Information theory

Kristian Wichmann

April 18, 2017

1 Self-information or surprisal

Let X be a random variable. Consider an event A. We may ask ourselves how much information $I(A)$ - also known as *self-information* or *surprisal* - we have gained by having this event occuring. It is clear, that such a quantity must depend only on the probability of the event:

$$I(A) = I(P(A)) \tag{1.1}$$

Therefore, we can express self-information through a function $f(p)$, so that if $P(A) = p$, then $I(A) = f(p)$.

If the outcome of an event A is certain, i.e. if $P(A) = 1$ then we have gained no information. So we must have $P(A) = 1 \Rightarrow I(A) = 0$. or in other words $f(1) = 0$. Non-certain events occuring, on the other hand, should give us non-zero information. So for $p < 1$ we should have $f(p) > 0$.

Further, if two events A and B are independent it seems reasonable to require that self-information is additive is the following sense:

$$I(A \cap B) = I(A) + I(B) \tag{1.2}$$

So if two independent events happen at the same time, self-information should simply add up. Because of independence, we also have:

$$P(A \cap B) = P(A) \cdot P(B) \tag{1.3}$$

Applying f to both sides of this equation we get:

$$I(A \cap B) = f(P(A) \cdot P(B)) \tag{1.4}$$

Combine this with equation (1.2) to get:

$$f(P(A) \cdot P(B)) = f(P(A)) + f(P(B)) \tag{1.5}$$

The only functions having this property are logarithms. Hence, the self-information must be of the form:

$$f(p) = -k \cdot \log(p) \tag{1.6}$$

The minus sign comes from requiring $f(p) > 0$ for $p < 1$. This means that k will be positive, but apart from that can be chosen freely. Since all logarithms are proportional to each other, this is equivalent to choice of base b being free:

$$f(p) = -\log_b(p) \tag{1.7}$$

1.1 Continous distributions?

The section above deals with discrete random variables? However, we run into problems if we try to mindlessly generalize to continuous variables: The "obvious" analogue of the self-information for the outcome $X = x$ the would be $-\log_b f(x)$, where $f(x)$ is the probability density function of X. But since this function need not be below 1, the associated "surprisal" may actually be negative! Clearly, something is fishy.

1.1.1 Surprisal for an interval

The problem is in essence, that a probability density is not a probability. Instead, the probability of finding X realized in an interval of size Δx close to x is approximately:

$$P(x \leq X < x + \Delta x) \approx f(x)\Delta x \tag{1.8}$$

The surprisal for this event is thus approximately:

$$-\log_b(f(x)\Delta x) = \log_b f(x) - \log_b \Delta x \tag{1.9}$$

In the limit $\Delta x \to 0$ this should become exact. But here the logarithm tends to minus infinity, and so surprisal is infinite! Which makes sense: Since there's a continuum of outcomes, any specific outcome has probability zero, and hence is infinitely surprising.

2 Entropy

The *entropy* of a discrete random variable X is the expectation value of the self-information:

