A primer on information theory

Michael Franke

Notions covered: information content / surprisal, entropy (joint, conditional, cross), Kullback-Leibler divergence, mutual information.

Information-theoretic notions —like entropy, Kullback-Leibler divergence or mutual information— have many useful applications in fields like statistics, machine learning or computational linguistics. These notions are anchored in mathematical results related to efficient coding of information and communication via noisy channels. However, in order to understand which notion to use when and why in concrete applications, this theoretical grounding in deep and important, but abstract mathematical results is not always relevant, sometimes possibly even a distraction. For the student interested in applications, explanations in terms of "minimal number of questions asked" or similar can be needlessly confusing at first. This primer therefore attempts to give a systematic overview of the most salient notions of information theory in terms of what is arguably the pre-theoretically most intuitive perspective: vocabulary about the subjective beliefs of agents.

1 *A measure of information gained*

The centerpiece of information theory is a numerical measure of the amount of information gained, so-called *information content* or *surprisal*.

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, Smith learns more from the observation of sunny weather (is more surprised by it) than Jones, because Jones had deemed sunny weather rather likely, while Smith had though that this event is rather unlikely.

This example demonstrates that a measure of the "information gained" from the observation of an even $x \in X$ should be sensitive to a baseline probability distribution $P \in \Delta(X)$ such as an agent's prior expectations. 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: $I_P(x_1 \& x_2) = I_P(x_1) + I_P(x_2)$.

1.2 *Information content (surprisal)*

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.

Let $P \in \Delta(X)$ be a probability distribution over (finite) set X.[1] 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 shows the information content $I_P(x)$ as a function of $P(x)$. Notice that an agent who assigns $P(x) = 0$ to some event, and sees x happening, is infinitely perplexed (has their mind blown).

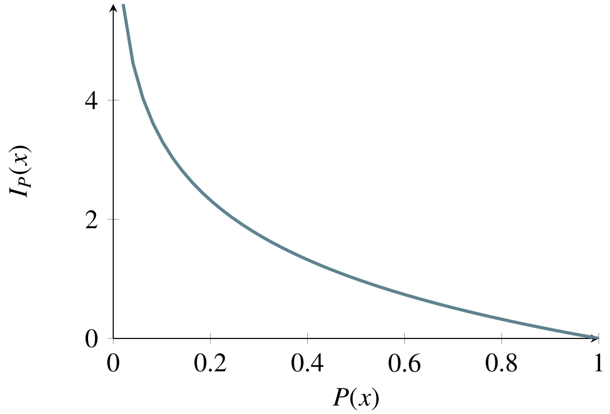

Figure 1: Information content / surprisal (to base 2) for an event x as a function of its prior probability $P(x)$.

[1] All definitions in this primer are given only for finite sets. Analogous definitions exist for uncountable sets. Intuitions behind concepts remain the same and are the focus at present.

1.3 *Excursion: Background on logarithms*

The logarithm is the inverse function to exponentiation:

$$x = a^y \quad \Leftrightarrow \quad y = \log_a x$$

Here, $a > 0$ is called the *base* of the logarithm. Since a^y is always positive, the logarithm takes as input only positive numbers x. Figure 2 shows the logarithm to base 2.

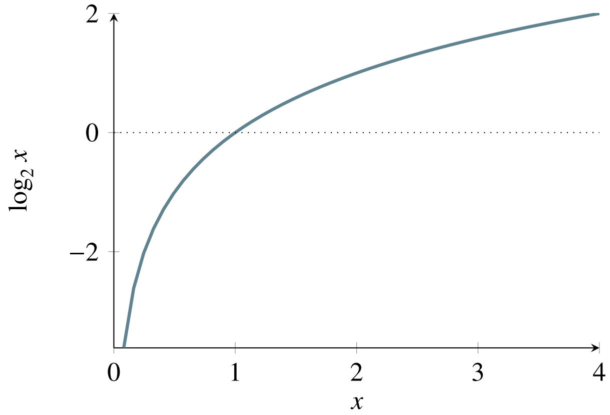

Figure 2: Logarithm to base 2.

