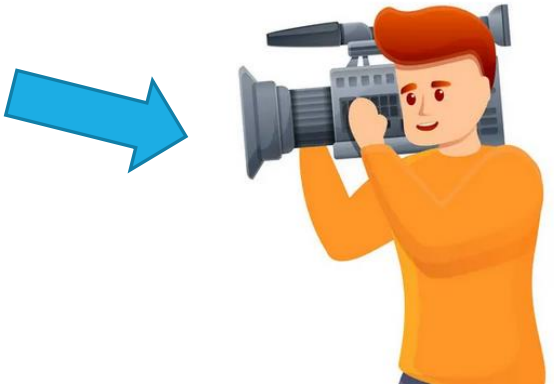
CHƯƠNG 4: XỬ LÝ DỮ LIỆU

TS. TRỊNH VĂN CHIẾN (SOICT-HUST)

HỆ THỐNG XỬ LÝ DỮ LIỆU

Khung cảnh



Xử lý dữ liệu

Dữ liệu lưu

CHUỖI MARKOV (1)

- Một chuỗi Markov được định nghĩa:

$$X \to Y \to Z$$

Nếu chúng ta có thể biểu diễn:

$$p(x, y, z) = p(z \mid y)\, p(y \mid x)\, p(x)$$

- Cách hiểu khác:

$$X \to Y \to f(Y)$$

CHUỖI MARKOV (2)

- Ví dụ: Tung một con xúc xắc với xác suất xuất hiện mặt thứ m là θ_m
  - Giả sử chúng ta tung n lần được kết quả $\{X_1, X_2, \dots, X_n\}$.
  - Tính giá trị trung bình của n lần tung

$$\bar{X} = \frac{1}{n}\left(X_1 + \dots + X_n\right)$$

  - Chuỗi Markov được định nghĩa như sau

$$\left\{\theta_1, \theta_2, \dots, \theta_6\right\} \rightarrow \left\{X_1, \dots, X_n\right\} \rightarrow \bar{X}_n$$

HỆ QUẢ (1)

- Chuỗi Markov $X \to Y \to Z \Leftrightarrow X$ và Z độc lập nếu Y được cho trước

Chứng minh: Dựa vào xác suất có điều kiện ta có

$$p(x,z \mid y) = \frac{p(x,y,z)}{p(y)} = \frac{p(z \mid y)\,p(y \mid x)\,p(x)}{p(y)} = \frac{p(z \mid y)\,p(x,y)}{p(y)} = p(x \mid y)\,p(z \mid y)$$

- Hệ quả này có thể mở rộng cho chuỗi Markov gồm n chiều
- Hệ quả này có thể dùng để kiểm tra một chuỗi là chuỗi Markov

HỆ QUẢ (2)

- Nếu chuỗi $X \to Y \to Z$ là chuỗi Markov thì chuỗi $Z \to Y \to X$ cũng là chuỗi Markov

Chứng minh:

$$p(x,y,z) = p(x)p(y \mid x)p(z \mid y) = p(x)p(y \mid x)\frac{p(y,z)}{p(y)}$$

$$= p(x,y)\frac{p(y \mid z)p(z)}{p(y)} = p(x \mid y)p(y)\frac{p(y \mid z)p(z)}{p(y)} = p(x \mid y)p(y \mid z)p(z)$$

BẤT ĐẲNG THỨC CHO XỬ LÝ DỮ LIỆU (1)

Định lý: Xem xét một chuỗi Markov $X \to Y \to Z$, ta có

$$I(X;Y) \geq I(X;Z); \quad I(Y;Z) \geq I(X;Z)$$

Dấu bằng xảy ra khi $I(X;Y \mid Z) = 0$

- Nếu chúng ta xử lý thông tin, chúng ta sẽ bị mất thông tin
- Trong một vài trường hợp, đẳng thức về thông tin tương hỗ vẫn đạt được khi loại bỏ một số thông tin

BẤT ĐẲNG THỨC CHO XỬ LÝ DỮ LIỆU (2)

