

CHƯƠNG 5: LÝ THUYẾT THÔNG TIN TIỆM CẬN

TS. TRỊNH VĂN CHIẾN (SOICT-HUST)

VÍ DỤ

- Lý thuyết: Sử dụng lượng tiền X_0 để đầu tư, tỉ lệ thu được hàng ngày là r_i . Do đó sau t ngày, tổng số tiền thu được là

$$X_t = X_0 r_1 r_2 \dots r_t$$

- Nếu r_i tuân theo (i.i.d. independent and identically distributed) với:

$$r_i = \begin{cases} 4, & \text{with probability } 1/2 \\ 0, & \text{with probability } 1/2 \end{cases}$$

- Theo thống kê $E\{r_i\} = 2$, do đó:

$$E\{X_t\} = X_0 E\{r_1 r_2 \dots r_t\} = X_0 (E\{r_1\})^t = X_0 2^t$$

- Thực tế: Với $X_0 = 1$, lợi nhuận thu được như sau: 1 4 16 0 0...

Điều gì sai?

LUẬT SỐ LỚN (YẾU)

- Luật số lớn yếu (weak law of large number): Cho một tập các biến ngẫu nhiên (X_1, \dots, X_i) có tính chất là độc lập và đồng nhất (i.i.d), tính chất sau hội tụ về mặt xác suất

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \rightarrow E\{X\}$$

- Hội tụ về mặt xác suất có nghĩa là với $\varepsilon > 0$

$$P\left\{\left|\overline{X}_n - E\{X\}\right| > \varepsilon\right\} \rightarrow 0$$

- Điều này có nghĩa là

$$P\left\{\left|\overline{X}_n - E\{X\}\right| \leq \varepsilon\right\} \rightarrow 1, n \rightarrow \infty$$

MỘT SỐ QUY TẮC HỘI TỤ

- Hội tụ về mặt trung bình bậc 2: Nếu $n \rightarrow \infty$, ta có

$$E\{(X_n - X)^2\} \rightarrow 0$$

- Hội tụ với xác suất là 1 (hội tụ gần như chắc chắn-almost sure): Nếu $n \rightarrow \infty$, ta có

$$P\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1$$

- Hội tụ về mặt phân bố: Nếu $n \rightarrow \infty$, ta có

$$\lim_{n \rightarrow \infty} F_n \rightarrow F$$

- Với F_n và F là hai phân bố xác suất của 2 biến ngẫu nhiên X_n và X

TÍCH CỦA CÁC BIẾN NGẪU NHIÊN

- Xem xét biến ngẫu nhiên sau:

$$\sqrt[n]{\prod_{i=1}^n X_i}$$

- Sử dụng mối quan hệ giữa trung bình hình học và trung bình số học ta có:

$$\sqrt[n]{\prod_{i=1}^n X_i} \leq \frac{1}{n} \sum_{i=1}^n X_i$$

- Nhiều ứng dụng trong thực tế sử dụng tích của các biến ngẫu nhiên:

- Diện tích, thể tích,...
- Đầu tư.....
- Phân bố đồng thời của các biến ngẫu nhiên i.i.d:

$$p(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i)$$

LUẬT SỐ LỚN CHO TÍCH CÁC BIẾN NGẪU NHIÊN

- Viết lại biến ngẫu nhiên

$$X_i = e^{\log X_i}$$

- Do đó ta có

$$\sqrt[n]{\prod_{i=1}^n X_i} \leq e^{\frac{1}{n} \sum_{i=1}^n \log X_i}$$

- Theo quy tắc của luật số lớn ta có

$$\sqrt[n]{\prod_{i=1}^n X_i} \leq e^{\frac{1}{n} \sum_{i=1}^n \log X_i} = e^{E\{\log X\}} \leq e^{\log E\{X\}} = E\{X\}$$

ỨNG DỤNG

- ▣ Đầu tư có rủi ro cao

$$E\{\log r_i\} = \frac{1}{2}\log 4 + \frac{1}{2}\log 0 = -\infty$$

$$E\{X_t\} \rightarrow X_0 e^{E\{\log r_i\}t} = 0, t \rightarrow \infty$$

- ▣ Xác định đường bao của kì vọng

$$X = \begin{cases} a, & \text{with probability } 1/2 \\ b, & \text{with probability } 1/2 \end{cases}$$

$$E\left\{\frac{X_1 + X_2}{2}\right\} \leq E\left\{\frac{X_1 + X_2}{2}\right\} = E\{X\}$$