$$H(X) = E[I(X)] = E[-\log_b(X)] \tag{2.1}$$

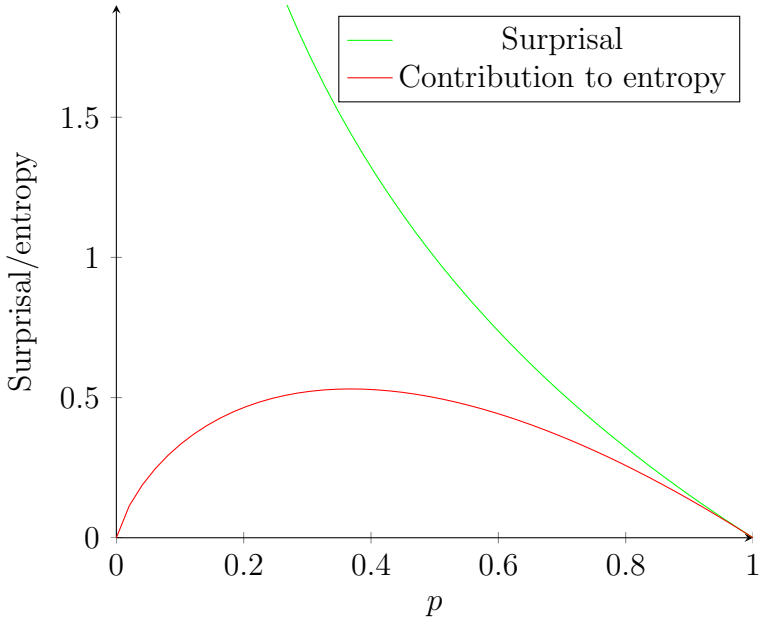

Figure 1: Surprisal and contribution to entropy as a function of p. Here for $b = 2$.

Here, $I(X)$ is itself a stochastic variable. Thus, entropy can be interpreted as the expected surprisal. Since X is discrete, we may write:

$$H(X) = -\sum_x p(x) \log_b p(x) \tag{2.2}$$

Figure 1 shows how much an outcome of p contributes to the total entropy. Since the limit for $p \to 0$ tends to zero, we will extend the definition to outcomes with zero probability; these do not contribute to the entropy.

2.1 Different choices of b

As mentioned above, we're free to choose b, but some choices are common. Each carry its own unit of entropy with it:

- $b = 2$: The corresponding entropy is known as *Shannon entropy*, and the unit is Shannon or simply bits.
- $b = e$: The corresponding unit is known as a nat.
- $b = 10$: The corresponding unit is known as a Hartley.

Unless explicitly mentioned, we will use Shannon entropy from now on.

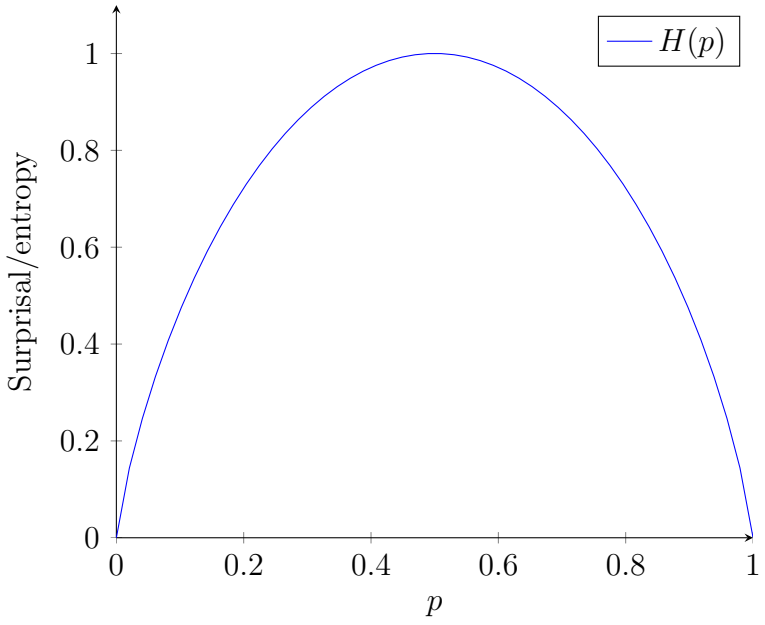

Figure 2: Entropy for an unfair coin toss.

2.2 Example: Entropy of a coin toss

Let's consider the simplest possible non-trivial situation: and experiment with two outcomes, A and B. If the probability of A is p, then the probability of B must be $1-p$. For a fair coin, both probabilities are $\frac{1}{2}$, and the entropy is:

$$H = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = \frac{1}{2} + \frac{1}{2} = 1 \tag{2.3}$$

If the coin is not fair, but the probability of tails (A) is p, instead we get:

$$H(p) = -p\log_2 p - (1-p)\log_2(1-p) \tag{2.4}$$

This function is plotted in figure 2.

