as the brain. We will find in this unit that it is not a complete theory of computation, we will use ideas from information theory along with other ideas about learning and computation. It is a good place to start.

A motivating example

The key insight in information theory is to think about randomness in the right way. Imagine you are applying for a job and you have to fill in your final grade; for simplicity a first, a second or a third, on the application form. Now, your grade isn't random, there might be a random element, but it is also the result of your ability to do well in exams. Furthermore, your potential employer is interested in your grade precisely because it isn't random, some employers, perhaps not to the degree you imagine, think is related to your ability to perform in the role they are offering. However, until they read what you have written they do not know your grade and so, to them, it is like performing a random experiment and it can be modelled using a random variable.

In fact, most situation we use a random variable for are like this; the variable models something we don't know rather than something that is truly random. The example of a coin flip often used when describing random variables is misleading.

Now, returning to the scenario above, consider how much the potential employer learns from reading your grade. This, of course, depends on how well the grading is aligned to the potential employer's, that is a complicated question, but there is a simpler issue related to the randomness of the vari-

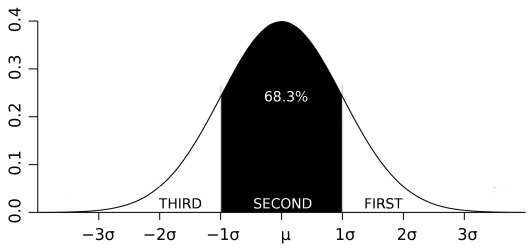

Figure 1: A simplified picture of marking to a curve.

able, the degree to which the potential employer can't guess the answer before reading what is written.

Think about how exams are marked. In America they are marked 'to a curve'; we don't do that here and the description here isn't a picture of how your exams are marked, it is just used to motivate information theory. In a cartoon sketch of marking to a curve, everyone is marked and the grades fall into a normal curve and two divisions are made at the points where the curves are steepest dividing those who took the exam into three groups, firsts, seconds and thirds. For definiteness say the distribution of marks is

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

and the divisions are made at $x = \mu \pm \sigma$ then 68.3% of students will get a second, 15.9% will get either a first or a third. This is described in Fig. 1.

Now, think of the potential employer: sometimes when they ask what grade a prospective employee got they will find out something very significant, if the student got a first they are in the top 15.9% of exam takers, if they got a third they are in the bottom 15.9%; however, most of the time, almost seven times in ten, 68.3% of the time to be more exact, they will learn that the student got a second. This isn't very informative, it just says the student got the same grade as 68.3% of students. Thus, although some of the time the prospective employer learns something very informative, most of the time they learn that the student is about the same as most students. On average this isn't a very informative distribution. Clearly the grading system would be more informative if one third of students got each of the three grades. If that was the case the employer would have no idea what the

answer to the question 'what grade did you get?' is going to be; as it is, they have a reasonable idea the answer might be 'a second'.

As another similar example; before I left Ireland I had a colleague called Stefan. Whenever the economy, this was during the so called Celtic tiger, was discussed he used to say 'it will crash'. Now it turns out he was right; but his answer wasn't very informative because I knew what he was going to say before he said it, I knew he would say 'it will crash' and so, for this and other reasons, talking to him was very boring and not very informative, even though on this particular topic he was correct.

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

Shannon's entropy

Shannon's Entropy was introduced by Claude Shannon in his 1948 paper which basically created the field of information theory (**?**). It is a single quantity that measures this idea of informativeness, balancing how useful a piece of information is with how likely you are to get it. We will focus for now on finite discrete sample spaces; going from finite spaces to infinite is easy, but going from discrete to continuous is tricky. For a finite discrete distribution with random variable X, possible outcomes $\{x_1, x_2, \dots x_n\} \in \mathcal{X}$ and a probability mass function p_X giving probabilities $p_X(x_i)$, the entropy is

$$H(X) = -\sum_{x_i \in \mathcal{X}} p_X(x_i) \log_2 p_X(x_i) \tag{2}$$

In this definition $p \log_2 p = 0$ when $p = 0$; this makes sense since

$$\lim_{p \to 0} p \log_2 p = 0 \tag{3}$$

The first thing to note about this definition is that is works for any sample space. Many quantities in statistics rely on the sample space having some sort of structure, for example the mean of distribution@

$$\langle x \rangle = \sum_{x_i \in \mathcal{X}} p_X(x_i) x_i \tag{4}$$

only works if it makes sense to multiple the x_i by real number and add them to each other. In other words, it assumes the sample space is a vector space.

