PLSA EM 算法推导

1. PLSA 和 LDA

PLSA(Probability Latent Semantic Analysis)的图模型如图 1 所示。

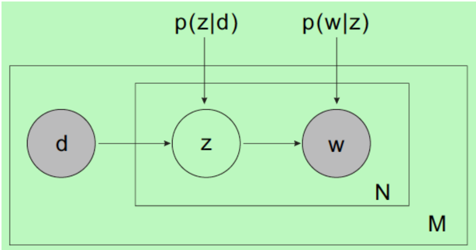

图 1. PLSA 的图模型

文档的生成过程如下，对于一篇文档 d，在每个词项的位置，首先选择一个 Topic，然后在 Topic 的词分布中，选择一个词，作为当前位置的词项 w。

PLSA 主题模型是比较老的模型，现在已经逐渐被 LDA(Latent Dirichlet Allocation)模型替代了。

2. EM 算法描述

假设现在有 5 个文档，6 个单词，那么我们可以得到 5×6 的文档-词项矩阵，如表 1 所示。

表 1. 文档-词项矩阵

	Apple	banana	grape	car	truck	train
D_1		……				
D_2			共现频率 $n(d_i,w_j)$			
D_3						
D_4						
D_5						

PLSA 的 EM 算法就根据这个文档-词项矩阵，计算出两个重要的分布，即每个文档在主题上的分布，以及每个主题在单词上的分布。假设主题的数量为 k=2，最后的结果可能如下。

表 2.文档的主题分布

	K_1	K_2
D_1	1.0	0
D_2	1.0	0
D_3	0	1.0
D_4	0	1.0
D_5	0.67	0.33

表 3.主题的单词分布

	W_1apple	W_2banana	W_3grape	W_4car	W_5truck	W_6train
K_1	0.5	0.2	0.15	0.05	0.05	0.05
K_2	0.05	0.05	0.05	0.5	0.2	0.15

3. EM 算法

输入的样本为(d_i, w_j)，这是可以观察的参数，需要估计的参数为：d_i 在主题上的分布 $p(z_k|d_i)$，以及主题在词项上的分布 $p(w_j|z_k)$。

对于(d_i,z_k,w_j)这样的完全的(Complete)样本，我们根据生成过程，有联合概率为 $p(d_i, z_k, w_j) = p(d_i)\, p(z_k|d_i)\, p(w_j|z_k)$。

在观察到(d_i, w_j)的情况下，z_k 的后验概率如下：注意$\gamma(z_{ijk})$即$p(z_k|d_i, w_j)$

$$\gamma(z_{ijk}) = p(z_k|d_i, w_j)$$

$$=\frac{p(d_i)p(z_k|d_i)p(w_j|z_k)}{\sum_{k=1}^{K}p(d_i)p(z_k|d_i)p(w_j|z_k)}=\frac{p(z_k|d_i)p(w_j|z_k)}{\sum_{k=1}^{K}p(z_k|d_i)p(w_j|z_k)}$$

这个概率是如何推导的呢？

首先，由贝叶斯公式，得到z_k的后验概率$p(z_k|w_j)=\frac{p(w_j|z_k)p(z_k)}{p(w_j)}=\frac{p(w_j|z_k)p(z_k)}{\sum_{k=1}^{K}p(w_j|z_k)p(z_k)}$。

上式同时除以d_i，有$p(z_k|d_i,w_j)=\frac{p(w_j|z_k)p(z_k|d_i)}{\sum_{k=1}^{K}p(w_j|z_k)p(z_k|d_i)}$(注意$d_i$对$w_j$是没有影响的)。

表 4. 文档-词项矩阵的共现频率的分配

	Apple	banana	grape	car	truck	train
D1						
D2			比如，$n(d_i,w_j)$按照 $p(z_1|d_i,w_j)$和$p(z_2|d_i,w_j)$分配给两个话题			
D3						
D4						
D5						

通过观察到的数据(d_i,w_j)，进行极大似然估计(目标函数是所有出现概率的乘积)

$$L=\prod_{i=1}^{N}\prod_{j=1}^{M}p(d_i,w_j)^{n(d_i,w_j)}$$

对 L 取对数，得到对数似然函数

$l=\log\ L=\sum_{i=1}^{N}\sum_{j=1}^{M}n(d_i,w_j)\log\ p(d_i,w_j)$

$=\sum_{i=1}^{N}\sum_{j=1}^{M}n(d_i,w_j)\log\ [\ p(w_j|d_i)p(d_i)\ ]$

(因为 $p(d_i,w_j)=\ p(w_j|d_i)p(d_i)$)

$=\sum_{i=1}^{N}\sum_{j=1}^{M}n(d_i,w_j)\log[\ (\sum_{k=1}^{K}p(z_k|d_i)p(w_j|z_k))p(d_i)\ ]$

(因为$p(w_j|d_i)=\sum_{k=1}^{K}p(z_k|d_i)p(w_j|z_k)$)

$=\sum_{i=1}^{N}\sum_{j=1}^{M}n(d_i,w_j)\log(\sum_{k=1}^{K}p(z_k|d_i)p(w_j|z_k))+\sum_{i=1}^{N}\sum_{j=1}^{M}n(d_i,w_j)\log p(d_i)$

在这里，可以认为$\log p(d_i)$为常数，目标函数剩下$\sum_{i=1}^{N}\sum_{j=1}^{M}n(d_i,w_j)\log(\sum_{k=1}^{K}p(z_k|d_i)p(w_j|z_k))$。

根据 Jensen 不等式，有

$\sum_i\log\ p(x^{(i)};\theta)=\sum_i\log\sum_{z^{(i)}}p(x^{(i)},z^{(i)};\theta)$

