

# Bài tập về nhà số 5

## Nền tảng toán học của các mô hình tạo sinh – PIMA

### Chủ đề: Mô hình năng lượng

Người giải: Võ Hoàng Nhật Khang

1. Xét một máy Boltzmann gồm  $k$  nút  $x := (x_1, x_2, \dots, x_k) \in \{0, 1\}^k$  với  $b_1, b_2, \dots, b_k$  lần lượt là các thiên vị của nút  $x_1, x_2, \dots, x_k$  và  $w_{ij}$  là trọng số của cặp nút  $x_i$  và  $x_j$ . Hàm mật độ xác suất của mô hình là

$$p(x; W, b) := \frac{1}{Z} e^{-E(x; W, b)},$$

trong đó

- $W = (w_{ij})$  là ma trận trọng số.
- $b = (b_1, b_2, \dots, b_k)$  là vector thiên vị.
- $E$  là hàm năng lượng và  $E(x; W, b)$  là năng lượng ứng với cấu hình  $x$  và bộ tham số  $W, b$ .
- $Z = \sum_{x \in \{0,1\}^k} e^{-E(x; W, b)}$  là hằng số chuẩn hóa.

Ta lấy  $D$  mẫu dữ liệu  $\mathcal{D} := \{d_1, d_2, \dots, d_D\}$  độc lập và có cùng phân phối thực nghiệm

$$p_{\text{data}}(x) = \frac{1}{D} \sum_{i=1}^D \delta(x, d_i),$$

với

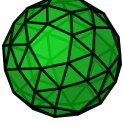
$$\delta(\alpha, \beta) = \begin{cases} 1, & \alpha = \beta \\ 0, & \alpha \neq \beta \end{cases}.$$

Vì sao việc tối thiểu hóa phân kỳ Kullback-Leibler từ  $p(x; \theta)$  đến  $p_{\text{data}}$  và tối đa hóa hàm log-hợp lý  $\ln \mathcal{L}(\theta) = \ln p(\mathcal{D}; \theta)$  của  $\mathcal{D}$  đối với bộ tham số  $\theta := (W, b)$  là tương đương? Tức là chứng minh

$$\operatorname{argmax}_{\theta} \ln \mathcal{L}(\theta) = \operatorname{argmin}_{\theta} D_{\text{KL}}(p_{\text{data}} \| p(x; \theta)),$$

trong trường hợp điểm cực trị  $\theta$  tồn tại.

**Lời giải:**



Để chứng minh sự tương đương, ta cần chỉ ra rằng tối đa hóa  $\ln \mathcal{L}(\theta)$  tương ứng với tối thiểu hóa  $D_{\text{KL}}(p_{\text{data}} \| p(x; \theta))$ .

Hàm log-hợp lý của tập dữ liệu  $\mathcal{D}$  được định nghĩa là

$$\ln \mathcal{L}(\theta) = \ln p(\mathcal{D}; \theta) = \sum_{i=1}^D \ln p(d_i; \theta), \quad (1)$$

trong đó  $p(d_i; \theta) = \frac{1}{Z} e^{-E(d_i; W, b)}$ .

Phân kỳ Kullback-Leibler từ  $p_{\text{data}}$  đến  $p(x; \theta)$  là

$$D_{\text{KL}}(p_{\text{data}} \| p(x; \theta)) = \sum_x p_{\text{data}}(x) \ln \left( \frac{p_{\text{data}}(x)}{p(x; \theta)} \right). \quad (2)$$

Thay  $p_{\text{data}}(x) = \frac{1}{D} \sum_{i=1}^D \delta(x, d_i)$  vào, ta được

$$D_{\text{KL}}(p_{\text{data}} \| p(x; \theta)) = \sum_x \left( \frac{1}{D} \sum_{i=1}^D \delta(x, d_i) \right) \ln \left( \frac{\frac{1}{D} \sum_{j=1}^D \delta(x, d_j)}{p(x; \theta)} \right). \quad (3)$$

Vì  $\delta(x, d_i) = 1$  chỉ khi  $x = d_i$ , nên

$$D_{\text{KL}}(p_{\text{data}} \| p(x; \theta)) = \frac{1}{D} \sum_{i=1}^D \ln \left( \frac{\frac{1}{D}}{p(d_i; \theta)} \right) = \frac{1}{D} \sum_{i=1}^D \left( \ln \frac{1}{D} - \ln p(d_i; \theta) \right) \quad (4)$$

$$= -\ln D - \frac{1}{D} \sum_{i=1}^D \ln p(d_i; \theta) \quad (5)$$

Do đó

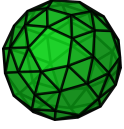
$$D_{\text{KL}}(p_{\text{data}} \| p(x; \theta)) = -\ln D - \frac{1}{D} \ln \mathcal{L}(\theta). \quad (6)$$

