Basics of information theory

Michael Franke

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

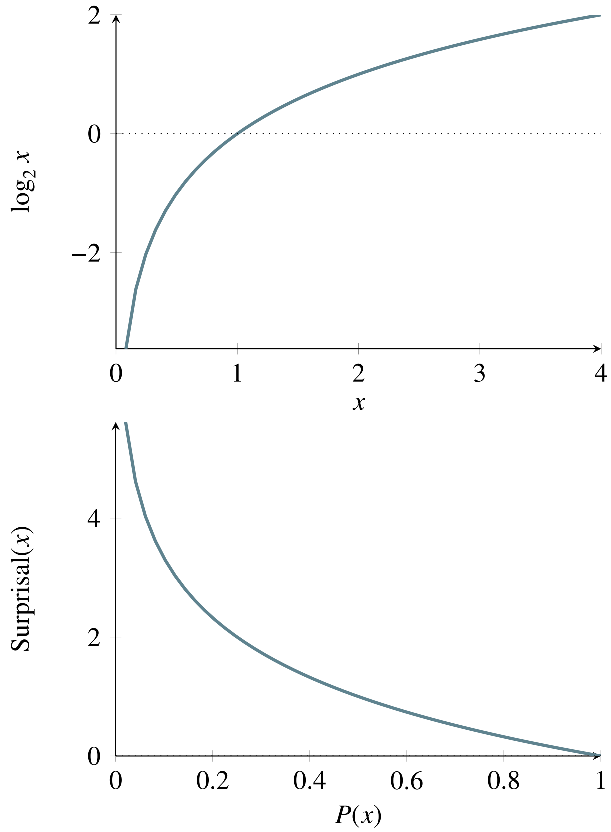

Figure 1: Logarithm and surprisal (to base 2).

1 *Information content (surprisal)*

Let $P \in \Delta(X)$ be a probability distribution over (finite) set X. For event $x \in \mathcal{X}$, the *information content* $I_X(x)$ of x (a.k.a. *surprisal* of x) under random variable X is defined as:

$$I_P(x) = -\log_2 P(x)$$

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

2 Entropy

Let $P \in \Delta(X)$ be a probability distribution over (finite) set X. The *entropy* $\mathcal{H}(P)$ of probability distribution P is the expected information content under the assumption that the true distribution is P:[1]

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

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

- example
- joint and conditional entropy

[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."

3 Cross entropy

Let $P, Q \in \Delta(X)$ be probability distributions over (finite) set X. The *cross entropy* $\mathcal{H}(P, Q)$ of probability distributions P and Q measures the expectation of information content given Q from the point of view of (assumed true) distribution P:

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

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

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 $Q \in \Delta(X)$ and the true distribution $P \in \Delta(X)$:

$$\begin{aligned}D_{KL}(P||Q) &= \sum_{x \in X} P(x)\ \left(I_Q(x) - I_P(x)\right) \\ &= \sum_{x \in X} P(x)\ \log \frac{P(x)}{Q(x)}\end{aligned}$$

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

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

$$D_{KL}(P\|Q) = \mathcal{H}(P,Q) - \mathcal{H}(P)$$

- examples
- not a metric

5 *Mutual information*

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P(Z)$$