For present purposes, the following calculation rules are important:[2]

$$\log_a(xy) = \log_a x + \log_a y$$
$$\log_a \frac{x}{y} = \log_a x - \log_a y$$

Notice that we can also transform bases of logarithms as follows:

$$\log_b x = \frac{\log_a x}{\log_a b}$$

Since $\log_a b$ is a constant (independent of x), the choice of base in the definition of information content is, mathematically speaking, arbitrary.[3]

2 *Measures of expected information content*

The most frequently used measures of information theory fall into one of two categories. For one there are *measures of expected information content*. These measures are discussed in this section. They follow the general format:

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

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.

[2] The first equation is what justifies using logarithms to fulfill the desideratum “additivity” given above. The second equation will be used in more compact formulations of Kullback-Leibler divergence and mutual information.

[3] Our choice of 2 does relate to ideas from theoretical computer science, though, namely the idea to encode information in binary code. But that is not necessary to understand the rationale for the use of information theoretic notions in general.

The other class of information-theoretic measures commonly used in applications are *measures of expected difference in information content*. These measures 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 P is the expected information content under the assumption that the true distribution is P:[4,5]

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

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 with beliefs P is about X.

Example 1. The entropy of Jones' beliefs in Table 1 is:

$$\begin{aligned}\mathcal{H}(P_J) = -\sum_{x \in \{\text{sunny,cloudy,rainy}\}} & P_J(x)\ \log_2 P_J(x) \\ &= -(0.6 \log_2 0.6 + 0.2 \log_2 0.2 + 0.2 \log_2 0.2) \\ &\approx 1.37\end{aligned}$$

A similar calculation for Smith's beliefs yields: $\mathcal{H}(P_S) \approx 1.16$. Smith is slightly less uncertain than Jones about the state of the weather.

2.2 *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:

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

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] 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."

[5] We follow common practice here and assume that $0 \log_a 0 = 0$. This is justified because $\lim_{x \to 0^+} x \log_a x = 0$.

Example 2. Suppose that Jones is exactly right. The beliefs P_J capture the precise probability of the weather in the limit (when not taking the weather conditions on the previous days into account). Smith's beliefs are therefore not quite in line with reality. The cross entropy $\mathcal{H}(P_J, P_S)$ then measures Smith's average surprisal by taking the real-world frequencies P_J to form the expectation and by using Smith's beliefs P_S to define the surprisal:

$$\begin{aligned}\mathcal{H}(P_J, P_S) &= -\sum_{x \in X} P_J(x) \ \log P_S(x) \\ &= -(0.6 \log_2 0.1 + 0.2 \log_2 0.2 + 0.2 \log_2 0.7) \\ &\approx 2.56\end{aligned}$$

2.3 *Joint entropy*

Joint entropy is entropy for joint probability distributions.[6] Let $R \in \Delta(X \times Y)$ by a joint probability distribution over the set of all pairs in the structured (finite) 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.[7] The *joint entropy* $\mathcal{H}(P, Q)$ of P and Q is defined as:[8]

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

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 3. Clark also has beliefs about the weather, but theirs are a joint belief $R \in \Delta(X \times Y)$ about the condition of the weather ($X \in \{\text{sunny}, \text{cloudy}, \text{rainy}\}$) and whether the swallows fly high or low in the evening ($Y \in \{\text{high}, \text{low}\}$), as shown in Table 2.

	sunny	cloudy	rainy	∑ rows
high	.6 × .4 = .24	.2 × .4 = .08	.2 × .4 = .08	.4
low	.1 × .6 = .06	.2 × .6 = .12	.7 × .6 = .42	.6
∑ columns	.3	.2	.5	

Table 2: Clark's joint probabilistic belief about the flying of swallows in the evening and the weather condition the next day..

