# BÀI 7 - LATENT DIRICHLET ALLOCATION

Hạn nộp bài: hết ngày 08/06/2024.

## 1 Bài tập lý thuyết

**Bài 1.** Trong phần phụ lục A.3 của bài báo Latent Dirichlet Allocation, tác giả tìm chặn dưới cho log hàm hợp lý của một văn bản $\mathbf{w}$ bất kì bằng bất đẳng thức

$$\log p(\mathbf{w} \mid \alpha, \beta) \geq \mathbb{E}_q[\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)] - \mathbb{E}_q[\log q(\theta, \mathbf{z})]. \tag{12}$$

Ký hiệu vế phải của bất đẳng thức trên là $\mathcal{L}(\gamma, \phi; \alpha, \beta)$, đây sẽ là một chặn dưới cho phân phối biến thiên $q(\theta, \mathbf{z} \mid \gamma, \phi)$.

a) Chứng minh rằng

$$\log p(\mathbf{w} \mid \alpha, \beta) = \mathcal{L}(\gamma, \phi; \alpha, \beta) + D_{KL}(q(\theta, \mathbf{z} \mid \gamma, \phi) \parallel p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)). \tag{13}$$

và

$$\begin{aligned}\mathcal{L}(\gamma, \phi; \alpha, \beta) = \mathbb{E}_q[\log p(\theta \mid \alpha)] + \mathbb{E}_q[\log p(\mathbf{z} \mid \theta)] + \mathbb{E}_q[\log p(\mathbf{w} \mid \mathbf{z}, \beta)] \\ - \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log q(\mathbf{z})].\end{aligned} \tag{14}$$

b) Khai triển vế phải của (14) để nhận được công thức (15) trong bài báo:

$$\begin{aligned}\mathcal{L}(\gamma, \phi; \alpha, \beta) = &\log \Gamma\left(\textstyle\sum_{j=1}^{k} \alpha_j\right) - \sum_{i=1}^{k} \log \Gamma(\alpha_i) + \sum_{i=1}^{k} (\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k} \gamma_j\right)\right) \\ &+ \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k} \gamma_j\right)\right) \\ &+ \sum_{n=1}^{N} \sum_{i=1}^{k} \sum_{j=1}^{V} \phi_{ni} w_n^j \log \beta_{ij} \\ &- \log \Gamma\left(\textstyle\sum_{j=1}^{k} \gamma_j\right) + \sum_{i=1}^{k} \log \Gamma(\gamma_i) - \sum_{i=1}^{k} (\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k} \gamma_j\right)\right) \\ &- \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log \phi_{ni}.\end{aligned} \tag{15}$$

## 2 Bài tập thực hành

**Bài 2.** Dữ liệu trong Research Articles chứa khoảng 30000 phần tóm tắt của các bài báo nghiên cứu.

a) Hãy đọc dữ liệu và xem danh sách các chủ đề trong dữ liệu. Trực quan hoá số lượng bài đăng về mỗi chủ đề.

b) Hãy tiền xử lý dữ liệu: ngắt từ, loại bỏ stopwords, tạo bigrams/trigrams.

c) Từ dữ liệu đã xử lý, hãy tạo dictionary và bag-of-word.

d) Xây dựng mô hình LDA để xử lý bài toán topic modeling cho dữ liệu này với số chủ đề là 20. In ra các từ khoá cho mỗi chủ đề. Tính perplexity và điểm coherence.

e) Dùng pyLDAvis để trực quan hoá các chủ đề. Nhận xét.

f) Ở bước tiền xử lý, hãy bổ sung thêm bước lemmatization. Xây dựng lại mô hình LDA và so sánh kết quả với mô hình khi không có lemmatization.

g) Điều chỉnh số lượng chủ đề, xây dựng lại mô hình LDA và tính điểm coherence. Chọn số lượng chủ đề tốt nhất cho dữ liệu này.