2.3 Entropy of a continuous distribution

As we saw above, the surprisal of any outcome from a continous probability distribution is infinite, but it's not inherently clear whether the entropy should diverge as well. This turns out to actually be true. Writing out the steps in the derivation will actually be useful later. So, let's pick an interval size Δx and write an approximate entropy as a sum:

$$H(X) \approx -\sum_{n\in\mathbb{Z}} P(x_n \le X < x_{n+1})\log_2 P(x_n \le X < x_{n+1}) \tag{2.5}$$

Here $x_n = n\Delta x$. The probability is approximately $f(x_n)\Delta x$:

$$H(X) \approx -\sum_{n\in\mathbb{Z}} f(x_n)\Delta x \log_2(f(x_n)\Delta x) \tag{2.6}$$

Use the usual formula for the product of a log to split it into two sums:

$$\underbrace{-\sum_{n\in\mathbb{Z}} f(x_n)\log_2(f(x_n))\Delta x}_{\to \int f(x)\log_2 f(x)\ dx} - \log_2(\Delta x)\underbrace{\sum_{n\in\mathbb{Z}} f(x_n)\Delta x}_{\to 1} \tag{2.7}$$

The underbraces show what happens when $\Delta x \to 0$. So, abusing the notation somewhat, the entropy of X is:

$$H(X) = -\int f(x)\log_2 f(x)\ dx - \lim_{\Delta x\to 0}\log_2 \Delta x \tag{2.8}$$

The first term is the *differential entropy*, which is the immediate generalization of equation 2.2. But we also know that this is not necessarily positive, since probability distributions are not bounded between 0 and 1. This is not absurd, because we need to take the second, infinite term - a *logarithmic divergence* - into account. The differential entropy, as absurd as it may seem, measures the deviation from this infinity is some sense. But the total entropy is infinite.

2.3.1 Maximum differential entropy

We might now ask the question: Is there a probability distribution which maximizes differential entropy? Naturally, that the answer to this question depends on our constraints, including the domain/dominating measure of the distribution function. For distributions with all real numbers/Lebesgue measure as dominating measure, with finite mean and variance, the maximum differential entropy distribution is none other than the normal distribution! This is one of the many ways to characterize the Gaussian.

3 Conditional entropy and mutual information

3.1 Entropy of a joint distribution

Given two discrete random variables X and Y, we may define their *joint entropy* simply as the entropy of their joint distribution:

$$H(X,Y) = -\sum_{x,y} P(X=x, Y=y)\log_2\left(P(X=x, Y=y)\right) \tag{3.1}$$

Or using the joint distribution function $p(x, y) = P(X = x, Y = y)$:

$$H(X, Y) = -\sum_{x,y} p(x, y) \log_2 p(x, y) \tag{3.2}$$

3.2 Conditional entropy

Similarly, we may define *conditional entropy*. If we already know the outcome of one random variable, this will limit the number of outcomes that contributes to the entropy. But the probabilities become conditional:

$$H(X|Y = y) = -\sum_{x} p(x|y) \log_2 p(x|y) \tag{3.3}$$

Since $p(x|y) = \frac{p(x,y)}{p(y)}$ this means:

$$H(X|Y = y) = -\sum_{x} \frac{p(x, y)}{p(y)} \log_2 \frac{p(x, y)}{p(y)} \tag{3.4}$$

The total conditional entropy is found by weighing all of these:

$$H(X|Y) = \sum_{y} p(y) \cdot H(X|Y = y) \tag{3.5}$$

Here $p(y) = \sum_x p(x, y)$ is the marginal probability for Y. (Similarly $p(x) = \sum_y p(x, y)$). Inserting equation 3.4 into equation 3.5 we get:

$$H(X|Y) = -\sum_{y} p(y) \sum_{x} \frac{p(x, y)}{p(y)} \log_2 \frac{p(x, y)}{p(y)} = -\sum_{xy} p(x, y) \log_2 \frac{p(x, y)}{p(y)} \tag{3.6}$$

