第一章　Aspect Model 简述

pLSA 和 pLSI 可以用于文档分类和话题检测。他们的核心是 aspect model，下面我根据自己对文献 [1] 的理解给出了 aspect model 的推导，希望对读者有益。

假设整个语料库中单词集合为 $\mathcal{W} = \{w_1.w_2, \dots, w_m\}$，文档集合为 $\mathcal{D} = \{d_1, d_2, \dots, d_n\}$。每个单词文档对 (w, d) 都对应隐变量 $\mathcal{Z} = \{z_1, z_2, \dots, z_k\}$ 的一种概率分布 $p(z|w, d)$。隐变量可以理解为话题，我们假设有 k 个话题，每个单词文档对 (w, d) 被认为是经过话题而联系在一起的，并可以用一个生成模型描述这个过程：文档 d 属于话题 z 的概率令为 $p(z|d)$，而话题 z 中词 w 的出现概率令为 $p(w|z)$，从而词 w 出现在文档 d 里的概率为：

$$p(w|d) = \sum_{z \in \mathcal{Z}} p(z|d)p(w|z) \tag{1.1}$$

而单词 w，文档 d，隐变量 z 的联合概率为：

$$p(w, d, z) = p(d)p(z|d)p(w|z) = p(z, d)p(w|z) = p(z)p(d|z)p(w|z) \tag{1.2}$$

这就是两个模型假设。从而每个文档单词对 (w, d) 的概率为:

$$p(w, d) = p(d)p(w|d) = p(d) \sum_{z \in \mathcal{Z}} p(z|d)p(w|z) \tag{1.3}$$

现在我们有一个矩阵，每个矩阵元素 $n(d, w)$ 表示词 w 在文档 d 中出现的次数，然后我们采用极大似然方法估计上述生成模型中的参数。注意模型中的参数即是 $\boldsymbol{\theta} = \{p(w|z), p(d|z), p(z)|d \in \mathcal{D}, w \in \mathcal{W}, z \in \mathcal{Z}\}$，我们可以在求解这些参数后导出文档属于每个话题的概率方分布 $p(z|d) \propto p(d|z)p(z)$。而隐变量则为 $\{p(z|w, d)|d \in \mathcal{D}, w \in \mathcal{W}, z \in \mathcal{Z}\}$。我们用标准的 EM 推导规则，有如下结果：

$$\begin{aligned}
\log \mathcal{L} &= \log \prod_{d \in \mathcal{D}} \prod_{w \in \mathcal{W}} p(d, w)^{n(d,w)} \\
&= \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} n(d, w) \log p(d, w) \\
&= \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} n(d, w) \log \sum_{z \in \mathcal{Z}} p(d, w, z) \\
&= \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} n(d, w) \log \sum_{z \in \mathcal{Z}} p(z|w, d) \frac{p(d, w, z)}{p(z|w, d)} \\
&\geq \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} n(d, w) \sum_{z \in \mathcal{Z}} p(z|w, d) \log \frac{p(d, w, z)}{p(z|w, d)} \\
&= \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} n(d, w) \sum_{z \in \mathcal{Z}} p(z|w, d) \log p(d, w, z) - \\
&\quad \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} n(d, w) \sum_{z \in \mathcal{Z}} p(z|w, d) \log p(z|w, d)
\end{aligned} \tag{1.4}$$

然后，我们就可以推导 EM 算法的迭代公式。

E step 固定参数，优化隐变量的分布使得使上述推导中的大于等于中的等号成立。根据 Jensen 不等式，等号在且只在随机变量取值都相同才成立 (这一表述不严格，不过通常这样够用)，在这里也就是下面式子成立：

$$\frac{p(d,w,z_1)}{p(z_1|w,d)} = \frac{p(d,w,z_2)}{p(z_2|w,d)} = \cdots = \frac{p(d,w,z_k)}{p(z_k|w,d)} = c$$

从而，

$$\begin{aligned}
&\Rightarrow 1 = \sum_{z\in\mathcal{Z}} p(z|w,d) = \sum_{z\in\mathcal{Z}} \frac{p(d,w,z)}{c} \\
&\Rightarrow c = \sum_{z\in\mathcal{Z}} p(d,w,z) \\
&\Rightarrow p(z|w,d) = \frac{p(d,w,z)}{c} = \frac{p(d,w,z)}{\sum_{z\in\mathcal{Z}} p(d,w,z)} = \frac{p(z)p(d|z)p(w|z)}{\sum_{z\in\mathcal{Z}} p(z)p(d|z)p(w|z)}
\end{aligned} \tag{1.5}$$

M step 固定隐变量的分布，优化参数，最大化似然函数的下界。问题变为：

$$\begin{cases}
\max & \sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} n(d,w) \sum_{z\in\mathcal{Z}} p(z|w,d) \log p(z)p(d|z)p(w|z) \\
\text{s.t} & \sum_{z\in\mathcal{Z}} p(z) = 1 \\
& \sum_{d\in\mathcal{D}} p(d|z) = 1, z\in\mathcal{Z} \\
& \sum_{w\in\mathcal{W}} p(w|z) = 1, z\in\mathcal{Z}
\end{cases} \tag{1.6}$$

