CS838-1 Advanced NLP: Information Theory

Xiaojin Zhu

2007
Send comments to jerryzhu@cs.wisc.edu

What is the *entropy* of English?

1 Entropy

Entropy of a discrete distribution $p(x)$ over the event space X is

$$H(p) = -\sum_{x \in X} p(x) \log p(x). \tag{1}$$

When the log has base 2, entropy has unit *bits*. Properties: $H(p) \geq 0$, with equality only if p is deterministic (use the fact $0 \log 0 = 0$). Entropy is the average number of 0/1 questions needed to describe an outcome from $p(x)$ (the Twenty Questions game). Entropy is a concave function of p.

For example, let $X = \{x_1, x_2, x_3, x_4\}$ and $p(x_1) = \frac{1}{2}, p(x_2) = \frac{1}{4}, p(x_3) = \frac{1}{8}, p(x_4) = \frac{1}{8}$. $H(p) = \frac{7}{4}$ bits.

This definition naturally extends to joint distributions. Assuming $(x, y) \sim p(x, y)$,

$$H(p) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y). \tag{2}$$

We sometimes write $H(X)$ instead of $H(p)$ with the understanding that p is the underlying distribution.

The conditional entropy $H(Y|X)$ is the amount of information needed to determine Y, if the other party knows X.

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y|x). \tag{3}$$

From above, we can derive the chain rule for entropy:

$$H(X_{1:n}) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_{1:n-1}). \tag{4}$$

Note in general $H(Y|X) \neq H(X|Y)$. When X and Y are independent, $H(Y|X) = H(Y)$. In particular when $X_{1:n}$ are independent and identically distributed (*i.i.d.*), $H(X_{1:n}) = nH(X_1)$.

2 Mutual Information

Recall the chain rule $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$, from which we see that

$$H(X) - H(X|Y) = H(Y) - H(Y|X). \tag{5}$$

This difference can be interpretted as the reduction in uncertainty in X after we know Y, or vice versa. It is thus known as the *information gain*, or more commonly the *mutual information* between X and Y:

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \tag{6}$$

Mutual information satisfies $I(X; Y) = I(Y; X) \geq 0$. Entropy is also called self-information because $I(X; X) = H(X)$: knowing X gives you all information about X!

3 KL-Divergence

The *Kullback-Leibler (KL) divergence*, also called relative entropy, FROM p TO q is

$$KL(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}. \tag{7}$$

It is often used as a measure of "distance" between the two distributions p, q. However KL-divergence is not a *metric* in that it is asymmetric, and it does not satisfy the triangle inequality:

$$\begin{aligned} KL(p\|q) &= KL(q\|p) \text{ NOT always true} &(8)\\ KL(p\|q) &\leq KL(p\|r) + KL(r\|q) \text{ NOT always true for all } r. &(9)\end{aligned}$$

It has the following properties: $KL(p\|q) \geq 0$, $KL(p\|q) = 0$ iff $p = q$. It is well-defined even if p has less support than q because $0 \log(0/q_i) = 0$. But it is unbounded if q has less support than p since $p_i \log(p_i/0) = \infty$.

If the data is generated from some underlying distribution p (e.g. words in a language), and one wants to find the Maximum Likelihood estimate (MLE) θ^{ML} of p under some model (e.g. unigram), in the limit of infinity data it is equivalent to minimizing the KL-divergence from p to θ:

$$\theta^{ML} = \arg\min_\theta KL(p\|\theta). \tag{10}$$

Mutual information and KL-divergence are connected:

$$I(X; Y) = KL(p(x, y)\|p(x)p(y)). \tag{11}$$

Intuitively, if X, Y are independent, $p(x, y) = p(x)p(y)$, and the KL-divergence is zero, and knowing X gives zero information gain about Y.

The Jensen-Shannon divergence (JSD) is symmetric. It is defined as

$$JSD(p, q) = 0.5KL(p\|r) + 0.5KL(q\|r), \tag{12}$$

where $r = (p + q)/2$. $\sqrt{JSD}$ is a metric.

4 Cross Entropy

Say $x \sim p(x)$ (e.g., the true underlying distribution of language), but we model X with a different distribution $q(x)$ (e.g., a unigram language model). The *cross entropy* between X and q is

$$H(X, q) = H(X) + KL(p\|q) = -\sum_x p(x) \log q(x). \tag{13}$$

This is the average length of bits needed to transmit an outcome x, if you thought $x \sim q(x)$ (and build an optimal code for that), but actually $x \sim p(x)$. $KL(p\|q)$ is the extra price (bits) you pay for the model mismatch.

5 The Entropy Rate of a Language

The entropy of a word sequence of length n is

$$H(w_{1:n}) = -\sum_{w_{1:n}} p(w_{1:n}) \log p(w_{1:n}). \tag{14}$$

This quantity depends on n, so a length normalized version is known as the *entropy rate* of a language L, when n approaches infinity:

$$H(L) = \lim_{n\to\infty} \frac{1}{n} H(w_{1:n}) = \lim_{n\to\infty} -\frac{1}{n} \sum_{w_{1:n}} p(w_{1:n}) \log p(w_{1:n}). \tag{15}$$

The Shannon-McMillan-Breiman theorem states that the above entropy rate can be computed with

$$H(L) = \lim_{n\to\infty} -\frac{1}{n} \log p(w_{1:n}), \tag{16}$$

when $w_{1:n}$ is sampled from p. Basically ONE typical sequence is enough. Note p appeared twice above: once to generate the sequence $w_{1:n}$, and once to compute the probability $p(w_{1:n})$.

In reality we never know p, but we have a corpus $w_{1:n}$ sampled from p. We nevertheless have a language model q, from which we can compute the *cross entropy rate* of the language:

$$H(L, q) = \lim_{n\to\infty} -\frac{1}{n} \log q(w_{1:n}). \tag{17}$$

It can be shown that $H(L, q) \geq H(L)$. The better q is, the tighter the upper bound. And because we only have a finite corpus, we end up with an approximation

$$H(L, q) \approx -\frac{1}{n} \log q(w_{1:n}). \quad (18)$$

For example, English letters (a-z, space) has been estimated to have the following cross entropy:

q	cross entropy (bits)
0-gram	4.76 (uniform, $\log_2 27$)
1-gram	4.03
2-gram	2.8
IBM word trigram	1.75
Shannon game (human)	1.3

A Shannon game demo can be found at `math.ucsd.edu/~crypto/java/ENTROPY`.

Perpelxity is related by $PP(L, q) = 2^{H(L,q)}$.