ĐẶC TÍNH XẤP XỈ TIỆM CẬN (1)

- Theo luật số lớn:

$$\frac{1}{n} \sum_{i=1}^n X_i \rightarrow E\{X\}$$

- Đặc tính tiệm cận:

$$-\frac{1}{n} \log p(X_1, \dots, X_n) \rightarrow H(X)$$

$$p(X_1, \dots, X_n) \approx 2^{-nH(X)}$$

- Đặc tính tiệm cận được dùng để:

- Chứng minh mã hóa không lỗi
- Chứng minh dung lượng kênh truyền thông
-

ĐẶC TÍNH XẤP XỈ TIỆM CẬN (2)

Định lý: Nếu các biến ngẫu nhiên X_1, \dots, X_n là i.i.d có phân bố xác suất $p(x)$, ta có:

$$-\frac{1}{n} \log p(X_1, \dots, X_n) \rightarrow H(X)$$

hội tụ theo xác suất

Chứng minh:

✓ Vì các biến ngẫu nhiên là i.i.d nên ta có:

$$\begin{aligned} -\frac{1}{n} \log p(X_1, \dots, X_n) &= -\frac{1}{n} \sum_{i=1}^n \log p(X_i) \\ &\rightarrow -E\{\log p(x)\} = H(X) \end{aligned}$$

TẬP TIÊU BIỂU (TYPICAL SET)

- Một tập tiêu biểu $A_\varepsilon^{(n)}$ chứa tất cả các mẫu giá trị (x_1, \dots, x_n) với đặc tính

$$2^{-n(H(X)+\varepsilon)} \leq p(x_1, \dots, x_n) \leq 2^{-n(H(X)-\varepsilon)}$$

- Đặc tính: Không phải tất cả các tập mẫu đều có giá trị như nhau
- Ví dụ: Tung đồng xu $X \in \{0,1\}$, $p(1)=0.8$, 06 lần và quan sát

$$p(1,0,1,1,0,1) = 0.8^4 0.2^2 = 0.0164$$

$$p(0,0,0,0,0,0) = 0.2^6 = 0.000064$$

Trong ví dụ trên, nếu chuỗi mẫu $(x_1, \dots, x_n) \in A_\varepsilon^{(n)}$ ta có

$$H(X) - \varepsilon \leq -\frac{1}{n} \log p(x_1, \dots, x_n) \leq H(X) + \varepsilon$$

Điều này nghĩa là một chuỗi nhị phân có nằm trong tập tiêu biểu nếu tần suất hiện giá trị “1” xấp xỉ k/n với k là số lần đếm số 1.

VÍ DỤ (1)

- Ví dụ 2: Tung đồng xu với xác suất $p(0)=0.1$ và $p(1)=0.9$:
 - a) Sau n lần tung, chuỗi mẫu (x_1, \dots, x_n) có xác suất hiện cao nhất là gì?
 - b) Chuỗi mẫu trên có nằm trong tập tiêu biểu không? Vì sao?

VÍ DỤ (2)

Ví dụ 2: Tung đồng xu với xác suất $p(0)=0.1$ và $p(1)=0.9$:

a) Sau n lần tung, chuỗi mẫu (x_1, \dots, x_n) có xác suất hiện cao nhất là gì?

b) Chuỗi mẫu trên có nằm trong tập tiêu biểu không? Vì sao?

Đáp án:

a) Vì $p(1) \gg p(0)$ nên chuỗi mẫu có xác suất xuất hiện cao nhất sau n lần tung $(x_1, \dots, x_n) = (1, \dots, 1)$

b) Tính:

$$H(X) = -0.1 \log_2 0.1 - 0.9 \log_2 0.9 = 0.452$$

$$-\frac{1}{n} \log_2 p(X = (1, \dots, 1)) = -\frac{1}{n} \log_2 (0.9^n) = 0.152$$

➔ Số lần tung tiến tới vô cùng thì 2 giá trị này không thể bằng nhau ➔ Chuỗi mẫu này không nằm trong tập tiêu biểu

VÍ DỤ (3)

Ví dụ 3: Xét đồng xu với xác suất $p(0) = 1 - p$ và $p(1) = p$. Tung đồng xu n lần, số lần xuất hiện giá trị 1 là m lần. Phân tích tập tiêu biểu.