$=\sum_i\log\sum_{z^{(i)}}Q_i(z^{(i)})\frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}$

$\geq\sum_i\sum_{z^{(i)}}Q_i(z^{(i)})\log\frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}$

于是，得到整体期望函数为

$\sum_{i=1}^{N}\sum_{j=1}^{M}n(d_i,w_j)\log(\sum_{k=1}^{K}p(z_k|d_i)p(w_j|z_k))$

$\geq\sum_{i=1}^{N}\sum_{j=1}^{M}n(d_i,w_j)\left(\sum_{k=1}^{K}Q_i(z^{(i)})\log\frac{p(z_k|d_i)p(w_j|z_k)}{Q_i(z^{(i)})}\right)$

于是有$Q=\sum_{i=1}^{N}\sum_{J=1}^{M}n(d_i,w_j)\sum_{k=1}^{K}\gamma(z_{ijk})(\log\frac{p(z_k|d_i)p(w_j|z_k)}{\gamma(z_{ijk})})$

注意$\gamma(z_{ijk})$即$p(z_k|d_i,w_j)$。

(1) 对于$p(z_k|d_i)$，有$\sum_{k=1}^{K}p(z_k|d_i)=1$，即 di 在个话题上的分布概率之和为 1，根据拉格朗日乘子法，有如下代价函数：

$L=Q(\theta,\theta^{old})+\lambda(\sum_{k=1}^{K}p(z_k|d_i)-1)$

代价函数对$p(z_k|d_i)$求偏导(注意 i 和 k 固定了)，令其为 0(注意 log(x)的导数为 1/x)

$\left(\sum_{j=1}^{M} n\left(d_i, w_j\right) \gamma\left(z_{ijk}\right)\right) \frac{1}{p(z_k|d_i)} + \lambda = 0$

有$-\sum_{j=1}^{M} n\left(d_i, w_j\right) \gamma\left(z_{ijk}\right) = \lambda p(z_k|d_i)$

上 述 式 子 ， 左 右 两 边 对 K 个 主 题 求 和 (注 意 $\sum_{k=1}^{K} p(z_k|d_i) = 1$，同时 $\gamma\left(z_{ijk}\right)$ 对 k 求和为 1)，可以得到$\lambda = -\sum_{j=1}^{M} n\left(d_i, w_j\right)$

于是，有

$p(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i,w_j)\gamma(z_{ijk})}{\sum_{j=1}^{M} n(d_i,w_j)} = \frac{\sum_{j=1}^{M} n(d_i,w_j) p\left(z_k|d_i, w_j\right)}{n(d_i)}$

(2) 对于$p\left(w_j|z_k\right)$，有$\sum_{j=1}^{M} p\left(w_j|z_k\right) = 1$，即 zk 在个单词上的分布概率之和为 1，根据拉格朗日乘子法，有如下代价函数：

$L = Q\left(\theta, \theta^{old}\right) + \lambda(\sum_{k=1}^{K} p\left(w_j|z_k\right) - 1)$

代价函数对$p\left(w_j|z_k\right)$求偏导(注意 k 和 j 固定了)，令其为0(注意 log(x)的导数为 1/x)

$\left(\sum_{i=1}^{N} n\left(d_i, w_j\right) \gamma\left(z_{ijk}\right)\right) \frac{1}{p\left(w_j|z_k\right)} + \lambda = 0$

有$-\sum_{i=1}^{N} n\left(d_i, w_j\right) \gamma\left(z_{ijk}\right) = \lambda\ p\left(w_j|z_k\right)$

上述式子，左右两边对 M 个词进行累加(注意$\sum_{j=1}^{M} p\left(w_j|z_k\right) = 1$)，可以得到

$\lambda = -\sum_{i=1}^{N}\sum_{j=1}^{M} n\left(d_i, w_j\right) \gamma\left(z_{ijk}\right)$

于是，有

$p\left(w_j|z_k\right) = \frac{\sum_{i=1}^{N} n(d_i,w_j)\gamma(z_{ijk})}{\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i,w_j)\gamma(z_{ijk})} = \frac{\sum_{i=1}^{N} n(d_i,w_j) p\left(z_k|d_i, w_j\right)}{\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i,w_j) p\left(z_k|d_i, w_j\right)}$

4. EM 算法

最后，针对 PLSA 模型的 EM 算法的 E 步骤和 M 步骤总结如下。

EM 算法的 E 步

$p\left(z_k|d_i, w_j\right) = \frac{p(z_k|d_i) p\left(w_j|z_k\right)}{\sum_{k=1}^{K} p(z_k|d_i) p\left(w_j|z_k\right)}$

EM 算法的 M 步

$p(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i,w_j)\gamma(z_{ijk})}{\sum_{j=1}^{M} n(d_i,w_j)} = \frac{\sum_{j=1}^{M} n(d_i,w_j) p\left(z_k|d_i, w_j\right)}{n(d_i)}$

$p\left(w_j|z_k\right) = \frac{\sum_{i=1}^{N} n(d_i,w_j)\gamma(z_{ijk})}{\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i,w_j)\gamma(z_{ijk})} = \frac{\sum_{i=1}^{N} n(d_i,w_j) p\left(z_k|d_i, w_j\right)}{\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i,w_j) p\left(z_k|d_i, w_j\right)}$

5.参考文献

1. Probabilistic Latent Semantic Analysis. https://arxiv.org/pdf/1212.3900.pdf, 2020.
2. PLSA 的 EM 推导. https://www.cnblogs.com/zjgtan/p/3887132.html, 2020.
3. PLSA 介绍与推导. https://blog.csdn.net/iothouzhuo/article/details/51470076, 2020.