Looking at figure 3, we would expect that $H(X, Y) - H(Y) = H(X|Y)$. Let's check that this is indeed the case:

$$H(X, Y) - H(Y) = -\sum_{x,y} p(x, y) \log_2 p(x, y) - \left(-\sum_{y} p(y) \log_2 p(y) \right) \tag{3.7}$$

Use the definition of $p(y)$:

$$\sum_{y} p(y) \log_2 p(y) = \sum_{xy} p(x, y) \log_2 p(y) \tag{3.8}$$

Hence:

$$H(X, Y) - H(Y) = -\sum_{x,y} p(x, y) \left(\log_2 p(x, y) - \log_2 p(y) \right) \tag{3.9}$$

This is the same as:

$$-\sum_{x,y} p(x, y) \log_2 \frac{p(x, y)}{p(y)} = H(X|Y) \tag{3.10}$$

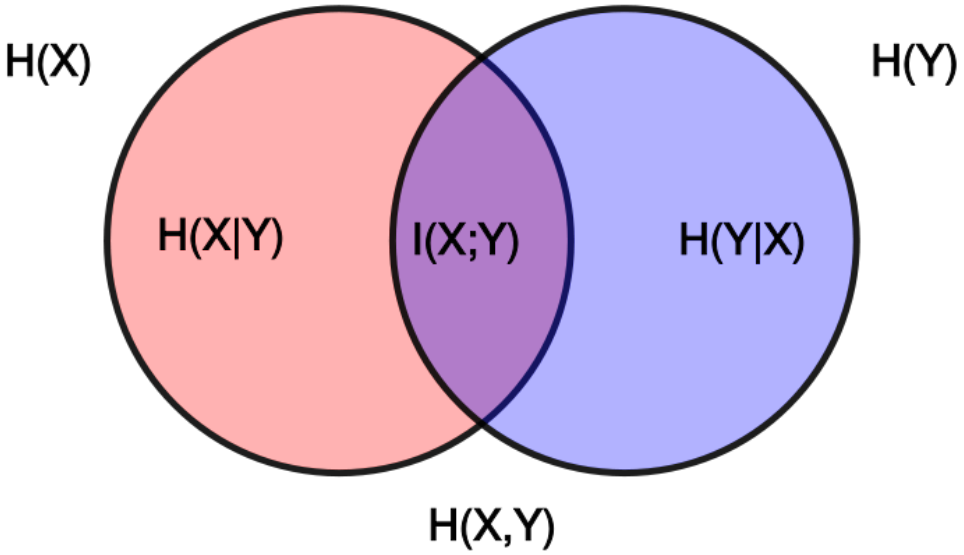

Figure 3: Visualization of the different entropies and mutual information. Image source: Wikipedia.

3.3 Mutual information

If we look at figure 3 once again, we also notice $I(X;Y)$, the *mutual information* between two discrete random variables X and Y. It should be equal to:

$$I(X;Y) = H(X) - H(X|Y) \tag{3.11}$$

Inserting we get:

$$-\sum_x p(x)\log_2 p(x) - \left(-\sum_{x,y} p(x,y)\log_2 \frac{p(x,y)}{p(y)}\right) \tag{3.12}$$

Now use $p(x) = \sum_y p(x,y)$ to rewrite the first term:

$$\sum_x p(x)\log_2 p(x) = \sum_{xy} p(x,y)\log_2 p(x) \tag{3.13}$$

Combine terms to get:

$$I(X;Y) = \sum_{x,y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)} \tag{3.14}$$

This is clearly symmetrical: $I(X;Y) = I(Y;X)$. Also, if X and Y are independent, then $p(x,y) = p(x)p(y)$ which makes the fraction 1 and logarithm zero, so the mutual information vanishes in this case.