Đáp án:

$$\text{Ta có } -\frac{1}{n} \log_2 p(X = (x_1, \dots, x_n)) = -\frac{1}{n} \log_2 p^m (1-p)^{n-m} = -\frac{m}{n} \log_2 p - \left(\frac{n-m}{n} \right) \log_2 (1-p)$$

Số lần xuất hiện 1 trung bình là $m = np$, do đó ta có

$$-\frac{1}{n} \log_2 p(X = (x_1, \dots, x_n)) = -p \log_2 p - (1-p) \log_2 (1-p) = H(X)$$

Quan sát:

- ▣ Nếu $p = 0.5$, thì mọi chuỗi mẫu đều nằm trong tập tiêu biểu
- ▣ Nếu $n = 10$, $p(0) = 0.1$, $p(1) = 0.9$, tập tiêu biểu chứa chuỗi mẫu có 1 bit 0

HỆ QUẢ (1)

- Các hệ quả của xấp xỉ tiệm cận

$$1) H(X) - \varepsilon \leq -\frac{1}{n} \log p(x_1, x_2, \dots, x_n) \leq H(X) + \varepsilon$$

$$2) P(A_\varepsilon^{(n)}) \geq 1 - \varepsilon$$

$$3) |A_\varepsilon^{(n)}| \leq 2^{n(H(X) + \varepsilon)}$$

$$4) |A_\varepsilon^{(n)}| \geq (1 - \varepsilon) 2^{n(H(X) - \varepsilon)}$$

HỆ QUẢ (2)

- Nếu số lần lấy mẫu đủ lớn: $P(A_\varepsilon^{(n)}) \geq 1 - \varepsilon$
- Chứng minh: Từ lý thuyết về xấp xỉ tiệm cận

$$-\frac{1}{n} \log p(X_1, \dots, X_n) \rightarrow H(X)$$

hội tụ theo xác suất, do đó

$$p \left\{ \underbrace{\left| -\frac{1}{n} \log p(X_1, \dots, X_n) - H(X) \right|}_{\in A_\varepsilon^{(n)}} \leq \varepsilon \right\} \geq 1 - \varepsilon$$

Nhận xét:

- Các chuỗi mẫu trong tập tiêu biểu là tiêu biểu nhất
- Các chuỗi mẫu này có xác suất xấp xỉ gần như nhau

HỆ QUẢ(3)

$$(1-\varepsilon)2^{n(H(X)-\varepsilon)} \leq |A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$$

Chứng minh:

Ta có:

$$1 = \sum_{(x_1, \dots, x_n)} p(x_1, \dots, x_n) \geq \sum_{(x_1, \dots, x_n) \in A_\varepsilon^{(n)}} p(x_1, \dots, x_n) \geq \sum_{(x_1, \dots, x_n) \in A_\varepsilon^{(n)}} p(x_1, \dots, x_n) 2^{-n(H(X)+\varepsilon)} = |A_\varepsilon^{(n)}| 2^{-n(H(X)+\varepsilon)}$$

Mặt khác $P(A_\varepsilon^{(n)}) \geq 1 - \varepsilon$, do đó

$$1 - \varepsilon \leq \sum_{(x_1, \dots, x_n) \in A_\varepsilon^{(n)}} p(x_1, \dots, x_n) = |A_\varepsilon^{(n)}| 2^{-n(H(X)-\varepsilon)}$$

