第二章

熵(Entropy)
與沈農(Shannon)第一定理

2.1 熵(Entropy)之定義

- 在自然界中，所有信息的來源都是隨機產生。然而，如果信息不是隨機的，其輸出的方式可以確切地得知，這時將沒有傳送的必要。
- 例如：這個句子「太陽從東方升起」它沒有任何的意義，因為每個人都知道這個事實。
- 我們根據下面三個句子來感覺其資訊量：

  (A) 約翰是一個男孩而瑪麗是一個女孩。

  (B) 瑪麗下一個小時就會來了。

  (C) 台中市在下雪了。

- 句子(A)跟前一個例子一樣不具有任何的資訊量。
- 句子(B) 含有一些的資訊量，因為瑪麗可能會來也可能不會來。
- 句子 (C)代表極大的資訊量，因為在台中市氣溫很少低於0 度C 以下的。

- 假設信息資料集合以隨機變數S表示，即

$$S = \{s_0, s_1, s_2, \cdots s_{M-1}\}$$

- 其對應之機率為

$$p_i = p(s_i)$$

- 這組信息資料其出現的機率必須符合下述條件

$$\sum_{i=0}^{M-1} p_i = 1$$

- 如果事件s_k出現的機率p_k較小，顯然它出現時將造成較大的驚訝，亦即提供較多的消息。
- 自我資訊量(self-information)定義如下：

$$I(s_k) = \log(1/\ p_k) = -\log\ p_k$$

這項定義有以下四種重要特性：

1. 必然出現的事件其包含之消息量為零。

$$I(s_k) = 0, \quad \text{for } p_k = 1$$

2. 任何事件的出現總會帶來若干消息。

$$I(s_k) \geq 0, \quad \text{for } 0 \leq p_k \leq 1$$

3. 越不容易出現的事件所帶的消息量就越大。

$$I(s_k) \geq I(s_i), \quad \text{for } p_k \leq p_i$$

4. 兩獨立事件的消息量可以相加。

$$I(s_k s_i) = I(s_k) + I(s_i)$$

- 因信號源含各個有不同的符號，而每個符號出現的機率可能均不同。故對於信號源而言，其平均消息量可用期望值計算得到

$$H(S) = E[I(s_k)]$$

$$= \sum_{k=0}^{M-1} p_k \cdot I(s_k)$$

$$= \sum_{k=0}^{M-1} p_k \cdot \log(1/\, p_k)$$

- H(S)稱為信號源的熵(Entropy)，它代表符號出現時所含的平均消息量。

2.2 熵之極值

- 假設集合S 為:

$$S = \{s_0, s_1, \cdots, s_{M-1}\}$$

- 其熵為

$$H(S) = \sum_{k=0}^{M-1} p_k \cdot \log(1/\, p_k)$$

- 當$0 \le k \le M-1$, $p_k = 1/M$ ，$H(\mathrm{S})$為極大值為

$$H(S) = \log M$$

- 當某一個機率為1, 其餘為0，則極小值為

$$H(S) = 0$$

- 因此，熵之極值為

$$0 \leq H(S) \leq \log M$$

2.3 熵與編碼效率

- 在通訊領域裡，如何將消息有效的信號源符號表示出來，這種程序我們稱為信號源編碼。而執行這項工作流程的裝置則稱為信號源編碼器。
- 要能有效的進行信號源編碼，就必須了解信號源的的各項統計資料。譬如較常出現的符號可以用較短的編碼字元表示，而不常出現的符號用較長的編碼字元代表，這種根據出現機率改變編碼長度的編碼方式稱為可變長度編碼。
- 如摩斯碼即為可變長度編碼的一種，由於在英文單字中“E”較“Q”常出現因此“E”以點(●)表示，而“Q”則以“--●-”表示。

- 若信號源包含M個符號，而第k個符號s_k出現的機率為p_k，其對應的二位元編碼長度為l_k。每個符號平均編碼字元長度定義為

$$\overline{L} = \sum_{k=0}^{M-1} p_k l_k$$

參數 $\overline{L}$ 代表每個符號平均位元數。

- 令$L_{\min}$為編碼長度的最小值，則編碼效率可定義如下

$$\eta = L_{\min} / \overline{L}$$

- 我們可以証明編碼長度的最小值$L_{\min}$ 等於$H(S)$ ，即

$$\eta = L_{\min} / \overline{L} = H(S) / \overline{L}$$

- 要如何提高編碼效率，即 $\overline{L} \to H(S)$
  - 減少編碼位元長度: $l_k \geq \log_2(1/p_k)$
  - 延伸前置編碼

