Bài tập 2.

Cho tập dữ liệu gồm nhiều văn bản thuộc các lớp C = $\{c_1, c_2, c_3, ...c_k\}$ và mỗi văn bản chứa các từ từ tập từ vựng V. Hãy sử dụng phương pháp MLE để tính:
2.1 Xác suát tiên nghiệm của lớp c_j

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

Trong đó:

- Count(c_j) : số văn bản thuộc lớp c_j

- N_{doc} : tổng số văn bản

Giải:

Gọi x_i là văn bản thuộc lớp c_j , với i $\in [1, m]$; m là số lượng văn bản thuộc lớp j

$\Rightarrow \sum_{i=1}^{m} x_i^j = Count(c_j)$

Xét các biến x_i là độc lập

Theo phân phối Categorical (trường hợp tổng quát của phân phối Bermoulli):

$$P(x_i|\hat{P}(c_j) = \prod_{i=1}^{m} \hat{P}(c_j)^{x_i}$$

Đánh giá $\hat{P}(c_j)$ dựa trên Maximum log-likelihood:

$\mathcal{L}\hat{P}(c_j) = \arg\max[P(x_1, x_2, x_3..., x_m|\hat{P}(c_j)] = \arg\max \prod_{i=1}^{m} P(x_i|\hat{P}(c_j)) = \arg\max \prod_{i=1}^{m}\prod_{j=1}^{k} \hat{P}(c_j)^{x_i}$

$= \arg\max \prod_{j=1}^{k} \hat{P}(c_j)^{\sum_{i=1}^{m} x_i} = \arg\max \prod_{j=1}^{k} \hat{P}(c_j)^{Count(c_j)} = \arg\max \sum_{j=1}^{k} Count(c_j)\log(\hat{P}(c_j))$

Điều kiện chuẩn :$\sum_{j=1}^{k} \hat{P}(c_j) = 1$

$\Rightarrow$ Áp dụng phương pháp nhân tử Lagrange:

$\mathcal{L}(\hat{P}(c_j), \Lambda) = \sum_{j=1}^{k} Count(c_j) \log(\hat{P}(c_j)) + \lambda(1 - \sum_{j=1}^{k} \hat{P}(c_j))$

Lấy đạo hàm riêng cho $\mathcal{L}$

$$\frac{\partial \mathcal{L}(\hat{P}(c_j), \lambda)}{\partial \hat{P}(c_j)} = \frac{Count(c_j)}{\hat{P}(c_j)} = 0$$

$$\Rightarrow \hat{P}(c_j) = \frac{Count(c_j)}{\lambda}(1)$$

$$\frac{\partial \mathcal{L}(\hat{P}(c_j), \lambda)}{\partial \lambda} = 1 - \sum_{j=1}^{k} \hat{P}(c_j) = 0$$

$$\Rightarrow \sum_{j=1}^{k} = 1(2)$$

Thế (1) vào (2) :

$$\Leftrightarrow \sum_{j=1}^{k} \frac{Count(c_j)}{\lambda} = 1$$

$$\Leftrightarrow \frac{1}{\lambda} \sum_{j=1}^{k} Count(c_j) = 1$$

$$\Leftrightarrow \frac{N_{doc}}{\lambda} = 1$$

$$\Rightarrow \lambda = N_{doc}$$

$$\Rightarrow \hat{P}(c_j) = \frac{Count(c_j)}{\lambda} = \frac{Count(c_j)}{N_{doc}}(dpcm)$$

2.2 Xác suất có điều kiện của từ w_i trong lớp c_j

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

Trong đó:

- $count(w_i, c_j)$: số lần từ w_i xuất hiện trong lớp c_j.
- $\sum_{w \in V} count(w, c_j)$: tổng tất cả các từ xuất hiện trong lớp c_j.

Giải

Xét các từ w_i là độc lập.
Theo phân phối categorical:
$P(count(w_i, c_j)|\hat{P}(w_i|c_j)) = \prod_{w_i \in V} \hat{P}(w_i|c_j)^{count(w_i,c_j)}$
Đánh giá $\hat{P}(w_i|c_j)$ dựa trên Maximum log-likelihood:

$$\mathscr{L}\hat{P}(w_i|c_j) = \arg\max[P(count(w_i, c_j)|\hat{P}(w_i|c_j))] = \arg\max[\prod_{w_i \in V} \hat{P}(w_i|c_j)^{count(w_i,c_j)}]$$

$$= \arg\max \sum_{w_i \in V} count(w_i, c_j) log(\hat{P}(w_i|c_j))$$

Điều kiện chuẩn : $\sum_{W_i \in V} \hat{P}(w_i|c_j) = 1$
$\Rightarrow$ Áp dụng phương pháp nhân tử Lagrange:

$$\mathscr{L}(\hat{P}(w_i|c_j), \lambda) = \sum_{w_i \in V} count(w_i, c_j) \log(\hat{P}(w_i|c_j)) + \lambda[1 - \sum_{w_i \in V} \hat{P}(w_i|c_j)]$$

Lấy đạo hàm riêng của L :

$$\frac{\partial \mathscr{L}(\hat{P}(w_i|c_j), \lambda)}{\partial \hat{P}(w_i|c_j)} = \frac{count(w_i, c_j)}{\hat{P}(w_i|c_j)} - \lambda = 0$$

$$\Rightarrow \hat{P}(w_i|c_j) = \frac{count(w_i, c_j)}{\lambda} \quad (1)$$

$$\frac{\partial \mathscr{L}(\hat{P}(w_i|c_j), \lambda)}{\partial \lambda} = 1 - \sum_{w_i \in V} \hat{P}(w_i|c_j) = 0$$

$$\Rightarrow \sum_{w_i \in V} \hat{P}(w_i|c_j) = 1 \quad \mathbf{(2)}$$

Thế **(1)** vào **(2)** :

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

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

$$\Leftrightarrow \lambda = \sum_{w_i \in V} count(w_i, c_j)$$

$$\Rightarrow \hat{P}(w_i, c_j) = \frac{count(w_i,c_j)}{\lambda} = \frac{count(w_i,c_j)}{\sum_{w_i \in V} count(w_i,c_j)}$$

$$hay \quad \hat{P}(w_i|c_j) = \frac{count(w_i,c_j)}{\sum_{w \in V}(count(w,c_j)}(dpcm)$$