

统计学习：第十八章

1 证明生成模型与共现模型是等价的。

答：假设单词 w 与文本 d 是关于 z 条件独立

$$\begin{aligned}\sum_{z \in Z} P(z)P(w|z)P(d|z) &= \sum_{z \in Z} P(w|z)P(d, z) \\ &= \sum_{z \in Z} P(w|z)P(z|d)P(d) \\ &= P(d) \sum_{z \in Z} P(w|z)P(z|d)\end{aligned}$$

等号由条件概率公式得到。

2 推导共现模型的 EM 算法。

答：

$$Q = \sum_{k=1}^K \sum_{j=1}^N \sum_{i=1}^M n(w_i, d_j) \log(P(z_k)P(d_j|z_k)P(w_i|z_k))P(z_k|w_i, d_j)$$

$P(z_k|w_i, d_j)$ 由上一步的贝叶斯公式得到，应用拉格朗日法得到，

$$\begin{aligned}P(z_k) &= \frac{\sum_{j=1}^N \sum_{i=1}^M n(w_i, d_j)P(z_k|w_i, d_j)}{\sum_{j=1}^N \sum_{i=1}^M n(w_i, d_j)} \\ P(d_j|z_k) &= \frac{\sum_{i=1}^M n(w_i, d_j)P(z_k|w_i, d_j)}{\sum_{j=1}^N \sum_{i=1}^M n(w_i, d_j)P(z_k|w_i, d_j)}, \\ P(w_i|z_k) &= \frac{\sum_{j=1}^N n(w_i, d_j)P(z_k|w_i, d_j)}{\sum_{j=1}^N \sum_{i=1}^M n(w_i, d_j)P(z_k|w_i, d_j)}\end{aligned}$$

3 对以下文本数据集进行概率潜在语义分析

答：见代码