Chương 3. Nén dữ liệu (data compression) và entropy

ntnhut@hcmus.edu.vn

Ví dụ "3.1" về nén dữ liệu

- Chuỗi nhị phân, số ký tự '0' nhiều gấp 9 lần '1':
  - P('0') = 0.9, P('1') = 0.1.
- Chia chuỗi thành từng khối để mã hoá.
- **<u>Trường hợp một khối = 2 ký tự:</u>**

Symbol	00	01	10	11
Probability	0.81	0.09	0.09	0.01

- Mã Huffman:

Symbol	00	01	10	11
Code	0	10	110	111

- Độ dài mã TB:

$$L_{\min} = 0.81 + 2(0.09) + 3(0.1) = 1.29.$$

- **Trường hợp một byte = 2 ký tự:**
  - Cần TB khoảng 1.29 bits để mã hoá 2 ký tự, hay 1.29/2 = **0.645 bits/ký tự.**
- **Trường hợp 1 byte = 3 ký tự:**

Symbol	000	100	010	001	110	101	110	111
Probab.	0.729	0.081	0.081	0.081	0.009	0.009	0.009	0.001
Code	0	100	101	110	11100	11101	11110	11111

$$L_{\min} = 0.729 + 3(0.243) + 5(0.028) = 1.598,$$

$$\frac{1.598}{3} = 0.533 \text{ bits/symbol.}$$

Câu hỏi: ***chúng ta có thể nén đến mức nào? Có thể nén 0.5 bits/ký tự hay ít hơn nữa được không?***

Ý tưởng về entropy

- 1948, Claude E. Shannon.
- Nén dựa vào *tính cú pháp (syntactic)* của văn bản, không phải *tính ngữ nghĩa (semantic)*.
- **Entropy của một nguồn thông tin S, H(S):**
  - H(S) = lượng thông tin cần thiết để xác định một ký tự nguồn.
- Tính chất của H(S):
  - $H(S) = H(p_1, p_2, \ldots, p_n)$.
  - H(S) *dương, liên tục, đối xứng*.

Định nghĩa entropy

Định nghĩa: *Entropy* của nguồn thông tin S có các phân phối xác suất $p_1, \ldots, p_n$ là:

$$H(S) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^{n} p_i \log_2 p_i \text{ (bits)}.$$

Ví dụ: Tung đồng xu, xác suất 2 mặt là như nhau. Entropy của nguồn thông tin này là

$$H(S) = \frac{1}{2}\log_2 2 + \frac{1}{2}\log_2 2 = 1 \text{ bit}.$$

Entropy nhị phân đối xứng

Chuỗi nhị phân, số ký tự '0' nhiều gấp 9 lần '1'

$$H(S) = -0.1 \log_2 0.1 - 0.9 \log_2 0.9 \approx 0.469 \text{ bits.}$$

Nguồn thông tin với 2 ký tự và xác suất (p,1 – p):

$$H(p, 1-p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p} \quad \text{(bits)}.$$

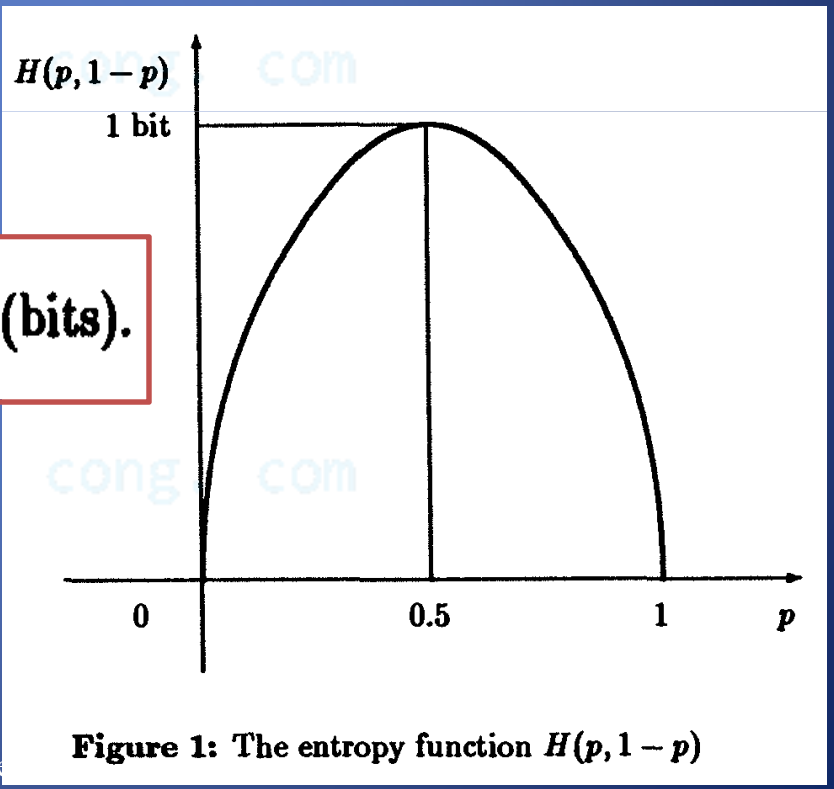

Figure 1: The entropy function $H(p, 1-p)$

ntnhut@hcmus.