The joint entropy of Clark's beliefs can be calculated as:

$$\begin{aligned}\mathcal{H}(R) &= -\sum_{x \in \{\text{sunny,cloudy,rainy}\}} \sum_{y \in \{\text{high,low}\}} R(x, y) \ \log_2 R(x, y) \\ &= -(.24 \log_2 .24 + .08 \log_2 .08 + .08 \log_2 .08 + .06 \log_2 .06 + \dots) \\ &\approx 2.22\end{aligned}$$

2.4 *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*

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

[7] $P(x) = \sum_{y \in Y} R(x, y)$.

[8] Frequently used definitions of joint entropy use random variables X and Y and read 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 is more general (beyond random variables), but also makes clear how joint entropy is really nothing special at all, except when you muffle the joint distribution through intransparency with random variable notation.

over (finite) X and Y respectively. The conditional entropy of Q given P is the expected surprisal of observing y after having already observed (and processed) a corresponding x:

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

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

Example 4. Suppose that Clark observes the swallows every night and then makes a weather prediction. If $P \in \Delta(X)$ captures Clark's marginal beliefs about the weather and $Q \in \Delta(Y)$ those about the swallows flying, then, from Clark's subjective point of view, the expected surprisal of their weather predictions after having observed the swallows is given by the conditional entropy:

$$\begin{aligned}\mathcal{H}(P \mid Q) = -\sum_{y\in Y} Q(y)\sum_{x\in X} R(x \mid y)\log_2 R(x \mid y)&\\ = Q(\text{high})\ [R(\text{sunny} \mid \text{high})\ \log_2 R(\text{sunny} \mid \text{high})&\\ + R(\text{cloudy} \mid \text{high})\ \log_2 R(\text{cloudy} \mid \text{high})&\\ + R(\text{rainy} \mid \text{high})\ \log_2 R(\text{rainy}\ \mid \text{high})]+&\\ Q(\text{low})\ [R(\text{sunny} \mid \text{low})\ \log_2 R(\text{sunny} \mid \text{low})&\\ + R(\text{cloudy} \mid \text{low})\ \log_2 R(\text{cloudy} \mid \text{low})&\\ + R(\text{rainy} \mid \text{low})\ \log_2 R(\text{rainy}\ \mid \text{low})]&\\ = 0.4[0.6\log_2 0.6 + 0.2\log_2 0.2 + 0.2\log_2 0.2]+&\\ 0.6[0.1\log_2 0.1 + 0.2\log_2 0.2 + 0.7\log_2 0.7]&\\ \approx 1.01&\end{aligned}$$

3 *Measures of expected difference in information content*

Measures of expected difference in information content essentially compare two probability distributions P and Q following the pattern:

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

In general, these measures can be thought of answers to a question like: "how much more, on average, would an agent using Q be (needlessly) surprised if the true distribution is P?"

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 $\mathcal{H}(P, Q)$:[9]

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

Example 5. Let's assume, again, that Jones' beliefs P_J are the true distribution. We want to compute the divergence of Smith's P_S beliefs from those of Jones'.

$$\begin{aligned} D_{KL}(P_J \,\|\, P_S) &= \sum_{x \in X} P_J(x) \log \frac{P_J(x)}{P_S(x)} \\ &= 0.6 \log \frac{0.6}{0.1} + 0.2 \log \frac{0.2}{0.2} + 0.2 \log \frac{0.2}{0.7} \\ &\approx 1.19 \end{aligned}$$

If, reversely, we assume that Smith's beliefs are the ground-truth, and compute the divergence of Jones' beliefs from Smith's the result is:

$$\begin{aligned} D_{KL}(P_S \,\|\, P_J) &= \sum_{x \in X} P_S(x) \log \frac{P_S(x)}{P_J(x)} \\ &= 0.1 \log \frac{0.1}{0.6} + 0.2 \log \frac{0.2}{0.2} + 0.7 \log \frac{0.7}{0.2} \\ &\approx 1.01 \end{aligned}$$

This shows that KL-divergence is not a symmetric measure.[10]

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 (finite) 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:

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

[9] This reformulation shows that KL-divergence and cross-entropy are exchangeable when the task is to find a distribution Q that approximates a true distribution P (e.g., in machine learning). KL-divergence and cross-entropy are exchangeable in such an optimization context since the only difference is a constant additive term $\mathcal{H}(P)$ which does not depend on the to-be-optimized Q.