这个问题的拉格朗日函数为：

$$\begin{aligned}
L = &\sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} n(d,w) \sum_{z\in\mathcal{Z}} p(z|w,d) \log p(z)p(d|z)p(w|z) + \alpha(\sum_{z\in\mathcal{Z}} p(z) - 1) + \\
&\sum_{z\in\mathcal{Z}} \beta_z(\sum_{d\in\mathcal{D}} p(d|z) - 1) + \sum_{z\in\mathcal{Z}} \gamma_z(\sum_{w\in\mathcal{W}} p(w|z) - 1)
\end{aligned} \tag{1.7}$$

由 KKT 条件，则：

$$
\begin{cases}
\dfrac{\partial L}{\partial p(z)} = \sum\limits_{d\in\mathcal{D}}\sum\limits_{w\in\mathcal{W}} n(d,w)\dfrac{p(z|w,d)}{p(z)} + \alpha = 0 \\
\dfrac{\partial L}{\partial p(d|z)} = \sum\limits_{w\in\mathcal{W}} n(d,w)\dfrac{p(z|w,d)}{p(d|z)} + \beta_z = 0 \\
\dfrac{\partial L}{\partial p(w|z)} = \sum\limits_{d\in\mathcal{D}} n(d,w)\dfrac{p(z|w,d)}{p(w|z)} + \gamma_z = 0 \\
\sum\limits_{z\in\mathcal{Z}} p(z) = 1 \\
\sum\limits_{d\in\mathcal{D}} p(d|z) = 1 \\
\sum\limits_{w\in\mathcal{W}} p(w|z) = 1
\end{cases}
\tag{1.8}
$$

那么，

$$
\begin{aligned}
&\sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} n(d,w)\frac{p(z|w,d)}{p(z)} + \alpha = 0 \Rightarrow p(z) = -\frac{\sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} n(d,w)p(z|w,d)}{\alpha} \\
&\Rightarrow 1 = \sum_{z\in\mathcal{Z}} p(z) = \sum_{z\in\mathcal{Z}} -\frac{\sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} n(d,w)p(z|w,d)}{\alpha} \\
&\Rightarrow -\alpha = \sum_{z\in\mathcal{Z}}\sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} n(d,w)p(z|w,d) \\
&\Rightarrow p(z) = \frac{\sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} n(d,w)p(z|w,d)}{\sum_{z\in\mathcal{Z}}\sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} n(d,w)p(z|w,d)}
\end{aligned}
\tag{1.9}
$$

类似地：

$$
\begin{aligned}
&\sum_{w\in\mathcal{W}} n(d,w)\frac{p(z|w,d)}{p(d|z)} + \beta_z = 0 \Rightarrow p(d|z) = \frac{\sum_{w\in\mathcal{W}} n(d,w)p(z|w,d)}{-\beta_z} \\
&\Rightarrow 1 = \sum_{d\in\mathcal{D}} p(d|z) = \sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} \frac{n(d,w)p(z|w,d)}{-\beta_z} \\
&\Rightarrow -\beta_z = \sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} n(d,w)p(z|w,d) \\
&\Rightarrow p(d|z) = \frac{\sum_{w\in\mathcal{W}} n(d,w)p(z|w,d)}{\sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} n(d,w)p(z|w,d)}
\end{aligned}
\tag{1.10}
$$

以及，

$$
p(w|z) = \frac{\sum_{d\in\mathcal{D}} n(d,w)p(z|w,d)}{\sum_{w\in\mathcal{W}}\sum_{d\in\mathcal{D}} n(d,w)p(z|w,d)} \tag{1.11}
$$

从而，算法总结如下：

E step 固定参数，优化隐变量的分布。

$$p(z|w,d) = \frac{p(z)p(d|z)p(w|z)}{\sum_{z\in\mathcal{Z}} p(z)p(d|z)p(w|z)}$$

M step 固定隐变量的分布，优化参数。

$$p(w|z) = \frac{\sum_{d\in\mathcal{D}} n(d,w)p(z|w,d)}{\sum_{w\in\mathcal{W}}\sum_{d\in\mathcal{D}} n(d,w)p(z|w,d)}$$

$$p(d|z) = \frac{\sum_{w\in\mathcal{W}} n(d,w)p(z|w,d)}{\sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} n(d,w)p(z|w,d)}$$

$$p(z) = \frac{\sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} n(d,w)p(z|w,d)}{\sum_{z\in\mathcal{Z}}\sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} n(d,w)p(z|w,d)}$$

参考文献

[1] Hofmann T. Probabilistic latent semantic indexing. Proceedings of Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999. 50–57.