3.4 Example: The Monty Hall problem

In the famous Monty Hall problem, you're presented with three doors, behind which is a prize. After making an initial choice of doors, Monty opens one empty door that you have not chosen, and offers you to change your choice.

As is known, when choosing the first door, all choices are equally likely to get you the prize. If we call this experiment X, the entropy is:

$$H(X) = \frac{1}{3}\log_2 3 + \frac{1}{3}\log_2 3 + \frac{1}{3}\log_2 3 = \log_2 3 \approx 1.58 \tag{3.15}$$

After the door has been opened, however, there's two options with probabilities of 1/3 and 2/3 respectively (it's advantageous to change your mind). If we call the event/experiment associated with opening the door Y, we might write this as:

$$H(X|Y) = \frac{1}{3}\log_2 3 + \frac{2}{3}\log_2 \frac{3}{2} = \tag{3.16}$$

$$\frac{1}{3}\log_2 3 + \frac{2}{3}(\log_2 3 - \log_2 2) = \log_2 3 - \frac{2}{3} \approx 0.92 \tag{3.17}$$

In other words, the mutual information is:

$$I(X;Y) = H(X) - H(X|Y) = \log_2 3 - \left(\log_2 3 - \frac{2}{3}\right) = \frac{2}{3} \tag{3.18}$$

So the information gain by Monty opening the door is 2/3 bits.

One might ask how much this is compared to the original uncertainty. This quantity is known as the *percentual information gain* or PIG for short:

$$\text{PIG} = \frac{I(X;Y)}{H(X)} = \frac{2/3}{\log_2 3} \approx 42.1\% \tag{3.19}$$

So Monty gives away about 42% of the information away by opening the door. Somewhere, Douglas Adams is smiling ...

Now, let's consider the situation from Monty's point of view. He has to choose a door. If he had to choose it without the player having chosen one yet, he would have two option with an equal probability - he cannot open the door with the prize behind it. So the entropy is the same as for a fair coin toss:

$$H(Y) = H = \frac{1}{2}\log_2 + \frac{1}{2}\log_2 = 1 \tag{3.20}$$

However, once the player has chosen a door, there's two situations to consider:

- If the player has indeed chosen the door with the prize behind it, the situation is the same as above:

$$H(Y|X = \text{prize}) = 1 \tag{3.21}$$

- If the player has chosen an empty door, there is only one choice for Monty, namely the other empty door. So the entropy vanished in this case:

$$H(Y|X = \text{no prize}) = 0 \tag{3.22}$$

We now use equation 3.5 to get the total conditional entropy:

$$H(Y|X) = \frac{1}{3} \cdot 1 + \frac{2}{3} \cdot 0 = \frac{1}{3} \tag{3.23}$$

Let's check that we get the same amount of mutual information here:

$$I(X;Y) = H(Y) - H(Y|X) = 1 - \frac{1}{3} = \frac{2}{3} \tag{3.24}$$

Indeed we do!

Finally, we may ask what Monty's percentual information gain by the player choosing a door is:

$$\text{PIG} = \frac{I(X;Y)}{H(Y)} = \frac{2/3}{1} = \frac{2}{3} \approx 66.7\% \tag{3.25}$$

3.5 Continuous distributions

3.5.1 Joint entropy

Following the same logic as in section 2.3, we may write the joint entropy is two continuous variables as:

$$H(X,Y) \approx -\sum_{n,m\in\mathbb{Z}} p_{mn} \log_2 p_{mn} \tag{3.26}$$

Here, $p_{nm} = P(x_n \leq X < x_{n+1}, y_m \leq Y < y_{m+1})$, which is in turn approximately $f(x,y)\Delta x \Delta y$, where f is the joint probability density. Splitting the logarithm, we end up with:

$$H(X,Y) = -\int f(x,y) \log_2 f(x,y)\ dx\ dy - \lim_{\Delta x\to 0} \log_2 \Delta x - \lim_{\Delta y\to 0} \log_2 \Delta y \tag{3.27}$$

Again, this is to be taken with the same grain of salt as before.

