Information theory

Kristian Wichmann

November 1, 2016

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 continous 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 probablity density function of X. But since this function need not be below 1, the associated surprisal is actually negative! Clearly, something is fishy. But for now, we will only consider discrete random variables.

2 Entropy

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

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

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:

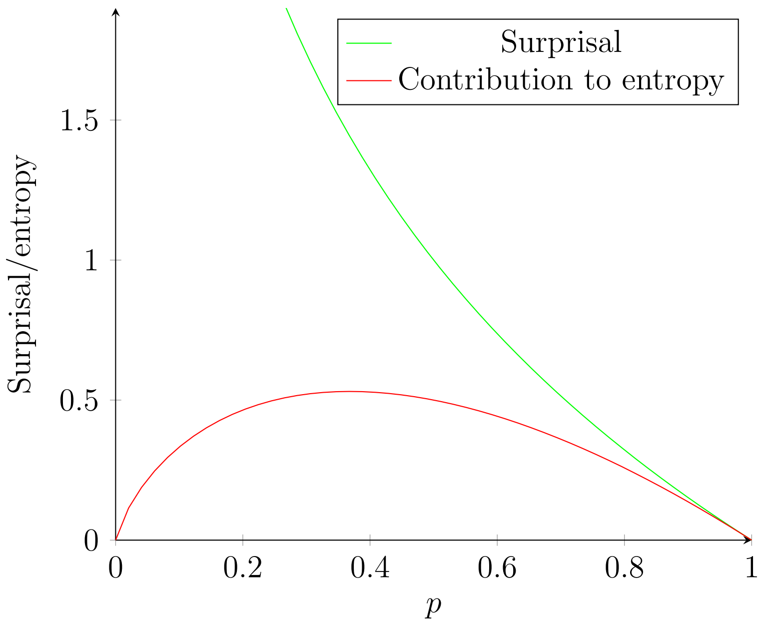

Figure 1: Surprisal and contribution to entropy as a function of p. Here for $b = 2$.

- $b = 2$: 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.