A Note on EM Algorithm and PLSA

Xinyan Lu

May 17, 2011

1 背景：PLSA模型

PLSA的全称是Probabilitic Latent Sematic Analysic，用作IR的时候也被称为PLSI (Probabilistic Latent Sematic Indexing)。[2]

下面简要介绍一下PLSA的基本思想。假设一共有 K 个主题，文档 d 的主题是这 K 个主题的叠加，即 $(p(z_1|d), p(z_2|d), \ldots, p(z_K|d))$，且 $\sum_{k=1}^{K} p(z_k|d) = 1$。也就是说，每篇文档可以属于不同的主题。每个主题 z 是一个概率分布，关于单词的分布，即单词的分布决定了它是什么主题。举个例子，“篮球”，“足球”这些词汇出现概率较高的话，会被认为是一个关于“体育”的主题。而文档 d 中每个单词 w 被认为是由如下过程产生的：先从 K 个主题中随机选择一个主题 z_k，注意其服从多元分布 Multi $(p(z_1|d), p(z_2|d), \ldots, p(z_K|d))$，然后根据 z_k 的单词分布 $p(w|z_k)$ 来随机选取一个单词。假设读者已有图模型相关知识[1]的简单基础，其图模型可参照Figure 1。

根据 PLSA 模型，待估计的参数为 $\theta = \{p(w|z_j), p(z_j|d)|w \in V, d \in C, 1 \leq j \leq k\}$，其中 C 为文档集合, V 表示在C中所有的词汇*，。而文档 C 的 log-likelihood 可以表示为：

$$\begin{aligned} L(\theta) = \log p(C|\theta) &= \sum_{d \in C} \sum_{w \in V} c(w,d) \times \log p(w,d) \\ &= \sum_{d \in C} \sum_{w \in V} c(w,d) \times \log \sum_{k=1}^{K} p(z_k|d) p(w|z_k) \end{aligned} \tag{1.1}$$

其中$c(w,d)$ 表示单词 w 在文档 d 中出现的次数。我们的任务是求得 θ 使得最大化此 log-likelihood。

试试对每个变量微分，然后求极值点。是不是很难？这个模型可以用 EM (Expectation Maximization) 算法来解得“最优”的 θ。注意在“最优”上面还有双引号，说明它不是真正意义上的最优解，它只是一个近似最优解，这点会在下面说明。

*实际上，往往会先把某些不含意义的词去掉，如 stopwords等，甚至有时候只留下具有某些特性的词语，根据实际需要来决定

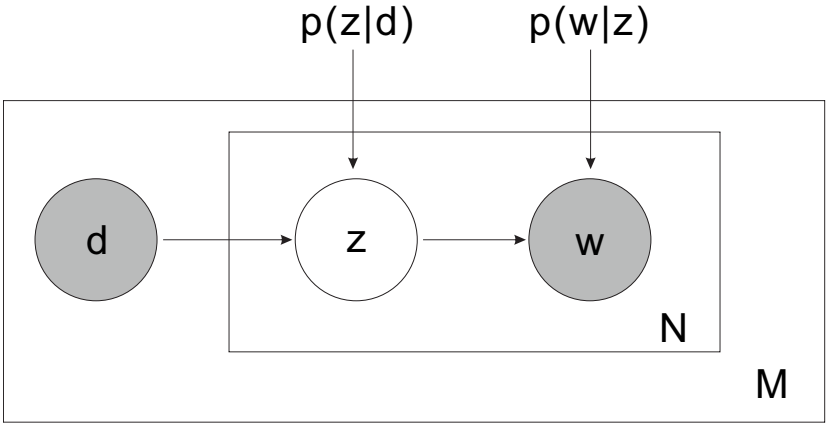

Figure 1: PLSA图模型

2 背景：EM 算法[3]

EM的基本思想是：随机选取†来初始化待估计的参数值 $\theta^{(0)}$， 然后不断迭代找更优的 $\theta^{(n+1)}$ 使得其 likelihood 或 log-likelihood，用 $L(\theta^{(n+1)})$ 来表示， 比原来的 $L(\theta^{(n)})$ 更大。 即我们现在已经得到了 $\theta^{(n)}$，想求 $\theta^{(n+1)}$ 使得：

$$\theta^{(n+1)} = \max_{\theta} L(\theta) - L(\theta^{(n)})$$

我们已经知道 log-likelihood 为 $L(\theta) = \log p(X|\theta)$，其中 X 表示已观察到的随机变量。现在转向 complete data log-likelihood 的概念。所谓的 complete data，是指图模型中的隐变量（latent variable）也被“观察”到了。 与 $L(\theta)$ 不同的是，

$$L_c(\theta) = \log p(X, H|\theta)$$

注意上式多了一个隐变量 H 在其中。$L_c(\theta)$ 与 $L(\theta)$ 的关系是：

$$L_c(\theta) = \log p(X, H|\theta) = \log p(X|\theta) + \log p(H|X, \theta) = L(\theta) + \log p(H|X, \theta)$$

引入上述概念后，则有：

$$L(\theta) - L(\theta^{(n)}) = L_c(\theta) - L_c(\theta^{(n)}) + \log \frac{p(H|X, \theta^n)}{p(H|X, \theta)}$$

对上式两边同时求关于 $p(H|X, \theta^n)$ 的期望，得：

$$\begin{aligned} L(\theta) - L(\theta^{(n)}) &= \sum_H L(\theta) p(H|X, \theta^n) - \sum_H L_c(\theta^{(n)}) p(H|X, \theta^n) \\ &\quad + \sum_H p(H|X, \theta^n) \log \frac{p(H|X, \theta^n)}{p(H|X, \theta)} \end{aligned}$$