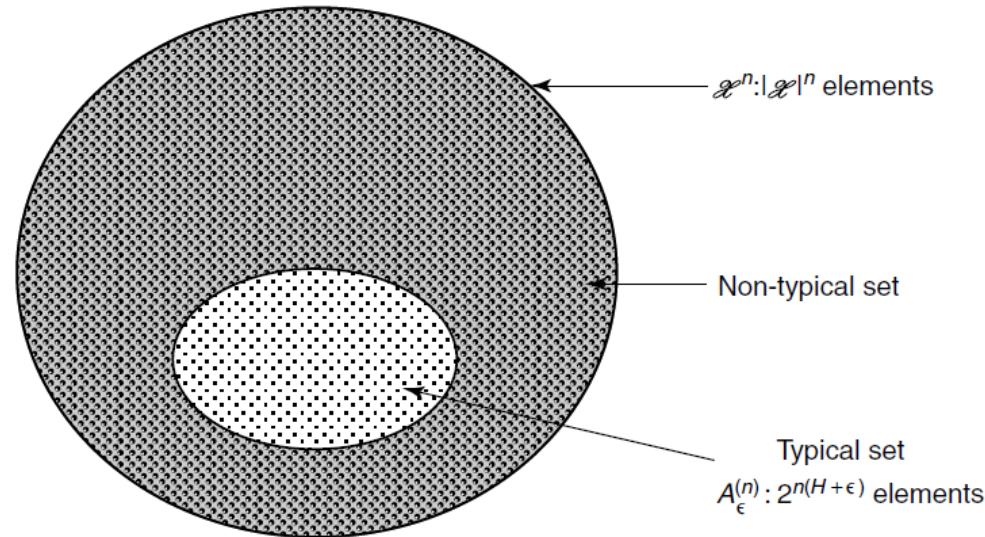
Nhận xét:

- Cỡ của tập tiêu biểu phụ thuộc vào lượng tin riêng
- Ví dụ tung đồng xu: Nếu $p=1/2$, $H(X)=1$, $2^{nH(X)} = 2^n$ tất cả chuỗi mẫu là chuỗi tiêu biểu

BIỂU ĐỒ TẬP TIÊU BIỂU

Chúng ta có thể chia không gian mẫu ra làm 2 tập hợp:

- ❑ Tập tiêu biểu (typical set): Xác suất xuất hiện cao, lượng tin riêng trong tập này gần bằng lượng tin riêng toàn bộ không gian mẫu → Chúng ta nên tập trung vào tập hợp này
- ❑ Tập không tiêu biểu (non-typical set): Xác suất xuất hiện nhỏ, có thể bỏ qua



BÀI TẬP

Bài tập: Xét tung đồng xu n lần (X_1, \dots, X_n) với $p(0)=0.4$ và $p(1)=0.6$

a) Tính lượng tin riêng $H(X)$

b) Với $n=25$, $\varepsilon=0.1$, tìm chuỗi mẫu trong tập tiêu biểu.

k	$\binom{n}{k}$	$\binom{n}{k} p^k (1-p)^{n-k}$	$-\frac{1}{n} \log p(x^n)$
0	1	0.000000	1.321928
1	25	0.000000	1.298530
2	300	0.000000	1.275131
3	2300	0.000001	1.251733
4	12650	0.000007	1.228334
5	53130	0.000054	1.204936
6	177100	0.000227	1.181537
7	480700	0.001205	1.158139
8	1081575	0.003121	1.134740
9	2042975	0.013169	1.111342
10	3268760	0.021222	1.087943
11	4457400	0.077801	1.064545
12	5200300	0.075967	1.041146
13	5200300	0.267718	1.017748
14	4457400	0.146507	0.994349
15	3268760	0.575383	0.970951
16	2042975	0.151086	0.947552
17	1081575	0.846448	0.924154
18	480700	0.079986	0.900755
19	177100	0.970638	0.877357
20	53130	0.019891	0.853958
21	12650	0.997633	0.830560
22	2300	0.001937	0.807161
23	300	0.999950	0.783763
24	25	0.000047	0.760364
25	1	0.000003	0.736966

BÀI TẬP

Đáp án:

a) $H(X) = -0.4 \log_2 0.4 - 0.6 \log_2 0.6 = 0.97095$ bits

b) Tập tiêu biểu nằm trong phạm vi $(H(X) - \varepsilon, H(X) + \varepsilon) = (0.87095, 1.07095)$ \rightarrow Chuỗi nằm trong cột từ 11 đến 19