Not all sample spaces are vector spaces: if we were looking at fruit purchased in a supermarket, the average fruit would make no sense since we would not know how to work out

$$0.25 \times \text{apple} + 0.125 \times \text{banana} + 0.1 \times \text{pear} \ldots \tag{5}$$

This is not a problem for Shannon's entropy, the one thing a sample space always has is a probability for each element, so $H(X)$ is always defined.

$H(X)$ has some other properties that seem suitable for the quantification of information; for example

$$H(X) \geq 0 \tag{6}$$

This follow from $0 \geq p_X(x_i) \leq 1$ so $\log_2 p_X(x_i)$ is never positive. Furthermore, it is easy to check that if all the events are equally likely:

$$p(x_i) = \frac{1}{\#(\mathcal{X})} \tag{7}$$

where $\#(\mathcal{X})$ is the number of points in the sample space, then

$$H(X) = \log \#(\mathcal{X}) \tag{8}$$

In fact it can be proved, though we won't prove it here, that

$$H(X) \leq \log \#(\mathcal{X}) \tag{9}$$

with equality only for the case above, where every x_i is equally likely. This is good; the most informative an experiment can be is when we have no idea before the experiment what the outcome will be, in other words, when all the outcomes are equally likely. Conversely if $p_X(x_k) = 1$ for one outcome and $p_X(x_i) = 0$ otherwise, that is for all $i \neq j$, then substituting in to the formula we see that $H(X) = 0$; this is what we want, we know in advance the outcome of the experiment would be x_k so the experiment itself provides no information. If there are two outcomes, a and b with $p(a) = p$ and $p(b) = 1-p$ then the entropy is

$$H = -p \log_2 p - (1-p) \log_2 (1-p) \tag{10}$$

which is plotted as Fig. 2.

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

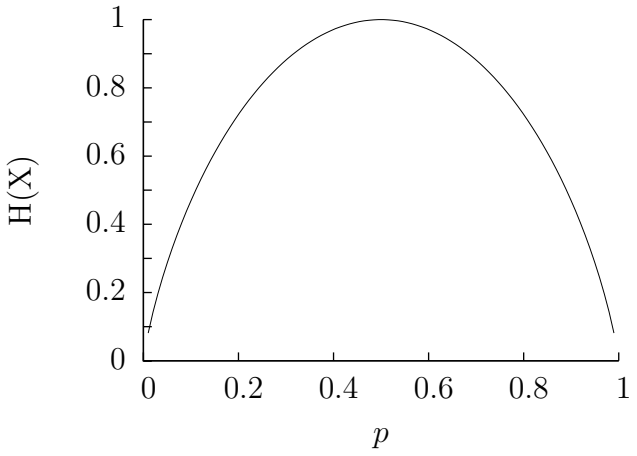

Figure 2: Information with two outcomes, p is plotted on the horizonal axis, $H(X)$ on the vertical.

Imagine storing a long sequence made up of the letters `A`, `B`, `C` and `D` as binary. The obvious way to do it would be to say that there are four letters so the sequence should be stored using two bits, a dictionary might look like

`A`	`B`	`C`	`D`
00	01	10	11

so the sequence `AABC` would be stored as 00000110, splitting this up into two: 00 00 01 10 allows the binary to be converted back into the original letters. Moreover, since each letter is coded using two bits, it is clear the code length is twice the number of letters.

Now, say we also knew that $p(\texttt{A}) = 0.5$, $p(\texttt{B}) = 0.25$, $p(\texttt{C}) = p(\texttt{D}) = 0.125$, in other words, in the message that will be encoded, `A` occurs half the time, `B` a quarter the time and `C` and `D` an eighth of the time. Now, consider this dictionary

`A`	`B`	`C`	`D`
0	10	110	111

Here, the sequence `AABC` become 010110110, this can be split up into 0 10 110 110 because the code word 0 is the only code word beginning with 0 and the

code word 10 is the only one beginning with 10. Now, some of the code words are longer than two, but, since `A` occurs half the time and has a code word of length one, `B` occurs a quarter the time and has a code word of length two and `C` and `D` each occur an eighth the time with code words of length three, the average code length for each letter is $0.5\times1+0.25\times2+0.125\times3+0.125\times3 = 1.75$. This is the same as the entropy

$$H(X) = -0.5\log_2(0.5) - 0.25\log_2(0.25) - 0.250\log_2(0.125) = 1.75 \quad (11)$$

In fact, there source coding theorem shows that the entropy $H(X)$ is a lower bound on the average length of a message using the most efficient code; it is a limit on the compressibility of the data. Here, the code attains the bound, but this works because the number of letters and their probabilities were chosen to make it work; usually there is a sort of rounding error because the code words have integer length. However, the source coding theorem guarantees that the most efficient code will attain or nearly attain its limit.