Vì  $-\ln D$  là hằng số không phụ thuộc  $\theta$ , việc tối thiểu hóa  $D_{\text{KL}}(p_{\text{data}} \| p(x; \theta))$  tương đương với tối thiểu hóa  $-\frac{1}{D} \ln \mathcal{L}(\theta)$ , tức là tối đa hóa  $\ln \mathcal{L}(\theta)$ . Vậy:

$$\operatorname{argmax}_{\theta} \ln \mathcal{L}(\theta) = \operatorname{argmin}_{\theta} D_{\text{KL}}(p_{\text{data}} \| p(x; \theta)). \quad (7)$$

2. Xét một máy Boltzmann hạn chế gồm  $m$  nút hiện  $v = (v_1, v_2, \dots, v_m) \in \{0, 1\}^m$  và  $n$  nút ẩn  $h = (h_1, h_2, \dots, h_n) \in \{0, 1\}^n$ , trong đó:

- $\theta = (W, b, c)$  là bộ tham số, với  $W = (w_{ij})$  là ma trận trọng số,  $w_{ij}$  là trọng số giữa hai nút  $v_i$  và  $h_j$ ;  $b = (b_1, b_2, \dots, b_m)$  là vector thiên vị của vector biến hiện  $v$ ;  $c = (c_1, c_2, \dots, c_n)$  là vector thiên vị của vector biến ẩn  $h$ .
- $E(v, h; \theta) = -\sum_{i=1}^m \sum_{j=1}^n w_{ij} v_i h_j - \sum_{i=1}^m b_i v_i - \sum_{j=1}^n c_j h_j$  là hàm năng lượng.
- $Z = \sum_{v, h} e^{-E(v, h; \theta)}$  là hằng số chuẩn hóa.



**a)** Việc giới hạn máy Boltzmann thành một đồ thị lưỡng phân, trong đó các nút ở cùng một lớp không còn tương tác làm cho thao tác tính toán các phân phối biên được thuận tiện hơn. Cụ thể, chứng minh:

$$p(v; \theta) := \sum_h p(v, h; \theta) = \frac{1}{Z} \prod_{i=1}^m e^{b_i v_i} \prod_{j=1}^n \left( 1 + e^{c_j + \sum_{i=1}^m w_{ij} v_i} \right).$$

**Lời giải:**

Ta bắt đầu từ định nghĩa

$$p(v; \theta) = \sum_h p(v, h; \theta) = \sum_h \frac{1}{Z} e^{-E(v, h; \theta)}. \quad (8)$$

Với  $E(v, h; \theta) = -\sum_{i=1}^m \sum_{j=1}^n w_{ij} v_i h_j - \sum_{i=1}^m b_i v_i - \sum_{j=1}^n c_j h_j$ , ta có

$$e^{-E(v, h; \theta)} = \exp \left( \sum_{i=1}^m \sum_{j=1}^n w_{ij} v_i h_j + \sum_{i=1}^m b_i v_i + \sum_{j=1}^n c_j h_j \right) \quad (9)$$

Tổng theo  $h$  là

$$\sum_h e^{-E(v, h; \theta)} = \sum_{h_1=0}^1 \cdots \sum_{h_n=0}^1 \exp \left( \sum_{i=1}^m b_i v_i + \sum_{j=1}^n h_j \left( c_j + \sum_{i=1}^m w_{ij} v_i \right) \right). \quad (10)$$

Tách biệt các phần

$$\sum_h e^{-E(v, h; \theta)} = \exp \left( \sum_{i=1}^m b_i v_i \right) \prod_{j=1}^n \sum_{h_j=0}^1 \exp \left( h_j \left( c_j + \sum_{i=1}^m w_{ij} v_i \right) \right). \quad (11)$$

Với  $h_j \in \{0, 1\}$ , ta tính

$$\sum_{h_j=0}^1 \exp \left( h_j \left( c_j + \sum_{i=1}^m w_{ij} v_i \right) \right) = 1 + \exp \left( c_j + \sum_{i=1}^m w_{ij} v_i \right). \quad (12)$$

Do đó

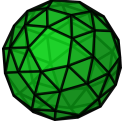
$$p(v; \theta) = \frac{1}{Z} \exp \left( \sum_{i=1}^m b_i v_i \right) \prod_{j=1}^n \left( 1 + e^{c_j + \sum_{i=1}^m w_{ij} v_i} \right) = \frac{1}{Z} \prod_{i=1}^m e^{b_i v_i} \prod_{j=1}^n \left( 1 + e^{c_j + \sum_{i=1}^m w_{ij} v_i} \right) \quad (13)$$

■

**b)** Xem xét thuật toán Gradient ascent trên hàm hợp lý được cập nhật như sau:

$$\theta^{(t+1)} = \theta^{(t)} + \eta \frac{\partial \ln \mathcal{L}(\theta^{(t)} | \tilde{v})}{\partial \theta^{(t)}},$$

trong đó



- $\theta^{(t+1)}$  và  $\theta^{(t)}$  lần lượt là bộ tham số được cập nhật ở bước  $t + 1$  và  $t$ .
- $\eta > 0$  là tốc độ học (hằng số).
- $\frac{\partial \ln \mathcal{L}(\theta^{(t)}|\tilde{v})}{\partial \theta^{(t)}}$  là gradient của hàm log-hợp lý  $\ln \mathcal{L}$  theo  $\theta^{(t)}$  cho mẫu  $\tilde{v}$ .

