Note

Derivation of Shannon Entropy

Yue-Xin Huang

School of Sciences, Great Bay University, Dongguan 523000, China.

信息熵，即香农熵，本质上刻画的是一个系统的不确定度，不确定度越高说明系统所包含的信息量越多，用 S 来表示信息量。比如抛骰子，如果每次事件开始就已经知道它是六点朝上，对于抛骰子的我们来说，每次事件不会给我们任何信息量 $S = 0$ 。而如果这个骰子是均匀的，每一面朝上的概率都是 $1/6$ ，不确定性最高，那么每次事件的信息量将是最大。

现在假定一个事件有 n 种可能的结果，每个结果对应的概率为 p_i ，由概率的归一性 $\sum_{i=1}^n p_i = 1$ ，我们的任务是寻找一个刻画这个事件不确定度的函数

$$S(p_1, p_2, \dots, p_n). \quad (0.1)$$

由上面的讨论，可以确定这个函数的一些性质：

- $S(p_1, p_2, \dots, p_n)$ 是一个连续函数；
- $S(p_1 = 1, p_2 = 0, \dots, p_n = 0) = 0$ ，相当于某一个结果已经确定发生，因此不包含任何信息量；
- 如果 $n = 2$ ，必然是 $p_i = 1/2$ 时 S 最大。如果不是，说明预先知道了某个事件的概率 p_k 较大，事件更偏向于其中的某个结果，因此信息量降低，这可以推广到任意 n ：对于一个固定 n ， $S(1/n, 1/n, \dots, 1/n)$ 取最大值；
- n 越大，可能的结果多，说明不确定度越高，因此 $f(n) = S(1/n, 1/n, \dots, 1/n)$ 是一个单调递增函数；
- 对于联合事件的信息熵（事件 A 和 B 发生的信息熵）可以看作是 A 的信息熵，加上发生 A 后 B 的条件信息熵，因此必然有 $S(AB) \leq S(A) + S(B)$ 。联合信息熵满足以下关系

$$S(AB) = S(A) + \sum_{k=1}^n p_k S(B|A_k), \quad (0.2)$$

$S(B|A_k)$ 表示发生 A_k 后的信息量。

现在讨论最后个性质。首先，如果事件 A 和 B 是独立事件，那么分别测量两个事件的信息量和同时测量两个事件的信息量必须是相同的，这说明信息熵有可加性，即

$$S(AB) = S(A) + S(B), \quad (0.3)$$

这个等式与(0.2)是相符的，因为此时 $S(B|A) = S(B)$ ，同时 $\sum_{k=1}^n p_k = 1$ 。对于不独立的事件，通过一个简单的例子说明。比如，一个篮子里装着五颜六色的小球 N 个，假定有 m 种颜色，第 k 种颜色的小球有 n_k 个，事件是在这个篮子中挑选出其中一个。也可以把这个事件分成两步，第一步先挑选一个颜色（事件 A ），如选了第 k 种，第二步再在这个颜色的 n_k 个小球中挑选一个小球（事件 B ）。两种策略的结果是一样的，所包含的信息量也应该相同。

第一步中，挑选中对应颜色的概率为 $p_k = \frac{1}{m}$ ，不确定度为 $S(p_1, p_2, \dots, p_m)$ ；第二步中挑选中特定小球的概率为 $1/n_k$ ，因此它的不确定度为 $p_k S(1/n_k, 1/n_k, \dots, 1/n_k)$ 。整体的不确定度

为两个步骤的不确定度之和

$$S(AB) = S(p_1, p_2, \dots, p_m) + \sum_{k=1}^m p_k S(1/n_k, 1/n_k, \dots, 1/n_k). \quad (0.4)$$

把上面讨论的事件 A 和 B 推广到任意事件，即是式(0.2)。再假定 $n_1 = n_2 = \dots = n_m = n$ ，并令 $S(p_1, p_2, \dots, p_m) = f(m)$ 。可以知道 $f(m)$ 表示的是 m 个等概率事件发生的不确定度，所以式(0.4)左边的 $S(AB) = f(N)$ ，即在 N 个小球中任选一个的不确定度，同时 $S(1/n_k, 1/n_k, \dots, 1/n_k) = f(n)$ 。式(0.4)化为

$$f(N) = f(m) + \sum_{k=1}^m p_k f(n) \implies f(mn) = f(m) + f(n). \quad (0.5)$$

因此，可以得到

$$f(m) = C \ln m, \quad (0.6)$$

$C > 0$ 为一个常数。在信息论中，对数通常会以 2 为底，这里用自然对数不会影响结果。如果每种颜色的小球数量不相同，式(0.4)的左边还是 $f(mn)$ ，把式(0.6)代入，可以得到

$$C \ln(mn) = S(p_1, p_2, \dots, p_m) + \sum_{k=1}^m p_k C \ln n_k, \quad (0.7)$$

$$\begin{aligned} \implies S(p_1, p_2, \dots, p_m) &= C \ln N - \sum_{k=1}^m p_k C \ln n_k \\ &= C \sum_{k=1}^m p_k \ln N - C \sum_{k=1}^m p_k \ln n_k \\ &= -C \sum_{k=1}^m p_k \ln \frac{n_k}{N} \\ &= -C \sum_{k=1}^m p_k \ln p_k, \end{aligned} \quad (0.8)$$

这个就是 Shannon entropy，可以验证它满足上文列出的所有性质。

如果令 $C = k_B$ ，同时有等概率假设 $p_k = 1/\Omega$ ， Ω 表示可能的状态数，Shannon 熵就是波尔兹曼熵

$$S = -k_B \sum_{k=1}^{\Omega} \frac{1}{\Omega} \ln \frac{1}{\Omega} = k_B \ln \Omega. \quad (0.9)$$