減少編碼位元長度

$$H(S) = \sum_{k=0}^{M-1} p_k \cdot \log_2(1/p_k)$$

$$\overline{L} = \sum_{k=0}^{M-1} p_k l_k$$

- 為使 $\overline{L} \to H(S)$ ，則

$$l_k = \lceil \log_2(1/p_k) \rceil$$

即為大於 $\log_2(1/p_k)$ 最小整數。

- 例如：

$$p_k = 0.06$$

$$l_k = \lceil \log_2 1/0.06 \rceil$$

$$= \lceil 4.06 \rceil$$

$$= 5$$

2.4 熵與延伸前置編碼

- 給定一離散無記憶信號源，其熵值為$H(S)$，每個字符平均編碼位元長度 $\overline{L}$ 為:

$$H(S) \le \overline{L} < H(S)+1$$

- 對任意之離散無記憶信號可進行延伸前置編碼嗎？所謂延伸前置編碼即為n個字符為一組加以編碼。

- 設 $\overline{L}_n$ 為延伸前置編碼的n字符平均編碼位元長度，它必需符合下列不等式

$$H(S^n) \le \overline{L}_n < H(S^n) + 1 \qquad (2.1)$$

$$S^n = S \times S \times \cdots \times S$$

- 若集合 $S = \{s_0, s_1, \cdots s_{M-1}\}$ ，集合的元素個數為 M ，即

$$|S| = M$$

- 所以，S^n所包括的元素總數為

$$|S^n| = M \times M \times \cdots \times M = M^n$$

- 針對一個離散無記憶信號源，

$$H(S^n) = nH(S)$$

將此式代入(2.1)，我們不難發現

$$H(S) \le \frac{\overline{L}_n}{n} = \overline{L} < H(S) + 1/n$$

- 當n接近無限大時，

$$\lim_{n->\infty} \frac{\overline{L}_n}{n} = \overline{L} = H(S)$$

- 因此我們可以說只要延伸前置編碼的長度n夠長，其每個字符平均編碼位元長度就可以接近其熵。

例題1：假設我們有一字符信號源$S = \{s_1, s_2\}$，其編碼及出現機率分別為

$$s_1 = 0, \quad p_1 = 2/3$$
$$s_2 = 1, \quad p_2 = 1/3$$

$$H(S) = 2/3 \times \log_2(3/2) + 1/3 \times \log_2(3)$$
$$= 0.9183$$

$$\overline{L} = 1 \times 2/3 + 1 \times 1/3 = 1$$

（每個字符平均編碼位元長度）

S 的第二次延伸S^2 = T =$\{t_1, t_2, t_3, t_4\}$ 而出現機率則分別為(請注意T包括 4個符號）

$$t_1 = s_1 s_1 = 00, \quad p_{t_1} = \frac{2}{3} \times \frac{2}{3} = \frac{4}{9}$$

$$t_2 = s_1 s_2 = 01, \quad p_{t_2} = \frac{2}{3} \times \frac{1}{3} = \frac{2}{9}$$

$$t_3 = s_2 s_1 = 10, \quad p_{t_3} = \frac{1}{3} \times \frac{2}{3} = \frac{2}{9}$$

$$t_4 = s_2 s_2 = 11, \quad p_{t_4} = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$$

$$H(T) = 2 \times H(S) = 1.8366$$

$$\overline{L}_2 = 2(4/9 + 2/9 + 2/9 + 1/9) = 2$$

（2個字符平均編碼位元長度）

$$\overline{L}_2 / 2 = 1$$ （每個字符平均編碼位元長度）

同理，當延伸為10，即 $n=10$，

$$\bar{L}_{10}=10((2/3)^{10}+1/3\times(2/3)^{9}+\cdots(1/3)^{10})=10$$

$$\bar{L}_{10}/10=1$$

當n變大時，並 $\bar{L}_{10}/10$ 沒有於趨近於H(S) ，why？

因為它們是固定編碼，非隨機率編碼，所以不會趨近於H(S)

