Basics of information theory

Michael Franke

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

The goal of this primer on information theory is to introduce the most salient notions of information theory relevant to common applications in fields like machine learning, computational linguistics or theoretical linguistics. Rather than appealing to deeper mathematical results (such as related to noisy communication channels and efficient coding), this primer explains and motivates the relevant notions from the perspective of (subjective) beliefs.

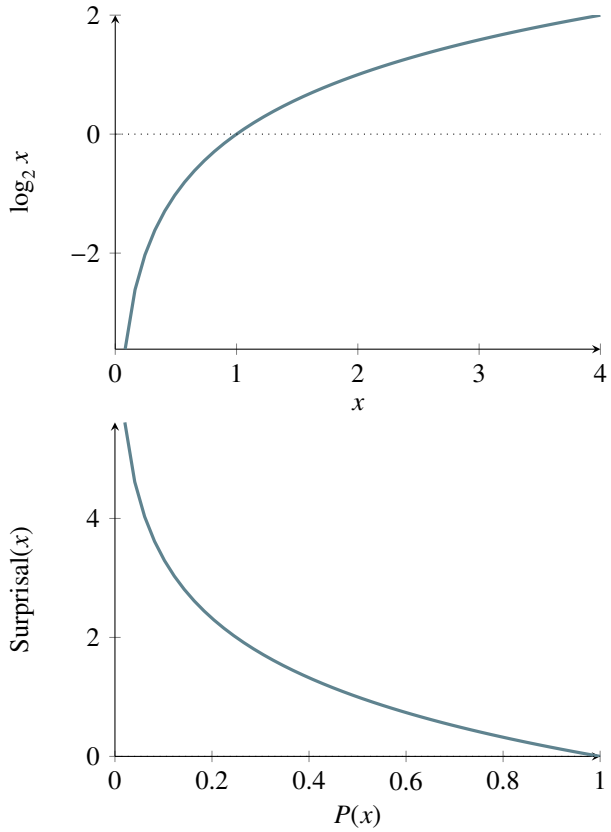

Figure 1: Logarithm and surprisal (to base 2).

1 Information content (surprisal)

Let $P \in \Delta(X)$ be a probability distribution over (finite) set X. For event $x \in X$, the *information content* $I_P(x)$ 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]

$$\begin{aligned}\mathcal{H}(P) &= \sum_{x \in X} P(x)\ I_P(x) \\ &= -\sum_{x \in X} P(x)\ \log_2 P(x)\end{aligned}$$

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

- example

2.1 Joint entropy

We can generalize the notion of entropy to more than one probability distribution, via a joint probability distribution that encompasses them. Let $R \in \Delta(X \times Y)$ by a joint probability distribution over the set of all pairs in $X \times Y$, and let $P \in \Delta(X)$ and $Q \in \Delta(Y)$ be the *marginal distributions* of X and Y respectively. The *joint entropy* $\mathcal{H}(P, Q)$ of P and Q is usually defined as:

$$\mathcal{H}(P, Q) = -\sum_{x \in X} \sum_{y \in Y} R(x, y) \log_2 R(x, y)$$

which is just the entropy of the joint probability distribution R:

$$\mathcal{H}(P, Q) = \mathcal{H}(R) = -\sum_{z \in X \times Y} R(z) \log_2 R(z)$$

2.2 Conditional entropy

Let $R \in \Delta(X \times Y)$ by a joint probability distribution over the set of all pairs in $X \times Y$, and let $P \in \Delta(X)$ and $Q \in \Delta(Y)$ be the *marginal distributions* of X and Y respectively. The conditional entropy of Q given P is the expected surprisal of observing an y after having updated our beliefs with the corresponding x:

$$\begin{aligned}\mathcal{H}(Q \mid P) &= -\sum_{\langle x,y \rangle \in X \times Y} R(x, y) \log_2 R(y \mid x) \\ &= -\sum_{\langle x,y \rangle \in X \times Y} R(y \mid x) P(x) \log_2 R(y \mid x) \\ &= -\underbrace{\sum_{x \in X} P(x)}_{\text{prob. of } x} \underbrace{\sum_{y \in Y} R(y \mid x) \log_2 R(y \mid x)}_{\text{entropy of } R(\cdot \mid x)}\end{aligned}$$

The conditional entropy therefore measures how surprised, on average, an agent with veridical beliefs R is about observations of y after having updated their beliefs already based on an observation of the corresponding x.

- example

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 $P \in \Delta(X)$:

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

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

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