- Chứng minh: Sử dụng định nghĩa thông tin tương hỗ và thông tin tương hỗ có điều kiện

$$I(X;Y,Z) = I(X;Z) + I(X;Y \mid Z) = I(X;Y) + \underbrace{I(X;Z \mid Y)}_{=0}$$

với $I(X;Z \mid Y) = 0$ bởi vì X, Z là độc lập nếu Y được cho trước. Do đó ta có

$$I(X;Y) \geq I(X;Z)$$

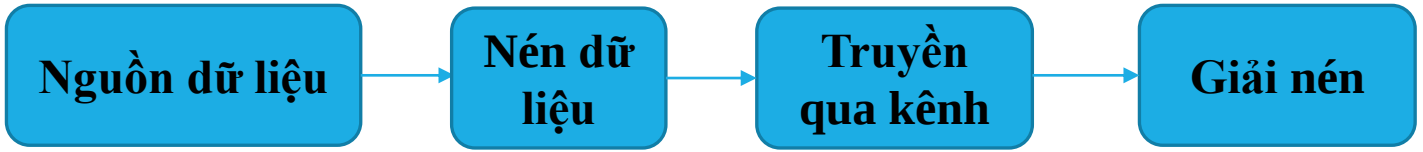
Dấu bằng xảy ra nếu $I(X;Y \mid Z) = 0$, nghĩa là $X \to Z \to Y$ là một chuỗi Markov.

Tương tự chúng ta có thể chứng minh

$$I(Y;Z) \geq I(X;Z)$$

MÔ HÌNH NÉN DỮ LIỆU

- Mô hình nén và giải nén giữ liệu



- Các bước thực hiện:
  - ✓ Nén dữ liệu W từ nguồn sử dụng $X = (X_1, ..., X_n)$
  - ✓ Truyền qua kênh truyền (ngẫu nhiên): Y
  - ✓ Giải nén thu được $\hat{W}$
  - ➔ Theo định lý về xử lý dữ liệu

$$I(W;\hat{W}) \leq I(X;Y)$$

HỆ QUẢ CỦA BẤT ĐẲNG THỨC XỬ LÝ DỮ LIỆU

- Nếu hàm g được cho trước trong chuỗi Markov: $X \to Y \to g(Y)$

$$I(X;Y) \geq I(X;g(Y))$$

- Xét chuỗi Markov $X \to Y \to Z$, ta có

$$I(X;Y \mid Z) \leq I(X;Y)$$

Chứng minh: Sử dụng thông tin tương hỗ và thông tin tương hỗ có điều kiện

$$I(X;Y,Z) = I(X;Z) + I(X;Y \mid Z) = I(X;Y) + \underbrace{I(X;Z \mid Y)}_{=0}$$

- Quan sát:
  - ✓ Sự phụ thuộc của X và Y sẽ giảm (hoặc không đổi) bằng việc quan sát Z
  - ✓ Nếu một tiến trình không tuân theo chuỗi Markov, có thể xảy ra $I(X;Y \mid Z) \geq I(X;Y)$
    - ✓ Ví dụ: Tung 2 đồng xu ứng với các sự kiện X, Y. Đặt $Z = X+Y$, ta có:

$$I(X;Y \mid Z) = 1/2; I(X;Y) = 0$$

THỐNG KÊ ĐỦ (SUFFICIENT STATISTICS) (1)

- Bất đẳng thức xử lý dữ liệu cung cấp thông tin về thống kê đủ (sufficient statistics)
- Cho một họ phân bố xác suất $\{f_\theta(x)\}$ được định danh bởi θ
- X là một mẫu dữ liệu (sampled) từ $f_\theta(x)$
- $T(X)$ là hàm thống kê mẫu dữ liệu X , ta có

$$\theta \to X \to T(X)$$

- Bất đẳng thức xử lý dữ liệu:

$$I\left(\theta; T(X)\right) \le I(\theta; X)$$

THỐNG KÊ ĐỦ (SUFFICIENT STATISTICS) (2)

- Một thống kê là đầy đủ cho θ nếu *T(X)* chứa tất cả thông tin của θ trong *X*

$$I(\theta; X) = I(\theta; T(X))$$

- Ví dụ:
  - Nếu $X = \{X_1, ..., X_n\} \sim \text{Ber}(\theta)$, $T(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$ là thống kê đầy đủ cho θ
  - Nếu $X = \{X_1, ..., X_n\} \sim \text{Uniform}(\theta, \theta+1)$, $T(X) = \{\min_i X_i, \max_i X_i\}$ là thống kê đầy đủ cho θ
- Lưu ý:
  - ✓ Một thống kê đầy đủ có thể là một chiều hoặc nhiều chiều
  - ✓ Một thống kê đầy đủ có thể không là duy nhất. Ví dụ: Bản thân *X* cũng là một thống kê đầy đủ của chính nó
  - ✓ Thống kê đầy đủ cực tiểu: là một hàm của toàn bộ các thống kê đầy đủ khác ➔ Thông tin nén tối đa của θ trong tập mẫu dữ liệu

BẤT ĐẲNG THỨC FANO (1)

