資訊熵

作者：呂佳軒
審稿：管希聖教授

Entropy

熵是描述系統混亂程度的物理量，他以特別的形式出現在物理定律中，熱力學第二定律是物理定律中極為少數的不等式。

事實上，熵的應用不只侷限在物理。本文想要簡介的資訊理論，就是熵的概念在電腦科學的延伸。

Information

資訊理論是一門相對年輕的學問， 其最早是由 Claude Shannon 於 1948 年所發表的《*A Mathematical Theory of Communication*》奠定基礎，在這篇經典文章中首先需要處理的問題就是要如何定義資訊，以及如何度量資訊。在文章中，資訊被利用其不確定程度來度量。

假設有一個資訊源 X，其具有 n 種不同的輸出，而輸出第 i 種可能的機率為 p_i，X 的不確定度記為 $H(X)$，被稱為資訊熵(Information Entropy)。Shannon 首先列出直觀上 $H(X)$ 應該要滿足的幾項性質：

當 X 輸出的機率稍有不同時，X 的不確定度應該沒有什麼改變，因此預期 $H(X)$ 不應該要有巨大的變化。也就是說，對於每個可能輸出 i 相對應的機率 p_i，$H(X)$ 應該是 p_i 的連續函數。

$\tilde{X}_n$ 是有 n 個輸出且機率是均勻的特殊情形，顯然當 $i > j$ 時，$\tilde{X}_i$ 應當相較 $\tilde{X}_j$ 有更大的不確定性。因此，對於均勻分佈 $p_i = 1/n$，$H(X)$ 是 n 的單調遞增函數。

當可以藉由連續的步驟得到輸出時，$H(X)$ 應該是各步驟的加權總和。

例如一個 X 有 (a, b, c) 三種輸出，其機率分別為

$$p_a = 1/2, p_b = 1/6, p_c = 1/3$$

除了直接輸出之外，也可以分成兩步驟先輸出 (a, bc) ，分別對應到

$$p_a = 1/2, p_{bc} = 1/2$$

當輸出是 bc 時再進一步分開為 (b, c)，此時分別對應到的機率為

$$p'_b = 1/3, p'_c = 2/3$$

因此

$$H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} H\left(\frac{1}{3}, \frac{2}{3}\right)$$

首先考慮有 n 個均勻輸出的簡單情形，令

$$A(s) = H\left(\frac{1}{s}, \dots, \frac{1}{s}\right)$$

對於每個正整數 s 跟 t ，我們可以找到足夠大的 m 跟 n，使得

$$s^m < t^n < s^{m+1}$$

同取對數後整理得

$$\frac{m}{n} < \frac{\log t}{\log s} < \frac{m}{n} + \frac{1}{n}$$

因此對任意的 ε，當 n 足夠大時有

$$\left|\frac{\log t}{\log s} - \frac{m}{n}\right| < \varepsilon$$

另外由性質三知道

$$\begin{aligned} A(s^n) &= H\left(\frac{1}{s^n}, \dots, \frac{1}{s^n}\right) \\ &= H\left(\frac{1}{s}, \dots, \frac{1}{s}\right) + \sum_{i=1}^{s} \frac{1}{s} \cdot H\left(\frac{1}{s}, \dots, \frac{1}{s}\right) \\ &= A(s) + A(s^{n-1}) = nA(s) \end{aligned}$$

由於性質二

$$A(s^m) < A(t^n) < A(s^{m+1})$$

整理後得到

$$\frac{\mathrm{m}}{\mathrm{n}} < \frac{A(t)}{A(s)} < \frac{m}{n} + \frac{1}{n}$$

同樣的對於任意的 ε ，當 n 足夠大時，$\left|\frac{A(t)}{A(s)} - \frac{m}{n}\right| < \varepsilon$。合併二式得到

$$\left|\frac{A(t)}{A(s)} - \frac{\log t}{\log s}\right| < 2\varepsilon$$

這蘊含 $A(t) = K \log t$ ，其中 K 是一常數。

接著考慮一個特殊的訊號源，他共有 N 個可能輸出，可以用 N 個整數 $n_1, \dots, n_N$ 來描述，每個可能輸出的機率為 $n_i / \sum_j n_j$。我們可以進一步將其展開至均勻分佈，這代表

$$A \sum n_i = H(p_1, \dots, p_N) + \sum_i p_i A(n_i)$$

展開 $A(s)$ 之後有

$$\begin{aligned} &K \log \sum n_i \\ &= H(p_1, \dots, p_N) + \sum K p_i \log n_i \end{aligned}$$

整理之後知道

$$H(p_1, \dots, p_N) = -K \sum\nolimits_i p_i \log \frac{n_i}{\sum_j n_j}$$

因此從這三個基本的假設出發，可以推出

$$H(X) = -K \sum_i p_i \log p_i$$

Coding

$H(X)$ 更深層的意義，可以藉由 Shannon 編碼定理來描述。該定理可以不嚴謹的描述成：對於長度為 N 的字串，如果每一個字元都是來自獨立且相同的資訊源 X，則該字串至多可以被壓縮成 $NH(X)$ 長的二進位序列；如果壓縮至 $H(X)$ 還少時，則不能保證可以還原出原本的字串。