Entropy cực tiểu và cực đại

Định lý: (1) *Entropy cực tiểu = 0 khi S chỉ có 1 ký tự.* (2) *Entropy đạt cực đại = $\log_2 n$ bits khi xác suất của các ký tự bằng nhau = 1/n.*

Chứng minh:

$$H(S) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^{n} p_i \log_2 p_i$$

(1) $p_i, \log_2(1/p_i) \geq 0$. '=' khi $p_i = 0$ hay $1/p_i = 1$.

(2) (bài tập).

Mở rộng của nguồn thông tin

Định nghĩa: Cho S là nguồn thông tin $\{a_1, \ldots, a_n\}$. ***Mở rộng bậc k của S***, ký hiệu S^k, là nguồn thông tin có các ký tự dạng '$a_{i1}a_{i2}\ldots a_{ik}$' với các xác suất $P(a_{i1}a_{i2}\ldots a_{ik}) = P(a_{i1})P(a_{i2})\ldots P(a_{ik})$. Trong đó, $i_1, i_2, \ldots, i_k \in \{1, 2, \ldots, n\}$.

Ví dụ 3.1 (tiếp): S: {0,1}; P(0) = 0.9; P(1) = 0.1.

S^2 và S^3:

Symbol	00	01	10	11
Probability	0.81	0.09	0.09	0.01

Symbol	000	100	010	001	110	101	110	111
Probab.	0.729	0.081	0.081	0.081	0.009	0.009	0.009	0.001
Code	0	100	101	110	11100	11101	11110	11111

Mối liên hệ giữa Entropy và Độ dài mã trung bình

$$H(S) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}$$

$$L = \sum_{i=1}^{n} d_i P(a_i).$$

Định lý: *mọi mã tức thời nhị phân của nguồn S đều có độ dài mã trung bình không nhỏ hơn entropy của S: $L \geq H(S)$.*

Chứng minh: (bài tập)

Định lý mã không nhiễu của Shannon

- Trong Ví dụ 3.1:
  - $H(S) = 0.469$ (bits)
  - $L_{min}(S^2)/2 = 0.645$ bits/symbol
  - $L_{min}(S^3)/3 = 0.533$ bits/symbol
  - … $\geq H(S)$.

Định lý: *Với mọi nguồn S, độ dài mã Huffman nhị phân* $L_{min}(S)$ *thoả :* $H(S) \leq L_{min}(S) \leq H(S) + 1$.

Với mở rộng S^k *của nguồn S ta có*:

$$\frac{L_{min}(S^k)}{k} \longrightarrow H(S) \quad for\ k \longrightarrow \infty.$$

Chứng minh: (bài tập)

Tóm tắt

1. Với mỗi nguồn S, entropy H(S) là lượng thông tin trung bình (tính bằng bit) của một ký tự. H(S) là số bit trung bình tối ưu để nén S.
2. Mã Huffman của S^k là cách nén tối ưu nhất.
3. Với bảng mã có r>2 ký tự mã:

$$H_r(S) = \sum p_i \log_r p_i,$$

$$H_r(S) \leq L_{\min}(S) \leq H_r(S) + 1,$$

$$\frac{L_{\min}(S^k)}{k} \longrightarrow H_r(S) \qquad \text{for } k \longrightarrow \infty.$$

Homework

- Đọc lại:
  - Chương 3 [1].
  - Chương 2 [2].
- Đọc trước:
  - Chương 4 [1]

Bài tập 1

- Một văn bản được viết bằng bảng ký tự {A, B, C, D}, trong đó ký tự A xuất hiện nhiều gấp 7 lần mỗi ký tự còn lại. Tìm một mã nhị phân sử dụng trung bình không quá 1.4 bits/ký tự.

- *Gợi ý: dùng mở rộng* S^k.

Bài tập 2

- Tính entropy của nguồn thông tin sau

Symbol	1	2	3	4	5	6
Probability	0.1	0.1	0.45	0.05	0.2	0.1

- **Bài tập Thực hành:**

Tính entropy của một nguồn thông tin cho trước.

Bài tập 3

- Một kênh truyền các ký tự 0 và 1 đồng xác suất. Tính xác suất nhận được chuỗi ‘01101’. Tính entropy của một văn bản dùng các chuỗi 5 ký tự.

Bài tập 4

- Một nguồn thông tin gồm 128 ký tự đồng xác suất. Tính độ dài của chuỗi có entropy bằng 42 bits.

Bài tập 5

- *Hiệu quả E(S)* của một nguồn thông tin S được định nghĩa là tỷ số giữa entropy H(S) và độ dài trung bình của mã Huffman nhị phân $L_{min}(S)$.

a) CMR: $0 \leq E(S) \leq 1$.

b) Nhận xét các cực trị của E(S).

c) Tính E(S) của các nguồn sau

Symbol	A	B	C	D	E	F	G	H
Prob. (1st source)	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
Prob. (2nd source)	0.1	0.2	0.1	0.3	0.05	0.1	0.05	0.1
Prob. (3rd source)	0.15	0.15	0.15	0.15	0.1	0.1	0.1	0.1

Bài tập 6

- Một văn bản nhị phân dài chứa số ký tự ‘0’ nhiều gấp đôi ‘1’. Tìm mã nén:

a) Sử dụng tối đa 0.94 bits/ký tự

b) Sử dụng tối đa 0.9 bits/ký tự.