Chúng minh rằng:

$$\frac{\partial \ln \mathcal{L}(\theta|\tilde{v})}{\partial \theta} = \mathbb{E}_{v, h \sim p(v, h)} \left[ \frac{\partial E(v, h)}{\partial \theta} \right] - \mathbb{E}_{h \sim p(h|\tilde{v})} \left[ \frac{\partial E(\tilde{v}, h)}{\partial \theta} \right].$$

**Lời giải:**

Hàm hợp lý cho mẫu  $\tilde{v}$  là

$$\mathcal{L}(\theta|\tilde{v}) = p(\tilde{v}; \theta) = \frac{\sum_h e^{-E(\tilde{v}, h; \theta)}}{Z}. \quad (14)$$

Log-hợp lý

$$\ln \mathcal{L}(\theta|\tilde{v}) = \ln \left( \sum_h e^{-E(\tilde{v}, h; \theta)} \right) - \ln Z. \quad (15)$$

Gradient theo  $\theta$

$$\frac{\partial \ln \mathcal{L}(\theta|\tilde{v})}{\partial \theta} = \frac{\partial}{\partial \theta} \ln \left( \sum_h e^{-E(\tilde{v}, h; \theta)} \right) - \frac{\partial \ln Z}{\partial \theta}. \quad (16)$$

Tính từng phần

$$\frac{\partial}{\partial \theta} \ln \left( \sum_h e^{-E(\tilde{v}, h; \theta)} \right) = -\mathbb{E}_{h \sim p(h|\tilde{v})} \left[ \frac{\partial E(\tilde{v}, h)}{\partial \theta} \right], \quad (17)$$

và

$$\frac{\partial \ln Z}{\partial \theta} = -\mathbb{E}_{v, h \sim p(v, h)} \left[ \frac{\partial E(v, h)}{\partial \theta} \right]. \quad (18)$$

Do đó

$$\frac{\partial \ln \mathcal{L}(\theta|\tilde{v})}{\partial \theta} = \mathbb{E}_{v, h \sim p(v, h)} \left[ \frac{\partial E(v, h)}{\partial \theta} \right] - \mathbb{E}_{h \sim p(h|\tilde{v})} \left[ \frac{\partial E(\tilde{v}, h)}{\partial \theta} \right] \quad (19)$$

■

**c)** Thực tế, đại lượng  $\mathbb{E}_{h \sim p(h|\tilde{v})} \left[ \frac{\partial E(\tilde{v}, h)}{\partial \theta} \right]$  có thể được tính toán dễ dàng (vì sao?). Ngược lại, trung bình  $\mathbb{E}_{v, h \sim p(v, h)} \left[ \frac{\partial E(v, h)}{\partial \theta} \right]$  lại mất nhiều chi phí tính toán vì phân phối đồng thời  $p(v, h)$  lấy trên tất cả cấu hình của  $v$  và  $h$ . Hãy đề xuất một phương pháp xấp xỉ  $\mathbb{E}_{v, h \sim p(v, h)} \left[ \frac{\partial E(v, h)}{\partial \theta} \right]$ .

**Lời giải:**

Đại lượng  $\mathbb{E}_{h \sim p(h|\tilde{v})} \left[ \frac{\partial E(\tilde{v}, h)}{\partial \theta} \right]$  dễ tính vì khi  $\tilde{v}$  cố định, các  $h_j$  độc lập, cho phép tính kỳ vọng riêng lẻ theo  $p(h_j|\tilde{v})$ .

Ngược lại,  $\mathbb{E}_{v, h \sim p(v, h)} \left[ \frac{\partial E(v, h)}{\partial \theta} \right]$  đòi hỏi tổng trên tất cả cấu hình  $v, h$ , rất tốn kém.

Phương pháp xấp xỉ: Sử dụng *phương pháp tương phản* (Contrastive Divergence, CD- $k$ ). Bắt đầu từ  $\tilde{v}$ , chạy Markov chain  $k$  bước (thường  $k = 1$ ) để lấy mẫu  $v^{(k)}, h^{(k)}$  và dùng  $\frac{\partial E(v^{(k)}, h^{(k)})}{\partial \theta}$  để xấp xỉ kỳ vọng. ■