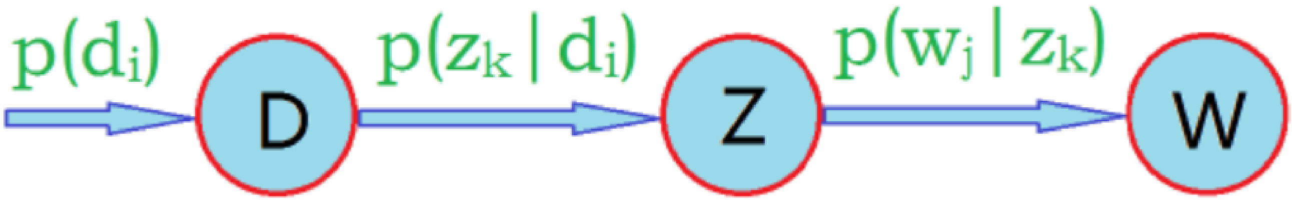
主题模型

今天来聊聊主题模型，这也是一个很好的把EM算法，贝叶斯理论串起来的好素材。在主题模型领域，LDA模型被人们熟知，但是在讲LDA之前，要先介绍一下它的前身pLSA。

1. pLSA

pLSA的全称叫做probabilistic Latent Semantic Analysis，概率语义分析模型。pLSA是一种简单的贝叶斯网络，可以用下面这幅图来描述。



定义D表示文档集合，Z表示主题集合，W表示词语集合。

定义概率$P(d_i)$表示文档d_i出现的概率；$P(z_k|d_i)$表示给定文档d_i，主题z_k出现的概率；$P(w_j|z_k)$表示给定主题z_k，词语w_j出现的概率。

每个主题在词语上服从多项分布，每个文档在主题上服从多项分布，即$P(z_k|d_i)$和$P(w_j|z_k)$均服从多项分布。假设一共有V个词，K个主题，则$P(z_k|d_i)$服从K点分布，$P(w_j|z_k)$服从V点分布。

给定一个语料库，里面有很多篇文档，可以观察词语和文档的对，第i篇文档的第j个词语(d_i, w_j)，它的联合概率分布可以表示为

$$P(d_i, w_j) = P(w_j|d_i)P(d_i)$$

而给定文档d_i，词语w_j的概率$P(w_j|d_i)$可以写成

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

这里多说一句，$P(w_j|z_k) = P(w_j|z_k, d_i)$，可以发现给不给$d_i$不影响$w_j$，因为已经给定了$z_k$，那么$d_i$和$w_j$就是一个head-to-tail的关系，这两个互相独立了。因此有这个等式关系。

可以发现$\sum_{k=1}^{K}$中的两项条件概率就是上文提到的给定文档的主题分布$P(z_k|d_i)$和给定主题的词分布$P(w_j|z_k)$，假定这两个分布由参数θ控制，因此可以写成$P(z_k|d_i, \theta)$和$P(w_j|z_k, \theta)$，但是为了简单起见，后文中如无特别情况，都简写为$P(z_k|d_i)$和$P(w_j|z_k)$。

因此为了求解这两个分布的参数，首先写出似然函数以及对数似然函数，将似然函数写成关于$P(z_k|d_i)$和$P(w_j|z_k)$为参数的形式

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

其中N和M分别表示语料库中文档的个数和词语的个数，$n(d_i, w_j)$表示d_i和w_j的文档-词语对的个数。

对数似然函数为

$$
\begin{aligned}
l &= \sum_i \sum_j n\left(d_i, w_j\right) \log P\left(d_i, w_j\right) \\
&= \sum_i \sum_j n\left(d_i, w_j\right) \log \left(P\left(w_j \mid d_i\right) P\left(d_i\right)\right) \\
&= \sum_i \sum_j n\left(d_i, w_j\right) \log \left(\sum_{k=1}^K P\left(w_j \mid z_k\right) P\left(z_k \mid d_i\right) P\left(d_i\right)\right)
\end{aligned}
$$

发现对数似然函数包括了$P(z_k|d_i)$和$P(w_j|z_k)$。观察对数似然函数，发现包含隐变量z_k，因此采用EM算法求解该似然函数最大。

E步，写出隐变量的后验概率$P(z_k|w_j,d_i)$以及Q函数的表达。假设迭代到了第t轮，参数为θ_t，那么在θ_t的控制下，$P(z_k|d_i)$和$P(w_j|z_k)$是已知的，那$P(z_k|w_j,d_i,\theta_t)$可以写成

$$
P(z_k|w_j,d_i,\theta_t) = \frac{P(w_j|z_k)P(z_k|d_i) = P(w_j,z_k|d_i)}{\sum_{l=1}^K P(w_j|z_l)P(z_l|d_i) = P(w_j|d_i)}
$$

回顾似然函数，

$$
\begin{aligned}
l &= \sum_i \sum_j n\left(d_i, w_j\right) \log P\left(d_i, w_j\right) \\
&= \sum_i \sum_j n\left(d_i, w_j\right) \log \left(P\left(w_j \mid d_i\right) P\left(d_i\right)\right) \\
&= \sum_i \sum_j n\left(d_i, w_j\right) \left(\log P(w_j|d_i) + \log P(d_i)\right) \quad \text{把 } \log \text{相 乘 写 开 成 相 加} \\
&= \left[\sum_i \sum_j n\left(d_i, w_j\right) \log P(w_j|d_i)\right] + \left[\sum_i \sum_j n\left(d_i, w_j\right) \log P(d_i)\right]
\end{aligned}
$$

