TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

BÀI TẬP MÔN
XỬ LÝ NGÔN NGỮ TỰ NHIÊN

BÀI TẬP QUÁ TRÌNH 01

Giảng viên hướng dẫn: Nguyễn Đức Vũ

Họ và tên	MSSV	Mã lớp
Trần Đình Khánh Đăng	22520195	CS221.P12

TP. Hồ Chí Minh, ngày 5 tháng 10 năm 2024

Bài tập 1

Chứng minh rằng tổng xác suất của tất cả các chuỗi có thể từ tập từ vựng $\mathcal{V}$ trong hai trường hợp sau:

- Không có từ kết thúc </s>: Mô hình ngôn ngữ có thể sinh chuỗi dài vô hạn.

$$\sum_{n=1}^{\infty}\sum_{x_{1:n}} P(x_{1:n}) = \infty \tag{1}$$

- Có từ kết thúc </s> Mô hình ngôn ngữ phải dừng lại việc sinh chuỗi bằng từ kết thúc câu </s>..

$$\sum_{n=1}^{\infty}\sum_{x_{1:n}} P(x_{1:n} < /s >) = 1 \tag{2}$$

Bài tập 1a

Ta có:

$$\sum_{n=1}^{1}\sum_{x_{1:n}} P(x_{1:n}) = 1,$$

$$\sum_{n=2}^{2}\sum_{x_{1:n}} P(x_{1:n}) = 1,$$

$$\Leftrightarrow \sum_{n=1}^{2}\sum_{x_{1:n}} P(x_{1:n}) = 1 + 1 = 2,$$

$$\text{và} \quad \sum_{n=1}^{3}\sum_{x_{1:n}} P(x_{1:n}) = 1 + 1 + 1 = 3,$$

$$\dots$$

Giả sử:

$$\sum_{n=1}^{k}\sum_{x_{1:n}} P(x_{1:n}) = k,$$

Với một số $k \geq 1$.
Ta cần chứng minh:

$$\sum_{n=1}^{k+1}\sum_{x_{1:n}} P(x_{1:n}) = k + 1.$$

$$\sum_{n=k+1}^{k+1}\sum_{x_{1:n}} P(x_{1:n}) = 1,$$

$$\text{Ta có} \sum_{n=1}^{k+1}\sum_{x_{1:n}} P(x_{1:n}) = \sum_{n=1}^{k}\sum_{x_{1:n}} P(x_{1:n}) + \sum_{n=k+1}^{k+1}\sum_{x_{1:n}} P(x_{1:n})$$

$$\Rightarrow \sum_{n=1}^{k+1}\sum_{x_{1:n}} P(x_{1:n}) = k + 1 \text{ (ĐPCM)}$$

Theo quy nạp, ta có:

$$\sum_{n=1}^{\infty}\sum_{x_{1:n}} P(x_{1:n}) = \infty.$$

Bài tập 1b

Ta có

$$\sum_{n=1}^{\infty}\sum_{x_{1:n}} P(x_{1:n}, < /s >) = 1$$

$$\Leftrightarrow \sum_{n=1}^{\infty} P(< /s >)\sum_{x_{1:n}} P(x_{1:n}) = 1$$

hoặc $\Leftrightarrow \sum_{n=1}^{\infty} P(< /s > | x_{1:n})\sum_{x_{1:n}} P(x_{1:n}) = 1$ (Trong trường hợp $P(< /s >)$ và $P(x_{1:n})$ không độc lập)

Xét

$$\begin{aligned}
&\sum_{x_{1:n}} P(x_{1:n}) \\
&= \sum_{x_{1:n}} \prod_{i} P(x_i) \\
&= \sum_{x_1}\sum_{x_2}\cdots\sum_{x_n}\prod_{i} P(x_i) \\
&= \sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} P(x_1)P(x_2)\dots P(x_n) \\
&= 1 \quad (3)
\end{aligned}$$

dễ thấy,

$$\sum_{n=1}^{\infty} P(< /s >) = 1 \quad (4)$$

$$\sum_{n=1}^{\infty} P(< /s > | x_{1:n}) = 1$$ (Trong trường hợp $P(< /s >)$ và $P(x_{1:n})$ không độc lập)

Từ (3) và (4) $\Rightarrow \sum_{n=1}^{\infty}\sum_{x_{1:n}} P(x_{1:n} < /s >) = 1$

Bài tập 2

Cho tập dữ liệu gồm nhiều văn bản thuộc các lớp $C = \{c_1, c_2, \dots, c_k\}$ và mỗi văn bản chứa các từ từ tập từ vựng $\mathcal{V}$. Hãy sử dụng phương pháp **MLE** để tính:

- **Xác xuất tiên nghiệm của lớp c_j:**

$$\hat{P}(c_j) = \frac{\text{count}(c_j)}{N_{doc}}$$

Trong đó:

[label=•]count(c_j): số văn bản thuộc lớp c_j. N_{doc}: tổng số văn bản.

- **Xác suất có điều kiện của từ w_i trong lớp c_j:**

$$\hat{P}(w_i|c_j) = \frac{\text{count}(w_i, c_j)}{\sum_{w\in\mathcal{V}}\text{count}(w, c_j)}$$

Trong đó:

[label=•]count(w_i, c_j): số lần từ w_i xuất hiện trong lớp c_j. $\sum_{w\in\mathcal{V}}\text{count}(w, c_j)$: tổng số lần xuất hiện của tất cả các từ trong lớp c_j.