這個定理保證了資料壓縮的可能性。舉

例來說，一篇英文文章在考慮大小寫字母，數字和標點，一共由 95 個不同的符號構成。一種最直觀的編碼方式，是利用每個符號 7 bits 長的 ASCII Code ($2^7 = 128 > 95$)， 因此一篇長 N 個符號的文章，可以用 $7N$ bits 來表示。但因為英文字母的出現頻率其實並非均勻的，像是 e 就顯然比 x 要來的常出現，因此當我們真正去計算文章 s 的熵時，會發現其實並不到 7 bits，也代表我們可以找到更有效率的二進位表示方式。

這個定理連結了 $H(X)$ 和我們現在在數位上常用的二進位表示，從上面的例子中不難發現，$H(X)$ 其實代表該資訊源的輸出平均要用多少位元才能儲存，資訊熵的單位為 bit。

Limit of Computation

到這邊為止，資訊熵都只是數學上的推算，合在通訊上的應用。要討論資訊熵和真實物理的連接，我們需要考慮真實世界我們處理資訊的方式。

現今的電腦運算上最小單位為邏輯閘，我們通常可以用真值表來描述邏輯閘。像是及閘 (AND Gate) 可以用以下的真值表來描述。

In1	In2	Out
1	1	1
1	0	0
0	1	0
0	0	0

這代表及閘是一個二輸入一輸出的運算，並且只有當兩個輸入同時為 1 時，輸出才會是 1。我們可以利用類似的方式來描述或閘(OR Gate)，反閘(NOT Gate)等常用邏輯閘。但包括及閘和或閘在內的常見邏輯閘都有一個很重要的性質，稱為不可逆性(irreversibility)。也就是說，當給定邏輯閘輸出時，我們其實並不能確定輸入是什麼，例如當及閘輸出為 1 時，輸入可能為 00，01 或 11。更仔細的觀察會知道，原本具有 2bits 的資訊，但在通過邏輯閘之後只剩下 1 bit，也就是說有 1 bit 的資訊在中間丟失了。

Rolf Landauer 是最早將丟失資訊和物理連接在一起的人，在熱力學中，熵是藉由內能的變化來定義的 $dS = dQ/T$，因此當熵有變化時，一定會連帶的有能量的變化。Landauer's principle 指出，當計算是不可逆的時候，必定得消耗能量，而所消耗的能量下限和其熵的變化有關

$$\Delta E \geq T\Delta S \text{ (bits)}$$

和 Moore's law 相似的，其實有一經驗性法則 Koomey's law 說明每隔一段時

間，同樣運算量所需要消耗的能量就會減半， 而 Landauer's principle 給出了該定律的極限，在現今的架構下，運算效率的改進是有極限的。也因此衍生出了對於可逆計算的研究。

Maxwell’s Demon

資訊理論也可以用來解釋許多傳統熱力學上的悖論，像是很知名的 Maxwell’s demon 悖論，這個悖論的描述是這樣的。

有一個絕熱的空腔，中間有一個絕熱的隔板，隔板上有一個閥門，空腔內有一些理想氣體。原先閥門是打開的，可以讓氣體分子自由通過，因此我們可以預期，空腔兩側的氣體應該會達到熱平衡，有著一樣的速度分布。

現在，有一種不知道從哪裡來的 demon 進入空腔，他擁有辨別氣體分子速度和操縱閥門的能力，這個調皮的 demon 決定選擇性的讓速度超過一定閾值的氣體分子到右側的空腔，低於該閾值的分子到左側空腔。隨著時間的經過，左右空腔的氣體平衡被破壞，右邊空腔的氣體速度分布較左邊來得高，我們可以很輕易的計算出這個過程氣體的熵是減少的。

如果將整個空腔連同 demon 當作是一個封閉的系統，他顯然會是一個封閉系統，根據熱力學第二定律，這個系統的熵應該是不減少的，這中間一定出了什麼問題吧？

將資訊理論加入討論之後，我們會發現 demon 整個操作必然得從氣體分子那得到一些資訊，而 demon 身上累積了資訊因此相對應的熵變化並不會是 0，這整個系統單單考慮氣體分子的熵變化是不夠的。

Reference

Cover, T. M., \& Thomas, J. A. (2005). Elements of Information Theory. https://doi.org/10.1002/047174882x

Koomey, J., Berard, S., Sanchez, M., \& Wong, H. (2011). Implications of Historical Trends in the Electrical Efficiency of Computing. IEEE Annals of the History of Computing, 33(3), 46--54. https://doi.org/10.1109/mahc.2010.28

Landauer, R. (1991). Information is Physical. Physics Today, 44(5), 23--29. https://doi.org/10.1063/1.881299

Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27(3), 379--423. https://doi.org/10.1002/j.15387305.1948.tb01338.x