例題2：假設我們有一字符號源$S=\{s_1, s_2\}$，其編碼及出現機率分別為

$$s_1 = 0, \quad p_1 = 2/3$$
$$s_2 = 10, \quad p_2 = 1/3$$

$$H(S) = 2/3 \times \log_2(3/2) + 1/3 \times \log_2(3)$$
$$= 0.9183$$
$$\overline{L} = 1 \times 2/3 + 2 \times 1/3 = 4/3 = 1.333$$

（每個字符平均編碼位元長度）

S 的第二次延伸$S^2 = T = \{t_1, t_2, t_3, t_4\}$ 而出現機率則分別為(請注意T包括 4個符號）

$$t_1 = s_1 s_1 = 00, \quad p_{t_1} = \frac{2}{3} \times \frac{2}{3} = \frac{4}{9}, \quad l_1 = \lceil \log_2(9/4) \rceil = \lceil 1.17 \rceil = 2$$

$$t_2 = s_1 s_2 = 010, \quad p_{t_2} = \frac{2}{3} \times \frac{1}{3} = \frac{2}{9}, \quad l_2 = \lceil \log_2(9/2) \rceil = \lceil 2.17 \rceil = 3$$

$$t_3 = s_2 s_1 = 100, \quad p_{t_3} = \frac{1}{3} \times \frac{2}{3} = \frac{2}{9}, \quad l_3 = \lceil \log_2(9/2) \rceil = \lceil 2.17 \rceil = 3$$

$$t_4 = s_2 s_2 = 1100, \quad p_{t_4} = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}, \quad l_4 = \lceil \log_2(9) \rceil = \lceil 3.17 \rceil = 4$$

$$H(T) = 2 \times H(S) = 1.8366$$

$$\overline{L}_2 = 2(4/9) + 3(2/9) + 3(2/9) + 4(1/9) = 2.666$$

（2個字符平均編碼位元長度）

$$\overline{L}_2 / 2 = 1.333$$

（每個字符平均編碼位元長度）

S 的第三次延伸$S^3 = T = \{t_1, t_2, t_3, \ldots, t_8\}$ 而出現機率則分別為(請注意T包括 8個符號）

$$t_1 = s_1 s_1 s_1, \quad p_{t_1} = \frac{8}{27}, \quad l_1 = \lceil \log_2(27/8) \rceil = \lceil 1.75 \rceil = 2$$

$$t_2 = s_1 s_1 s_2, \quad p_{t_2} = \frac{4}{27}, \quad l_2 = \lceil \log_2(27/4) \rceil = \lceil 2.75 \rceil = 3$$

$$t_3 = s_1 s_2 s_1, \quad p_{t_3} = \frac{4}{27}, \quad l_3 = 3$$

$$t_4 = s_1 s_2 s_2, \quad p_{t_4} = \frac{2}{27}, \quad l_4 = \lceil \log_2(27/2) \rceil = \lceil 3.75 \rceil = 4$$

$$t_5 = s_2 s_1 s_1, \quad p_{t_5} = \frac{4}{27}, \quad l_5 = 3$$

$$t_6 = s_2 s_1 s_2, \quad p_{t_6} = \frac{2}{27}, \quad l_6 = 4$$

$$t_7 = s_2 s_2 s_1, \quad p_{t_7} = \frac{2}{27}, \quad l_7 = 4$$

$$t_8 = s_2 s_2 s_2, \quad p_{t_8} = \frac{1}{27}, \quad l_8 = \lceil \log_2(27/1) \rceil = \lceil 4.75 \rceil = 5$$

$$\overline{L}_3 = 2(8/27) + 3(4 \times 3/27) + 3(2 \times 4/27)$$

$$+ (5/27) = 81/27 \quad \text{（3個字符平均編碼位元長度）}$$

$$\overline{L}_3 / 3 = 1.0 \quad \text{（每個字符平均編碼位元長度）}$$

- S 的第五次延伸S^5 = T ={t_1, t_2, t_3, …, t_{32}}

$$\overline{L}_5 / 5 = 0.93333$$

作業1

假設我們有字符信號源$S = \{s_1, s_2, s_3\}$，出現機率分別為

$$p(s_1) = p_1 = 1/2$$
$$p(s_2) = p_2 = 1/4$$
$$p(s_3) = p_3 = 1/4$$

(1)若採固定方式編碼，其H(S)與 $\overline{L}$ 為何？

(2)若採機率方式編碼，其H(S) 、$\overline{L}$ 與$\overline{L}_2$ (第二次延伸中，2個字符平均編碼位元長度)為何？