Finally, to return to the exam grade example. The distribution we looked at has

$$H(X) = -0.684\log_2 .684 - 0.386\log_2 0.159 = 1.4 \quad (12)$$

If, instead, the division between the grades were arranged so that an equal number of people got a first, second and third then the probability would be

$$p(\text{first}) = p(\text{second}) = p(\text{third}) = \frac{1}{3} \quad (13)$$

and

$$H(X) = -\log_2 \frac{1}{3} = 1.58 \quad (14)$$

So, on average the employer would attain 0.18 bits of information if grade boundaries were chosen so that each grade was equally common. However, of course, this might not be the information the employer wants, maybe the specifically want to employ someone with a grade in the lowers 16%, maybe they thing these people are more creative. In this situation the new system would be worse, they may gain more information, but not the specific piece of information they need!

Joint entropy and conditional entropy

Typically we want to use information theory to study the relationship between two random variables. The object which quantifies this is mutual

information. However, before discussing mutual information it is useful to consider conditional entropy. Given two random variable X and Y the probability of getting the pair (x, y) is given by the **joint probability** $p_{(X,Y)}(x, y)$. The **joint entropy** is just the entropy of the joint distribution:

$$H(X, Y) = -\sum_{x,y} p_{X,Y}(x, y) \log_2 p_{X,Y}(x, y) \tag{15}$$

Now, recall that $p_{X|Y}(x|y)$ is the **conditional probability** of x given y; if we know that a pair drawn from the joint probability has $Y = y$ it gives the probability that the pair is (x, y). This means

$$p_{(X,Y)}(x, y) = p_{X|Y}(x|y)p_Y y \tag{16}$$

In other words, the probability of (x, y) is the probability of y multiplied by the probability of x given that we know $Y = y$. The marginal probabilities are related to the conditional probabilities as well:

$$p_X(x) = \sum_y p_{X|Y}(x|y) \tag{17}$$

Now, we can stick the conditional probability into the formula for the entropy:

$$H(X|Y = y) = -\sum_x p_{X|Y}(x|y) \log_2 p_{X|Y}(x|y) \tag{18}$$

This is the entropy of the variable X if we know $Y = y$. The **conditional probability** is the average of this, averaged over all the values of y; hence

$$H(X|Y) = \sum_y p_Y(y) H(X|Y = y) = -\sum_{x,y} p_{X,Y}(x, y) \log_2 p_{X|Y}(x|y) \tag{19}$$

The conditional entropy has some nice properties; for example, if X and Y are independent then $p_{X,Y}(x, y) = p_X(x)p_Y(y)$ and $p_{X|Y}(x|y) = p_X(x)$ so

$$H(X|Y) = -\sum_{x,y} p_{X,Y}(x, y) \log_2 p_{X|Y}(x|y) = H(X) \tag{20}$$

Conversely, if X is determined by Y, for example if $x = f(y)$ for some function f then $p_{X|Y}(x|y)$ is zero for every x except $f(y)$, in which case it is one and

$$H(X|Y) = 0 \tag{21}$$

We can also intrepret $H(X|Y)$ in a straightforward way, it is the average amount of information still in X when we know Y.

Lets do an example. Given the joint distribution for (X, Y):

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

Calculating the joint entropy is easy:

$$H(X,Y) = -\frac{1}{2}\log_2\frac{1}{4} - \frac{1}{2}\log_2\frac{1}{2} = 1.5 \tag{22}$$

We can also calculate the entropy of the marginal distribution: $p_X(x = x_0) = 3/4$ and $p_X(x = x_1) = 1/4$ so

$$H(X) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.81 \tag{23}$$

Now, to work out the conditional probability we need to work out the entropy conditioned on the two values of y; so $p_{X|Y}(x_0|Y = y_0) = p_{X|Y}(x_1|Y = y_0) = 1/2$ so clearly

$$H(X|Y = y_0) = 1 \tag{24}$$

On the other hand $p_{X|Y}(x_0|Y = y_1) = 1$ but $p_{X|Y}(x_1|Y = y_1) = 0$ so

$$H(X|y = y_1) = 0 \tag{25}$$

and

$$H(X|Y) = \frac{1}{2}H(X|y = y_0) + \frac{1}{2}H(X|y = y_1) = 0.5 \tag{26}$$

Thus, although it turns out $H(X|Y = y_0) > H(X)$

$$H(X|Y) < H(X) \tag{27}$$

and we will see that in general $H(X|Y) \leq H(X)$; this is as it should be, knowing about Y can only reduce the amount of information in X on average.