4 Cross-entropy and the Kullback-Leibler divergence

4.1 Comparing probability distributions

Let's say we have two probability distributions p and q for the discrete random variables X. The entropy of X will differ according to which distribution we use:

$$H_p(X) = E_p[-\log_2 p(x)] = -\sum_x p(x) \log_2 p(x) \tag{4.1}$$

$$H_q(X) = E_q[-\log_2 q(x)] = -\sum_x q(x) \log_2 q(x) \tag{4.2}$$

Here, we've introduced subscripts on the expectation operator, depending on which distributaion is used. Now, we might use the distribution of p to evalute the q-surprise expectation:

$$H_{pq}(X) = E_x[-\log_2 p(y)] = -\sum_x p(x) \log_2 q(x) \tag{4.3}$$

This is known as the *cross-entropy*. This is often written as $H(p,q)$, but this alternative notation is used here to to avoid confusion with the joint entropy. Note that in general this is not a symmetric function: $H_{pq}(X) \neq H_{qp}(X)$. Therefore, even is the cross-entropy in some sense measures a difference between the two distributions, it cannot be a metric.

4.2 Kullback-Leibler divergence

The *Kullback-Leibler divergence* from q to p (note the appearantly reversed order) is defined as:

$$D_{KL}(p||q) = H_{pq}(X) - H_p(X) \tag{4.4}$$

Here, the X is understood on the left hand side. Using equations 4.1 and 4.3 this is:

$$-\sum_x p(x)\log_2 q(x) + \sum_x p(x)\log_2 p(x) = \sum_x p(x)\log_2 \frac{p(x)}{q(x)} \tag{4.5}$$

In summary:

$$D_{KL}(p||q) = \sum_x p(x)\log_2 \frac{p(x)}{q(x)} \tag{4.6}$$

Like cross-entropy, this is asymmetric, so while it still can be used to compare distributions, it is not a metric on distribution space. Hence the name divergence.

4.3 Connection to mutual information

Remember that the mutual information (or information gain) between the random variables X and Y is:

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x,y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)} \tag{4.7}$$

Here, we've used the formula from equation 3.14. But this is the same as the KL divergence between the distribution functions $p(x,y)$ and $p(x)p(y)$:

$$I(X;Y) = H_{KL}(p(x,y)||p(x)p(y)) \tag{4.8}$$

Again we see that mutual information in some sense measures distance to independence.

5 Model selection

If we're in a situation where we have a model which outputs a probability distribution p based of the inputs, and we have a target distribution of y, then we can use the KL-divergence and cross-entropy as a criterion for selecting a model that is good in some sense.

5.1 Example: Classification

A supervised classification problem is a primary example of the situation described above. We have a target distribution given by the one-hot encoding of the label for a given sample: $y_i = \delta_{ij}$, where j is the label. The model will have a number of parameters which we wish to to tune, and the output distribution p depends on these. Based on the last section, a reasonable way to choose parameters, is to require that y deviates as little from p as possible. In other words, we seek to minimize the KL-divergence $D_{KL}(y||p)$. Let's write this out:

$$D_{KL}(y||p) = H_{yp} - H_y \tag{5.1}$$

The entropy of y is simply zero. And more importantly, it's a constant, so it will not matter when minimizing, even if the target distribution was more complex. So model selection has been reduced to choosing the parameters which minimizes the cross-entropy H_{yp}. For one sample, this is:

$$-\sum_x y(x) \log_2 p(x) \tag{5.2}$$

Here x runs over the possible cases the classification distinguishes between. When we have n samples, it is customary to use the average cross-entropy:

$$-\frac{1}{n} \sum_{i=1}^{n} \sum_x y_i(x) \log_2 p_i(x) \tag{5.3}$$

In machine learning, this is known as a *cost function*. This quantity is then minimized, often using some variation of gradient descent.