Bài tập 2a

- Đặt $\theta_j = P(c_j)$ là xác xuất của lớp c_j mà chúng ta muốn ước lượng.

Likelihood Function

Hàm khả năng cho số lượng tài liệu trong k lớp theo phân phối đa thức:

$$L(\theta_1, \theta_2, \ldots, \theta_k) = \prod_{j=1}^{k} \theta_j^{\text{count}(c_j)}$$

Hàm Log-Likelihood

Lấy logarit của hàm khả năng, ta được hàm Log-likelihood:

$$\log L(\theta_1, \theta_2, \ldots, \theta_k) = \sum_{j=1}^{k} \text{count}(c_j) \log \theta_j$$

Ràng buộc xác suất

Vì θ_j đại diện cho xác suất, ta có ràng buộc:

$$\sum_{j=1}^{k} \theta_j = 1$$

Sử dụng Lagrange Multipliers

Chúng ta sử dụng một hệ số Lagrange λ để đưa ràng buộc vào bài toán cực đại hóa. Hàm Lagrangian là:

$$\mathcal{L} = \sum_{j=1}^{k} \text{count}(c_j) \log \theta_j + \lambda \left(1 - \sum_{j=1}^{k} \theta_j\right)$$

Cực đại hóa Log-likelihood

Để cực đại hóa $\mathcal{L}$, ta lấy đạo hàm riêng theo từng θ_j và đặt bằng 0:

$$\frac{\partial \mathcal{L}}{\partial \theta_j} = \frac{\text{count}(c_j)}{\theta_j} - \lambda = 0$$

Giải tìm θ_j:

$$\theta_j = \frac{\text{count}(c_j)}{\lambda}$$

Giải cho λ

Sử dụng ràng buộc $\sum_{j=1}^{k} \theta_j = 1$, ta có thể giải cho λ:

$$\sum_{j=1}^{k} \frac{\text{count}(c_j)}{\lambda} = 1$$

$$\frac{1}{\lambda} \sum_{j=1}^{k} \text{count}(c_j) = 1$$

Vì $\sum_{j=1}^{k} \text{count}(c_j) = N_{doc}$, ta có:

$$\frac{N_{doc}}{\lambda} = 1 \quad \Rightarrow \quad \lambda = N_{doc}$$

Ước lượng cuối cùng cho θ_j

Thay $\lambda = N_{doc}$ vào biểu thức của θ_j:

$$\theta_j = \frac{\text{count}(c_j)}{N_{doc}}$$

Kết luận

Vì vậy, Maximum Likelihood Estimate (MLE) cho xác suất của lớp c_j là:

$$\hat{P}(c_j) = \frac{\text{count}(c_j)}{N_{doc}}$$

Bài tập 2b

Chúng ta muốn ước lượng xác xuất có điều kiện $P(w_i \mid c_j)$, xác suất của từ w_i khi biết lớp c_j.

Likelihood Function

Likelihood Functionlà:

$$L(\theta_{w_i,c_j}) = \prod_{w \in V} \theta_{w,c_j}^{\text{count}(w,c_j)}$$

Hàm Log-Likelihood

Lấy log của hàm khả năng, ta được hàm Log-likelihood:

$$\log L(\theta_{w_i,c_j}) = \sum_{w \in V} \text{count}(w, c_j) \log \theta_{w,c_j}$$

Cực đại hóa Log-Likelihood

Để cực đại hóa hàm Log-likelihood, ta áp dụng nguyên tắc MLE dưới ràng buộc:

$$\sum_{w \in V} \theta_{w,c_j} = 1$$

Sử dụng Lagrange multipliers, ta định nghĩa hàm Lagrangian:

$$\mathcal{L} = \sum_{w \in V} \text{count}(w, c_j) \log \theta_{w,c_j} + \lambda \left(1 - \sum_{w \in V} \theta_{w,c_j}\right)$$

Lấy đạo hàm riêng theo θ_{w_i,c_j} và đặt bằng 0:

$$\frac{\partial \mathcal{L}}{\partial \theta_{w_i,c_j}} = \frac{\text{count}(w_i, c_j)}{\theta_{w_i,c_j}} - \lambda = 0$$

Giải ra θ_{w_i,c_j}:

$$\theta_{w_i,c_j} = \frac{\text{count}(w_i, c_j)}{\lambda}$$

Giải tìm λ

Sử dụng ràng buộc $\sum_{w \in V} \theta_{w,c_j} = 1$:

$$\sum_{w \in V} \frac{\text{count}(w, c_j)}{\lambda} = 1$$

$$\frac{1}{\lambda} \sum_{w \in V} \text{count}(w, c_j) = 1$$

Do đó, $\lambda = \sum_{w \in V} \text{count}(w, c_j)$.

Ước lượng cuối cùng cho θ_j

Thay $\lambda = \sum_{w \in V} \text{count}(w, c_j)$ vào biểu thức của θ_j:

$$\theta_j = \frac{\text{count}(c_j)}{\sum_{w \in V} \text{count}(w, c_j)}$$

Kết luận

Vì vậy, Maximum Likelihood Estimate (MLE) cho xác xuất có điều kiện của lớp c_j là:

$$\hat{P}(c_j) = \frac{\text{count}(c_j)}{\sum_{w \in V} \text{count}(w, c_j)}$$