



# 熵之估計與無失真資料壓縮

## 第 3 章



# 英文之結構與熵

- ▶ 平均編碼長度可以無限制地逼近符號源之熵值。然而，基本問題是如何決定熵值。
- ▶ 以英文為例，我們限制符號源只包含**26**個英文字母及一個空白，以 “^” 表示，共**27**個符號。
- ▶ 對於這樣一個符號源，最簡單的描述模式是 *DMS* 並且假設所有符號的出現機率都一樣，即二十七分之一。就這個模式而言，英文的熵值為

$$H(S) = \log_2(27) = 4.75 \text{ 位元/符號},$$

# 英文之結構與熵

- ▶ 這個模式完全沒有考慮英文裡面所含的結構，其結果是所求得之熵（不確定性）高，無法降低存在於這個語言中的任何冗贅。
- ▶ 比較好的模式是考慮進去每個符號實際上的出現機率（如表3.1所列）。這些機率值的取得是藉由匯整一般的典型英文文章，然後再做統計獲得。

符號	機率	符號	機率
空白	0.1859	N	0.0574
A	0.0642	O	0.0632
B	0.0127	P	0.0152
C	0.0218	Q	0.0008
D	0.0317	R	0.0484
E	0.1031	S	0.0514
F	0.0208	T	0.0796
G	0.0152	U	0.0228
H	0.0467	V	0.0083
I	0.0575	W	0.0175
J	0.0008	X	0.0013
K	0.0049	Y	0.0164
L	0.0321	Z	0.0005
M	0.0198		

表3.1 英文字母與其出現的機率。

- ▶ 利用這些符號的出現機率所求得之熵為

$$H_2(S) = \sum_{s^2} p(s_{i_1}, s_{i_2}) \log_2 \left( \frac{1}{p(s_{i_2} | s_{i_1})} \right) = 3.32$$

- ▶ 這個模式所產生的字已比前一個模式更實際地表現出母音與子音所佔的比例。它主要的缺點是它並沒有把英文字母間存在的相關性考慮進去。

# 英文之結構與熵

- ▶ 將連續符號間之相關性考慮進去的最簡單模式是第一階馬可夫過程，需要二十七個機率表，每一個機率表對應於前一個出現的英文字母或空白，表示馬可夫過程的狀態（前一個出現符號）。

- ▶ 由這個第一階馬可夫過程模式所估算得的熵為

$$H_2(S) = \sum_{s^2} p(s_{i_1}, s_{i_2}) \log_2 \left( \frac{1}{p(s_{i_2} | s_{i_1})} \right) = 3.32 \text{ 位元符號}$$

- ▶ 第二階馬可夫過程的熵為

$$H_2(S) = \sum_{s^3} p(s_{i_1}, s_{i_2}, s_i) \log_2 \left( \frac{1}{p(s_i | s_{i_1}, s_{i_2})} \right) = 3.1 \text{ 位元符號}$$

# 英文之可預測性與熵

- ▶ 先對一個受實驗者展示一段他不熟悉的文章，共 $N-1$ 個符號，然後請他猜下一個字母是什麼，直到猜對為止。
- ▶ 基於受實驗者對於英文的了解，受實驗者在猜測過程中心裡會有一組條件機率（已知前 $N-1$ 個符號，下一個字母是某某某的機率）並且根據這組條件機率從大到小選擇他的猜測。

# 英文之可預測性與熵

- ▶ 實驗重覆 $n$ 次。令  $q_i^N$  表示在看過前面 $N-1$ 個字母後，受實驗者仍然需要猜 $i$ 次才能猜出正確字母的次數。  
*Shannon*證明這段文章的熵滿足

$$\sum_{i=1}^{27} i \left( \frac{q_i^N}{n} - \frac{q_{i+1}^N}{n} \right) \log_2 i \leq H_2(S) \leq - \sum_{i=1}^{27} \frac{q_i^N}{n} \log_2 \left( \frac{q_i^N}{n} \right)$$

- ▶ 由這個式子，*Shannon*得到

$$0.6 \text{ bits/s} \leq H(S) \leq 1.3 \text{ bits/s}$$

- ▶ 人類對英文的知識構成英文符號源的冗贅，語音等其他訊號也存在著其他種類的冗贅。



# 自然影像 (Natural images)

## 之可預期性與熵

- ▶ *Kersten*使用一個和*Shannon*的猜字遊戲類似的過程來估算自然影像之熵值與冗贅。在他的研究裡，使用了不同複雜度的影像，每張影像的取樣值皆為128×128個像素，而且每一個像素都量化為4個位元(16個灰階度)。
- ▶ 影像中一定比例的像素先被去掉，然後由受實驗者試著去將這些像素的原灰階度填回去。受實驗者可以要求電腦將不同的灰階度顯示在影像下方，直到他選擇到自己滿意的灰階度為止。



# 自然影像 (Natural images) 之可預期性與熵



- ▶ 這個實驗重複做一百次，然後統計所使用的猜測次數。一張影像的冗贅量則被定義成 $1 - \text{熵} / \text{實際上每個像素使用之位元數}$ （在這個實驗是4）
- ▶ 根據 *Shannon* 所提出之熵的上、下限，*Kersten* 結論出，這八張影像的冗贅量由最低的46%（最複雜的影像）到最高的74%（人臉影像）。