Entropy 101

Boaz Barak

October 2, 2012

Note: for more on information theory and its applications to theoretical computer science, I recommend the web pages of the courses by Anup Rao and Mark Braverman on this topic, see `http://www.cs.princeton.edu/courses/archive/fall11/cos597D/`, `https://catalyst.uw.edu/workspace/anuprao/15415/86751`. In particular you can see Lecture 1 in Mark's course and Lecture 3 in Anup's course for a lot of the material below.

Entropy If we have a random variable X (always in our case over $\{0,1\}^k$ for some k) then the *entropy* of X, denoted $H(X)$, aims to capture how many bits of information it contains, or how much uncertainty it has.

For example, if X is uniform over $\{0,1\}^k$ then $H(X) = k$. On the other hand, if $X = x...x$ for random $x \in \{0,1\}$ (i.e., the same bit repeated k times) then $H(X) = 1$.

More generally, if X is uniform over a subset $S \subseteq \{0,1\}^k$ and $|S| = 2^m$ then $H(X) = m$. A distribution of this form is called a *flat* distribution, and in many cases one can pretend that the distribution we deal with is flat without much loss in understanding.

Formal definition The formal definition of entropy for every X ranging on some domain D is

$$H(X) = \sum_{x \in D} \Pr[X = x] \cdot \log\left(1/\Pr[X = x]\right)$$

Exercise: $H(X) \leq \log|D|$ with equality iff X is the uniform distribution over D.

Entropy and independence Suppose we have two (possibly correlated) random variables X, Y ranging over $\{0,1\}^k$. What can we say about $H(XY)$ where XY is the variable over $\{0,1\}^{2k}$ obtained by concatenating X and Y.

Assume the flat case, so X is uniform over S_X and Y is uniform over S_Y.

Clearly, XY is contained in $S_X \times S_Y$ and so by the exercise

$$H(XY) \leq \log(|S_X| \cdot |S_Y|) = H(X) + H(Y)$$

The case where equality holds is when X and Y are independent and then XY will be uniform over $S_X \times S_Y$.

The other extreme case is when $X = Y$ (i.e. X and Y are perfectly correlated), in which case we'll have

$$H(XY) = H(X)$$

Mutual information The example above hints that the relation between $H(XY)$ and $H(X) + H(Y)$ captures the amount of dependence between X and Y. This motivates the definition of *mutual information* between X and Y (denoted $I(X;Y)$). This aims to capture the question of "how many bits about X do you learn from seeing Y?" (or vice versa- it turns out to be symmetric). The definition is as follows

$$I(X;Y) = H(X) + H(Y) - H(XY)$$

We see from the examples above that if X and Y are independent then $I(X;Y) = 0$, while if $X = Y$ then $I(X;Y) = k$.

The following fact will be useful for us:

Exercise: If X and Y are independent and Z is arbitrary then $I(X;Z)+I(Y;Z) \leq I(XY;Z)$.

This exercise is a bit harder than the one above but the intuition behind this is that if you can learn m bits about XY from Z, then because X and Y are independent, then it must be that m' of those bits come from X and m'' of them come from Y where $m' + m'' = m$.

Note that this inequality can sometimes be strict as is evidenced by the following example: X and Y are independent random bits and $Z = X \oplus Y$ (can you see why the inequality is strict in this case?).

Conditional quantities Sometimes we'll need to consider a setting where we *condition* on the value of some other random variable W that may be correlated with the others. All the quantities we consider such as entropy, mutual information, etc.. can be extended to this case, where the idea is that we pick w at random from W and then considered the other random variables conditioned on the event $W = w$.

So, the entropy of X conditioned on W, denoted $H(X|W)$ is equal to $\mathbb{E}_{w \in W} H(X_w)$ where X_w is the random variable X conditioned on the event $W = w$. Sometimes we write $X|W = w$ for X_w.

The mutual information of X and Y conditioned on W, denoted $I(X;Y|W)$ is equal to $\mathbb{E}_{w \in W} I(X_w; Y_w)$. By linearity of expectation this equals $\mathbb{E}_w H(X_w) + \mathbb{E}_w H(Y_w) - E_w H(X_w Y_w) = H(X|W) + H(Y|W) - H(XY|W)$.

For example, if X and Y are random bits and $W = X \oplus Y$ then one can see that X and W are independent, as well as X and Y. However, if we know that $W = w$, the variable XY has only 2 possibilities (the two pairs who XOR to w) and so $H(XY|W) = 1$. Thus, $I(XY|W) = 2 - 1 = 1$. In other words, given W, you can learn the bit X from Y and vice versa.