†也可能是其它方法来确定，初值越接近真实值越好

注意上式左边项跟原来的是一样的，这是因为 H 没有出现在似然 L里面。 上式右手边最后一项可以看作是关于 $p(H|X,\theta^n)$ 和 $p(H|X,\theta)$ 的 KL-divergence，它的值总是非负的。则有：

$$L(\theta) - L(\theta^{(n)}) \geq \sum_H L(\theta)p(H|X,\theta^n) - \sum_H L_c(\theta^{(n)})p(H|X,\theta^n)$$

$$L(\theta) \geq \sum_H L_c(\theta)p(H|X,\theta^n) + L_c(\theta^{(n)}) - \sum_H L_c(\theta^{(n)})p(H|X,\theta^n) \tag{2.1}$$

因此我们就得到 $L(\theta)$ 的一个下界。EM算法的思想是不断最大化这个下界，从而不断最大化 $L(\theta)$。 注意到式（2.1）中的最后两项是跟 θ 无关的，它们被认为是常数，因此我们的最终任务是最大化：

$$E_{p(H|X,\theta^n)}[L_c(\theta)] = \sum_H L_c(\theta)p(H|X,\theta^n)$$

上式在 EM 中也被称为"Q-function"，记作 $Q(\theta;\theta^n)$。

综上所述， EM 算法的一般步骤为：

1. 随机选取或者根据某种先验知识来初始化 θ^0
2. 迭代地执行以下两步
   (a) E-step (expectation)：计算 $Q(\theta;\theta^n)$
   (b) M-step (maximization)：重新估计参数 θ，即求 $\theta^{(n+1)}$ 使得

$$\theta^{(n+1)} = \arg_\theta \max Q(\theta;\theta^n)$$

3. 直至 $L(\theta)$ 收敛（即 $Q(\theta;\theta^n)$ 收敛）的时候才停止，否则继续执行第二步。

注意 EM 并不保证收敛到最优解，而是收敛到局部极值点。 可以尝试取多次初始值，或者是将初始值取到跟最优解较近的位置（先用其它算法估计一个近似值）。

3 用 EM 求解 PLSA

在 PLSA 中，所谓的隐变量是 z，即每个单词 w 对应的主题。 为方便起见，我们采用以下表示方法：$\mathbf{t}$ 是一个 K 维的向量 $[t_1, t_2, \ldots, t_K]$，其中第 k 个元素 t_k的值为 1，其它元素的值均为 0。 当单词 w 对应的主题 $z = z_k$ 时，则相应的 $t_k = 1$。 现在假设我们“观察”到了每个单词背后的主题 t。与此对应的 complete data log-likelihood 为：

$$\begin{aligned} L_c(\theta) &= \log p(C,H|\theta) \\ &= \sum_{d\in C}\sum_{w\in V} c(w,d) \times \log \prod_{k=1}^{K} (p(z_k|d)p(w|z_k))^{t_k} \\ &= \sum_{d\in C}\sum_{w\in V} c(w,d) \times \sum_{k=1}^{K} t_k \log p(z_k|d)p(w|z_k) \end{aligned}$$

注意上式和（1.1）式不同的地方，当我们观察到 z_k 时，意味着单词出现的概率变为了 $p(z_k|d)p(w|z_k)$, 而不是原来的累加和，再利用 t 的性质，则有 $p(z_k|d)p(w|z_k) = \prod_{k=1}^{K}(p(z_k|d)p(w|z_k))^{t_k}$。写成这样的形式为了方便计算后面 E-step 中的期望。

E-step中很重要的一步，是在计算 Q-function 前，要先计算后验概率，有的时候可能会很难计算[‡]，所幸的是在 PLSA 中还是很容易计算的，通过简单的全概率公式得：

$$p(z_k|w,d) = \frac{p(z_k|d)p(w|z_k)}{\sum_{k'=1}^{K} p(z_{k'}|d)p(w|z_{k'})}$$

从而得到：

$$\begin{aligned} Q(\theta;\theta^n) &= \sum_{H} L_c(\theta)p(H|X,\theta^n) \\ &= \sum_{d\in C}\sum_{w\in V} c(w,d) \times \sum_{k=1}^{K} p(z_k|w,d)\log p(z_k|d)p(w|z_k) \end{aligned}$$

要最大化 Q-function，记得有约束项 $\sum_{k=1}^{K} p(z_k|d) = 1$ 以及 $\sum_{w\in V} p(w|z_k) = 1$，构造得拉格朗日乘子：

$$Q(\theta;\theta^n) + \sum_{d\in C}\tau_d(1 - \sum_{k=1}^{K} p(z_k|d)) + \sum_{k=1}^{K}\rho_k(1 - \sum_{w\in V} p(w|z_k))$$

对每个活动参数进行求导并使之等于0，联立方程解之,易解得：

$$p(z_k|d_i) = \frac{\sum_w c(w,d_i)p(z_k|d_i,w)}{\sum_w c(w,d_i)} \quad k = 1..K, d_i \in C$$

$$p(w_j|z_k) = \frac{\sum_d c(w_j,d)p(z_k|d,w_j)}{\sum_d\sum_w c(w,d)p(z_k|d,w)} \quad k = 1..K, w_j \in V$$

即为 EM 算法中的 M-step。重复 E-step 和 M-step 直至算法收敛到某一个 log-likelihood 值。

参考文献

[1] C. Bishop and S. O. service). Pattern recognition and machine learning, volume 4. Springer New York, 2006.

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

[3] C. Zhai. A note on the expectation-maximization (em) algorithm. Course note of CS410.

[‡]这时候就要用上更“高级”的方法如Approximation Inference了