- Bất đẳng thức Fano (1942) liên kết xác suất lỗi P_e với lượng tin riêng
- Xác suất lỗi có quan hệ mật thiết với lượng tin riêng bởi vì: Phía phát truyền X, phía thu nhận Y và giải mã $\tilde{X}$ có thể xuất hiện lỗi

$$P_e = P\left(\hat{X} \neq X\right)$$

- Chuỗi Markov: $X \to Y \to \hat{X}$
- Quan sát: Giải mã X từ Y với xác suất bằng 0 nếu $H(X \mid Y) = 0$
- Bất đẳng thức Fano mở rộng quan sát trên: Giải mã X từ Y với xác suất lỗi nhỏ nếu

$H(X \mid Y)$ nhỏ

BẤT ĐẲNG THỨC FANO (2)

Định lý: Với bất kỳ phiên bản giải mã $\hat{X}$ thỏa mãn chuỗi Markov $X \to Y \to \hat{X}$

$$H(P_e) + P_e \log|\tilde{X}| \geq H(\hat{X}|X) \geq H(X|Y)$$

- Hệ quả:

$$P_e \geq \frac{H(Y|X)-1}{\log|\tilde{X}|} = \frac{H(X)-I(X;Y)-1}{\log|\tilde{X}|}$$

- Nếu phương pháp giải mã *g(Y)* xem xét trực tiếp giá trị trong $\tilde{X}$, bất đẳng thức Fano sẽ viết lại

$$H(P_e) + P_e \log(|\tilde{X}|-1) \geq H(X|Y)$$

BÀI TẬP (1)

Bài 1: Cho hai biến ngẫu nhiên (X,Y) có xác suất xảy ra đồng thời p(x,y) như sau

X \ Y	a	b	c
1	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$
2	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$
3	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{6}$

Giả sử $\hat{X}(Y)$ là một phương pháp giải mã X từ Y với xác suất lỗi $P_e = P(\hat{X}(Y) \neq X)$

a) Định nghĩa phương pháp giải mã $\hat{X}(Y)$ và định nghĩa xác suất lỗi Pe cho phương pháp giải mã đề xuất

b) Xác định bao của Pe từ bất đẳng thức Fano và đưa ra kết luận

BÀI TẬP (2)

a) Phương pháp giải mã đề xuất:

$$\hat{X}(Y) = \begin{cases} 1, & y = a \\ 2, & y = b \\ 3, & y = c \end{cases}$$

Xác suất lỗi Pe được định nghĩa như sau

$$P_e = P(1,b) + P(1,c) + P(2,a) + P(2,c) + P(3,a) + P(3,b) = 1/2$$

b) Sử dụng bất đẳng thức Fano: $P_e \geq \frac{H(Y \mid X) - 1}{\log\left(\mid \tilde{X} \mid -1\right)}$

$$H(X \mid Y) = 1.5 \; bits$$

➔ Bất đẳng thức Fano $P_e \geq \frac{1.5 - 1}{\log(3-1)} = 0.5$

Phương pháp giải mã khá tốt

BÀI TẬP (3)

Bài 2: Xem xét ba biến ngẫu nhiên X,Y, Z có quan hệ phụ thuộc lẫn nhau. Chứng minh

$a)\quad H(X,Y \mid Z) \geq H(X \mid Z)$

$b)\quad I(X,Y;Z) \geq I(X;Z)$

$c)\quad H(X,Y,Z) - H(X,Y) \leq H(X,Z) - H(X)$

$d)\quad I(X;Z \mid Y) \geq I(Z;Y \mid X) - I(Z;Y) + I(X;Z)$

BÀI TẬP (4)

a) $$H(X,Y|Z) = H(X|Z) + H(Y|X,Z) \geq H(X|Z)$$

b) $$I(X,Y;Z) = H(X,Y) - H(X,Y|Z) = H(X) + H(Y|X) - H(X|Z) - H(Y|X,Z)$$
$$= I(X;Z) + H(Y|X) - H(Y|X,Z) \geq I(X;Z)$$

c) Sử dụng quy tắc chuỗi

$$H(X,Y,Z) - H(X,Y) = H(Z \mid X,Y)$$
$$H(X,Z) - H(X) = H(Z \mid X)$$

d) Ta có:

$$I(X;Z|Y) - I(Y;Z|X) = H(Z|Y) - H(Z|X,Y) - H(Z|X) + H(Z|X,Y)$$
$$= H(Z|Y) - H(Z|X)$$

$$I(X;Z) - I(Y;Z) = H(Z) - H(Z|X) - H(Z) + H(Z|Y) = H(Z|Y) - H(Z|X)$$

TỔNG KẾT

- Xử lý dữ liệu: Có thể mất mát thông tin hoặc không
- Thống kê đủ (sufficient statistics) bảo toàn thông tin
- Khi ước lượng thông tin từ tín hiệu quan sát, lỗi mất mát thông tin có thể được bao bởi bất đẳng thức Fano