Clearly, if Y is independent of X then $H(Y|X) = H(X)$. On the other hand if $Y = X$ then $H(Y|X) = 0$. So, we see that if X and Y are independent then $H(XY) = H(X) + H(Y) = H(X) + H(Y|X)$, while if $X = Y$ then we have $H(XY) = H(X) = H(X) + H(Y|X)$. In fact this equality holds in all other cases as well and is known as the "entropy chain rule"

Exercise: Prove that for all X, Y, $H(XY) = H(X) + H(Y|X)$.

Distances between distributions When are two random variables X, Y over the same domain D close to one another? It turns out "closeness' can be defined in several ways, though for our purposes they'll all be equivalent. We let $P_X, P_Y \in \mathbb{R}^D$ be the vectors of probabilities corresponding to X and Y. That is for every $w \in D$, $P_X(w) = \Pr[X = w]$ and $P_Y(w) =$

$\Pr[Y = w]$. We say that X and Y are "close" if the vectors P_X and P_Y are close to one another. Let us now define this more formally:

Statistical/ Total variation distance The *total variation* (also known as *statistical* or L_1) distance between X and Y is defined as

$$\Delta_{\mathsf{TV}}(X,Y) = \tfrac{1}{2}|P_X - P_Y|_1 = \tfrac{1}{2}\sum_{w\in D}|P_X(w) - P_Y(w)|$$

(The factor 1/2 will not be significant for any of our discussions; it added for normalization to ensure that $\Delta_{\mathsf{TV}}(X,Y) \leq 1$ for all X, Y.)

Hellinger distance The *Hellinger distance* between X and Y is defined as

$$\Delta_{\mathsf{Hel}}(X,Y) = \tfrac{1}{2}\|\sqrt{P_X} - \sqrt{P_Y}\|_2 = \left(\tfrac{1}{2}\sum_{w\in D}|\sqrt{P_X(w)} - \sqrt{P_Y}(w)|^2\right)^{1/2}$$

(Again the 1/2 factor is insignificant and added for normalization)

Note that the vectors $\sqrt{P_X}$ and $\sqrt{P_Y}$ are unit vectors and so the Hellinger distance is (up to the 1/2 factor) just their Euclidean distance.

The following is implied by standard properties of the L_1 and L_2 norms:

Exercise: Prove that both Δ_{TV} and Δ_{Hel} are valid distance functions. That is, they satisfy symmetry ($\Delta(X,Y) = \Delta(Y,X)$), positivity ($\Delta(X,Y) \geq 0$ with $\Delta(X,Y) = 0$ iff $X = Y$), and triangle inequality ($\Delta(X,Z) \leq \Delta(X,Y) + \Delta(Y,Z)$).

Another useful fact is obtained by the standard relations between the 2 norm and inner product:

Exercise: $\Delta_{\mathsf{Hel}}(X,Y)^2 = 1 - \langle\sqrt{P_X}, \sqrt{P_Y}\rangle = 1 - \sum_{w\in D}\sqrt{P_X(w)P_Y(w)}$

Relation between Hellinger and TV distance It turns out that for our purposes, the Hellinger and TV distance are equivalent— when one is small then so is the other, as is shown by this exercise

Exercise: $\frac{1}{2}H(X,Y)^2 \leq \Delta_{\mathsf{TV}}(X,Y) \leq \frac{1}{2}H(X,Y)$.

Entropy, independence, distance If two random variables X and Y are close to being independent, then intuitively the distribution XY should be close to being the product distribution $X \times Y$ (obtained by taking an independent copy of X and an independent copy of Y). In fact, one can show this is true in the following quantitative sense:

Exercise: (harder) $\Delta_{\mathsf{Hel}}(XY, X\times Y)^2 \leq I(XY; X\times Y)$

Note/hint: the way this is typically proven is by showing that $\Delta_{\mathsf{Hel}}(XY, X\times Y)^2 \leq \Delta_{\mathsf{KL}}(X\|Y) = I(XY; X\times Y)$ where for Δ_{KL} is the *Kullback-Leibler divergence*, defined as

$$\Delta_{\mathsf{KL}}(Z\|W) = \sum_{w\in D}\Pr[Z=w]\log\tfrac{\Pr[Z=w]}{\Pr[W=w]}$$