A primer on information theory (for the subjectively perplexed)

Michael Franke

Surprisal, entropy (joint, conditional, cross), 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.

1 A measure of information gained

At the heart of information theory lies a numerical measure of the amount of information learned, so-called *information content* or *surprisal*. Let us motivate this measure based on three intuitive desiderata.

1.1 Motivation

Suppose that Jones and Smith are uncertain about the weather tomorrow at noon. There are only three possible states of the weather $X = \{\text{sunny}, \text{misty}, \text{rainy}\}$. Jones' and Smith's subjective beliefs are given in Table 1.

	sunny	cloudy	rainy
Jones' beliefs	0.6	0.2	0.2
Smith's beliefs	0.1	0.2	0.7

Table 1: Subjective beliefs about the weather.

The next day, Jones and Smith both observe that it is in fact sunny. Who learns more? Who is more surprised by this turn of events? — Intuitively, since Jones had expected sunny weather, he gains less information than Smith, who had though that this event is rather unlikely. In sum, a measure of the "information learned" should be sensitive to an agent's prior expectations $P \in D(X)$. We therefore aim to define a numerical measure $I_P(x)$, which assigns a number representing the information gained by observing x *in*X, given prior beliefs $P \in \Delta(X)$.

There are three further desiderata on the measure $I_P(x)$:

1. **Lower bound.** If an agent is already maximally certain that x would occur, so that $P(x) = 0$, the agent learns nothing from observing x, so that $I_P(x) = 0$.

2. **Monotonicity.** The lower the prior probability of $P(x)$ the more information is gained from (the more surprised the agent is by) observing an event

$x \in X$.

3. **Additivity.** If an agent observes two independent events $x_1, x_2 \in X$, the total information gained should be the sum of the information gained from each individual information: $I_P(x_1 \& x_2) = I_P(x_1) I_P(x_2)$.

The family of functions that satisfies these desiderata is that of negative logarithms (up to a scaling factor). Most frequently, the logarithm to base 2 is used.

1.2 *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.

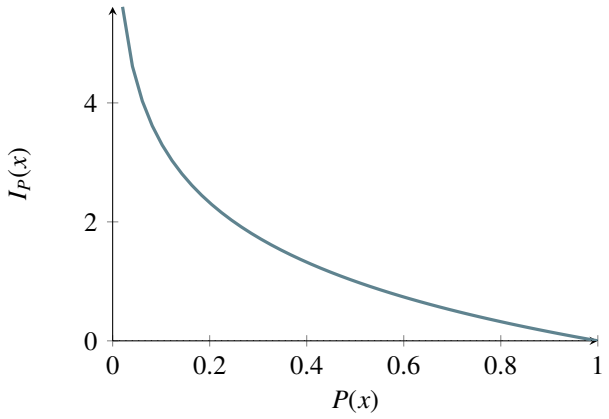

Figure 1: Information content / surprisal (to base 2) for an event x as a function of its prior probability $P(x)$.

Notice that an agent who assigns $P(x) = 0$ to some event, and sees x happening, is infinitely perplexed (has their mind blown).

2 *Measures of expected information content*

The basic measures used in applications of information theory fall in one of two categories. For one there are *measures of expected information content*. These measures follow the general format:

$$\sum_{x \in X} \text{Probability of } x \times \text{Info-content of } x$$

and are discussed in this section. Different measures of this format differ in what kind of distributions are used to define the probability of x and the information content of x.

The other class of information-theoretic measures commonly used in applications are *measures of expected difference in information content*. These measures instantiating are dealt with in Section 3, and they instantiate the template:

$$\sum_{x \in X} \text{Probability of } x \times (\text{ Inf.cont } x \text{ wrt. Q } - \text{ Inf.cont. } x \text{ wrt. P })$$

where information content is measured based on two different distributions P and Q.

2.1 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. Entropy can also be interpreted as a measure of uncertainty: the higher the entropy of P the more uncertain an agent an agent with beliefs P is about X.

- example

2.2 Joint entropy

Joint entropy is entropy for joint probability distributions.[2] Let $R \in \Delta(X \times Y)$ by a joint probability distribution over the set of all pairs in the structured event space $X \times Y$, and let $P \in \Delta(X)$ and $Q \in \Delta(Y)$ be the *marginal distributions* over X and Y respectively.[3] The *joint entropy* $\mathcal{H}(P, Q)$ of P and Q is defined as:[4]

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

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

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

- example

[2] Here, we only look at structured spaces with two dimensions. Joint entropy can be generalized to further dimensions in analogous manner.

[3] So, $P(x) = \sum_{y \in Y} R(x, y)$ and similarly for Q.

[4] A usual definition of joint entropy is in terms of random variables X and Y and reads as $\mathcal{H}(X, Y) = -\sum_{x \in X} \sum_{y \in Y} P(x, y) \log_2 P(x, y)$, where it is implicitly assumed that there is an encompassing joint probability distribution $P(x, y)$ over pairs of numbers. Giving the definition in the way we do here makes clear how joint entropy is really nothing special at all, except when you muffle the joint distribution through intransparency with random variable notation.

2.3 *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* over X and Y respectively. The conditional entropy of Q given P is the expected surprisal of observing y after having updated beliefs with the corresponding x:

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

We can also think about conditional entropy as the average uncertainty of an agent about dimension Y after observing dimension X.

- example

2.4 *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.

- example

3 *Measure of expected difference in information content*

3.1 *Kullback-Leibler divergence (relative entropy)*

The *Kullback-Leibler (KL) divergence* (also known as *relative entropy*) measures the expected 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

4 *Mutual information*

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* over X and Y respectively. The *mutual information* $I(P, Q)$ of P and Q is the expected excess surprisal if the dimensions X and Y are assumed to be stochastically independent:

$$I(X, Y) = \sum_{\langle x,y\rangle \in X \times Y} R(x, y) \log \frac{R(x, y)}{P(x)\ Q(y)}$$

We can write this more intelligibly in terms of the Kullback-Leibler divergence between true distribution R and the distribution $S \in \Delta(X \times Y)$ which is derived from P and Q by assuming that the dimensions X and Y are stochastically independent, so that $S(x, y) = P(x)\ Q(x)$:

$$I(X, Y) = D_{KL}(R \,\|\, S) = \sum_{\langle x,y\rangle \in X \times Y} R(x, y) \log \frac{R(x, y)}{S(x, y)}$$

Intuitively, we may think of mutual information as a measure of how much more (needlessly) surprised an agent is who believes X and Y are stochastically independent (while having correct beliefs of then marginal distributions).