Finally, there is a chain rule for conditional entropy; recall

$$p_{X,Y}(x, y) = p_{X|Y}(x|y)p_Y y \tag{28}$$

Now

$$H(X,Y) \quad = \quad -\sum_{x,y} p_{X,Y}(x,y)\log_2 p_{X,Y}(x,y)$$

$$\begin{aligned} &= -\sum_{x,y} p_{X,Y}(x,y) \log_2 p_{Y|X}(y|x) p_X(x) \\ &= -\sum_{x,y} p_{Y|X}(y|x) p_X(x) \log_2 p_{Y|X}(y|x) \\ &\quad -\sum_{x} p_X(x) \log_2 p_X(x) \end{aligned} \tag{29}$$

using

$$\begin{aligned} \sum_{x,y} p_{X,Y}(x,y) \log_2 p_X(x) &= \sum_{x} \left(\sum_{y} p_{X,Y}(x,y) \right) \log_2 p_X(x) \\ &= \sum_{x} p_X(x) \log_2 p_X(x) \end{aligned} \tag{30}$$

Thus

$$H(X,Y) = H(X) + H(Y|X) \tag{31}$$

This again makes sense; the amount of information in X and Y is the amount of information in X plus the amount of information remaining in Y if we already know X.

Mutual information

The mutual information is a measure of how related two distributions are; it is given by

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{32}$$

Thus, it is the amount of information in X and Y considered separately, minus the amount of information in them considered together. If the two distributions are independent, then $I(X,Y) = 0$ since $H(X,Y) = H(X) + H(Y)$: this follows from the chain rule, since $H(Y|X) = H(X)$ when X and Y are independent. Conversely, if Y is determined by X then $H(Y|X) = 0$ and $H(X,Y) = H(X)$ and $I(X,Y) = H(Y)$.

Combining the mutual information and the chain rule give two other expressions

$$I(X,Y) = H(X) - H(X|Y) \tag{33}$$

and

$$I(X,Y) = H(Y) - H(Y|X) \tag{34}$$

Rewriting the entropies in terms of their definitions and fiddling around with sums and mass functions gives a more direct formula

$$I(X,Y) = \sum_{x,y} p_{X,Y}(x,y) \log_2 \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \tag{35}$$

It can be proved that $I(X,Y) \geq 0$, with equality if and only if X and Y are independent. A consequence of this, as mentioned above, is that $H(X) \geq H(X|Y)$.

The mutual information is a powerful object; it determines the relatedness of random variables irrespective of how the variables are related. It is useful to compare it to the correlation; correlation is often used to measure how related to variables are:

$$\text{corr}(X,Y) = \frac{\langle (X-\mu_X)(Y-\mu_Y) \rangle}{\sigma_X \sigma_Y} \tag{36}$$

where μ_X is the average of X and σ_X is its standard deviation. Now the correlation relies on the sample space being a vector space, as we discussed above; the mutual information makes no such assumption. Beyond this though, the correlation only measures a certain type of relation and does not completely characterize the nature of the dependence between them. As an extreme example, consider the random variables X and $Y = X^2$ where X is -1 or one with probability a quarter each, and zero with probability a half. Now

$$\langle (X-\mu_X)(Y-\mu_Y) \rangle = \frac{1}{4}(-1)\frac{1}{2} + \frac{1}{4}\frac{1}{2} = 0 \tag{37}$$

even though the variables are completely dependent. $I(X,Y) = 0$. For real data, however, the mutual information can be difficult to estimate accurately.

The data processing inequality

A Markov chain is a triplet of random variables X, Y and Z where, roughly speaking Z only learns about X through Y; it isn't that X and Z are independent, just that if you know the value of Y knowing the value of X wouldn't tell you anything more. A game of snakes and ladders is an example: your game position after three goes is Z in this story, after two goes is Y and after one is X. Now Z depends on X, if you went up a ladder in your

first go you are more likely have a high value at the end of your third go. But if you knew your game position after two goes, you wouldn't be any better at predicting your position after three goes if you knew where you were after one.

More formally we write

$$X \to Y \to Z \tag{38}$$

and say X, Y, Z are a Markov chain if

$$p(x, z|y) = p(x|y)p(z|y) \tag{39}$$

In other words, if you know $Y = y$ the conditional distributions of X and Y are independent. The data processing inequality states that in this situation

$$I(X, Y) \geq I(X, Z) \tag{40}$$

Thus Z can't know any more about X than Y does. In a way, this is formulating the obvious: if the relationship between Z and X 'comes through' Y then no amount of cunning processing allows you to make Z more informative about X than Y is. However, knowing this as a fact is useful and clarifies our thinking, it means we know, for example, that V1 of the visual cortex has no more information about what you are looking at than your thalamus does, because the visual information comes from the outside world to your retina, from their to the thalamus and hence on to the V1. Thus the purpose of the visual pathway can't be to add to our information about the world. Instead, it is to extract from the information coming from the retina that part that is useful.