Basics of information theory

Michael Franke

Surprisal, entropy, Kullback-Leibler divergence, mutual information.

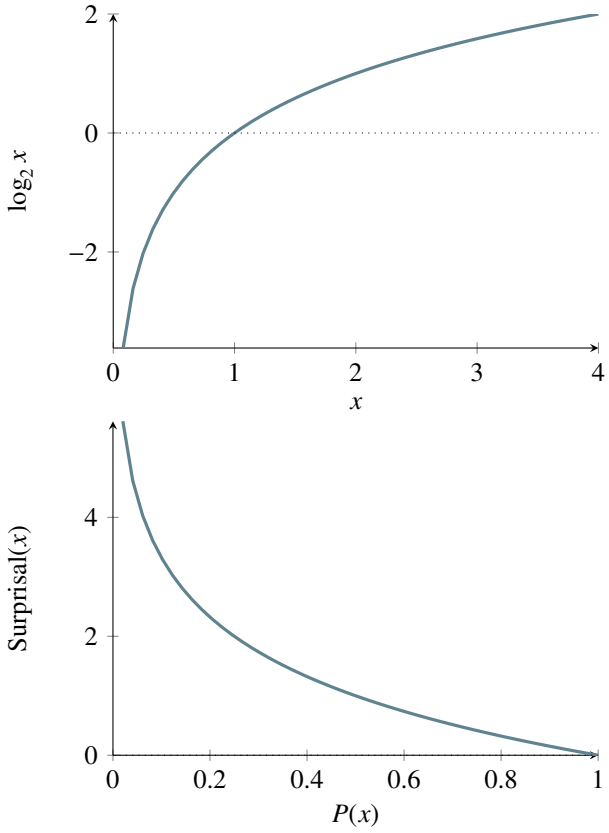

Figure 1: Logarithm and surprisal (to base 2).

1 Information content (surprisal)

Let X be a random variable with support χ. For event $x \in \chi$, the *information content* $I_X(x)$ of x (a.k.a. *surprisal* of x) under random variable X is defined as:

$$I_X(x) = -\log_2 P(X = x)$$

Intuitively speaking, the information content $I_X(x)$ is a measure of how surprised an agent with beliefs X is (alternatively: how much the agent learns) when they observe x.

2 *Entropy*

The *entropy* $\mathcal{H}(X)$ of a random variable is the expected information content under the assumption that the true distribution is X:[1]

$$\begin{aligned}\mathcal{H}(X) &= \sum_{x \in \mathcal{X}} P(X = x)\ I_X(x) \\ &= -\sum_{x \in \mathcal{X}} P(X = x)\ \log P(X = x)\end{aligned}$$

[1] Here and below, writing "true distribution" or similar formulations does not necessarily entail a strong commitment to actual truth. It is shorthand for more careful but cumbersome language like "the distribution used as a reference or baseline which we assume to be true or treat as-if true."

Intuitively speaking, the entropy $\mathcal{H}(X)$ measures the expected (or average) surprisal of an agent whose beliefs are X when the true distribution is X.

- example
- joint and conditional entropy

3 *Cross entropy*

The *cross entropy* $\mathcal{H}(X, Y)$ of random variables X and Y measures the expectation of information content given Y from the point of view of (assumed true) distribution X:

$$\begin{aligned}\mathcal{H}(X, Y) &= \sum_{x \in \mathcal{X}} P(X = x)\ I_Y(x) \\ &= -\sum_{x \in \mathcal{X}} P(X = x)\ \log P(Y = x)\end{aligned}$$

Intuitively speaking, the cross entropy $\mathcal{H}(X, Y)$ measures the expected (or average) surprisal of an agent whose beliefs are Y when the true distribution is X.

4 *Kullback-Leibler divergence (relative entropy)*

The *Kullback-Leibler (KL) divergence* (also known as *relative entropy*) measures the expected (or average) difference in information content between the distribution Y and the true distribution X:

$$\begin{aligned}D_{KL}(X \| Y) &= \sum_{x \in \mathcal{X}} P(X = x)\ (I_Y(x) - I_X(x)) \\ &= \sum_{x \in \mathcal{X}} P(X = x)\ \log \frac{P(X = x)}{P(Y = x)}\end{aligned}$$

Intuitively speaking, the KL-divergence $D_{KL}(X \| Y)$ measures how much more surprised an agent is, on average, when they hold beliefs described by Y instead of the true distribution X.

KL-divergence $D_{KL}(X \| Y)$ can be equivalently written in terms of the entropy $\mathcal{H}(X)$ of X and the cross entropy $\mathcal{H}(X, Y)$ of X and Y:

$$D_{KL}(X \| Y) = \mathcal{H}(X, Y) - \mathcal{H}(X)$$

- examples
- not a metric

5 *Mutual information*