[10] KL-divergence is therefore not a distance metric (like geometrical distance).

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 about the marginal distributions).

Example 6. Clark has the joint probability distribution $R \in \Delta(X \times Y)$ in Table 2, whose marginal distributions are $P \in \Delta(X)$ (beliefs about the weather) and $Q \in \Delta(Y)$ (beliefs about the swallows). Clark's beliefs are repeated here from above:

	sunny	cloudy	rainy	∑ rows
high	.24	.08	.08	.4
low	.06	.12	.42	.6
∑ columns	.3	.2	.5	

Jackson has a joint probability distribution $S \in \Delta(X \times Y)$ which assigns the same probabilities to the (marginal) events "weather condition" and "swallows' flying height," but which assumes that these dimensions are independent, so that $S(x, y) = P(x)\, Q(y)$. This means that Jackson's beliefs are these:

	sunny	cloudy	rainy	∑ rows
high	.12	.08	.20	.4
low	.18	.12	.30	.6
∑ columns	.3	.2	.5	

If Clark's beliefs are the reference distribution, how much more surprised will

Clark be about the occurrence of pairs $\langle x, y\rangle$ on average?

$$
\begin{aligned}
I(X,Y) = D_{KL}(R \,\|\, S) = \sum_{\langle x,y\rangle \in X\times Y} R(x,y) \log \frac{R(x,y)}{S(x,y)} \\
&= R(\text{sunny}, \text{high}) \ \log_2 \frac{R(\text{sunny}, \text{high})}{Q(\text{sunny}, \text{high})} \\
&\qquad + R(\text{cloudy}, \text{high}) \ \log_2 \frac{R(\text{cloudy}, \text{high})}{Q(\text{cloudy}, \text{high})} \\
&\qquad + R(\text{rainy}, \text{high}) \ \log_2 \frac{R(\text{rainy} \ , \text{high})}{Q(\text{rainy} \ , \text{high})} \\
&\quad + R(\text{sunny}, \text{low}) \ \log_2 \frac{R(\text{sunny}, \text{low})}{Q(\text{sunny}, \text{low})} \\
&\qquad + R(\text{cloudy}, \text{low}) \ \log_2 \frac{R(\text{cloudy}, \text{low})}{Q(\text{cloudy}, \text{low})} \\
&\qquad + R(\text{rainy}, \text{low}) \ \log_2 \frac{R(\text{rainy} \ , \text{low})}{Q(\text{rainy} \ , \text{low})} \\
&\approx 0.24
\end{aligned}
$$

Exercise 1. Roberts and Carpenter hold the following beliefs about the weather:

	sunny	cloudy	rainy	storm
Roberts' beliefs	0.1	0.6	0.2	0.1
Carpenter's beliefs	0.4	0.1	0.3	0.2

Calculate the entropy of their beliefs. Who is more uncertain?

Exercise 2. Let the true distribution over weather conditions (ignoring dependencies between consecutive days) be:

	sunny	cloudy	rainy	storm
ground truth	0.7	0.1	0	0.2

Calculate the cross-entropy of Robert's and Carpenter's belief with respect to this reference distribution. Can you guess who will be more suprised on average?

Exercise 3. Calculate the KL-divergence between the reference distribution from the previous exercise and Robert's and Carpenter's beliefs (each). (Hint: you might be able to use some results from other exercises.)

Exercise 4. Proof that indeed, as claimed above, $D_{KL}(P \,||\, Q) = \mathcal{H}(P, Q) - \mathcal{H}(P)$

Exercise 5. Decide for each of the following claims whether it is true or false?

(i) Kullback-Leibler divergence is a special case of mutual information.

(ii) Kullback-Leibler divergence is a measure of expected information content of the reference distribution.

(iii) Joint entropy measures the entropy of joint probability distributions.

(iv) Entropy is a measure of the information gained from a single observation.

(v) Mutual information could be used to measure the extent to which making an assumption of stochastic independence is severe.