发现等式右边那一项中仅包含$n(d_i,w_j)$和$\log P(d_i)$是可以通过语料库数出来的，可以看作常数省略，因此似然函数可以写成

$$
\begin{aligned}
l &= \sum_i \sum_j n\left(d_i, w_j\right) \log P(w_j|d_i) \\
&= \sum_i \sum_j n\left(d_i, w_j\right) \log \left[\sum_k P(w_j|z_k)P(z_k|d_i)\right] \\
&= \sum_i \sum_j n(d_i, w_j) \log \left[\sum_k P(z_k|w_j,d_i,\theta_t) \frac{P(w_j|z_k)P(z_k|d_i)}{P(z_k|w_j,d_i,\theta_t)}\right] \\
&\geq \sum_i \sum_j n\left(d_i, w_j\right) \sum_k P(z_k|w_j,d_i,\theta_t) \log \frac{P(w_j|z_k)P(z_k|d_i)}{P(z_k|w_j,d_i,\theta_t)}
\end{aligned}
$$

其中，θ_t表示第t轮迭代完成后，得到的模型参数。

可以看到对数似然函数的下界，定义为

$$
B(\theta,\theta_t) = \sum_i \sum_j n\left(d_i, w_j\right) \sum_k P(z_k|w_j,d_i,\theta_t) \log \frac{P(w_j|z_k)P(z_k|d_i)}{P(z_k|w_j,d_i,\theta_t)}
$$

通过不断最大化下界，来提升似然函数。观察下界函数$B(\theta,\theta_t)$中$\log$项的分母，对于$\log$函数，分母多除一个常数，不影响$\arg\max$，因此在$\log$项中多除以一个常数$P(w_j,d_i|\theta_t)$，因此分母变成了

$$P(z_k|w_j,d_i,\theta_t)P(w_j,d_i|\theta_t)=P(z_k,w_j,d_i|\theta_t)$$

对于$\sum_k$而言是个常数，因此可以省略掉。所以对下界函数求最大可以写成

$$\arg\max_{\theta} B(\theta,\theta_t)=\arg\max_{\theta}\sum_i\sum_j n\left(d_i,w_j\right)\sum_k P(z_k|w_j,d_i,\theta_t)\log P(w_j|z_k)P(z_k|d_i)$$

那么Q函数可以写成

$$Q(\theta,\theta_t)=\sum_i\sum_j n\left(d_i,w_j\right)\sum_k P(z_k|w_j,d_i,\theta_t)\log P(w_j|z_k)P(z_k|d_i)$$

M步，更新参数θ从而更新$P(z_k|d_i)$和$P(w_j|z_k)$。

已知目标函数为$Q(\theta,\theta_t)$，并且要求的参数$P(z_k|d_i)$和$P(w_j|z_k)$需要满足约束条件

$$\begin{cases}\sum_{j=1}^{M} P\left(w_j|z_k\right)=1\\ \sum_{k=1}^{K} P\left(z_k|d_i\right)=1\end{cases}$$

因此这是一个给定等式约束条件的最大化似然函数问题，可以使用Lagrange乘子法求解。

写出拉格朗日函数，

$$\begin{aligned}\mathcal{L}(\theta,\tau,\rho)=&\sum_i\sum_j n\left(d_i,w_j\right)\sum_{k=1}^{n} P\left(z_k|d_i,w_j\right)\log P\left(w_j|z_k\right)P\left(z_k|d_i\right)\\&+\sum_{k=1}^{K}\tau_k\left(1-\sum_{j=1}^{M} P\left(w_j|z_k\right)\right)+\sum_{i=1}^{N}\rho_i\left(1-\sum_{k=1}^{K} P\left(z_k|d_i\right)\right)\end{aligned}$$

参数θ只在$P(w_j|z_k)$和$P(z_k|d_i)$中，因此将拉格朗日函数对这两个参数求偏导数

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial P\left(w_j|z_k\right)}&=\frac{\sum_i n\left(d_i,w_j\right)P\left(z_k|d_i,w_j\right)}{P\left(w_j|z_k\right)}-\tau_k=0\\\frac{\partial\mathcal{L}}{\partial P\left(z_k|d_i\right)}&=\frac{\sum_i n\left(d_i,w_j\right)P\left(z_k|d_i,w_j\right)}{P\left(z_k|d_i\right)}-\rho_i=0\end{aligned}$$

对第一个式子进行分析

$$\sum_i n\left(d_i,w_j\right)P\left(z_k|d_i,w_j\right)=\tau_k P(w_j|z_k)$$

回顾词语个数一共有M个，因此想要把右边的$P(w_j|z_k)$积分掉，需要在两边同时加上$\sum_{j=1}^{M}$，即

$$\begin{aligned}\sum_{j=1}^{M}\sum_i n(d_i,w_j)P(z_k|d_i,w_j)&=\sum_{j=1}^{M}\tau_k P(w_j|z_k)\\&=\tau_k\sum_{j=1}^{M}P(w_j|z_k)=\tau_k\end{aligned}$$

将τ_k带回式子，将$P(w_j|z_k)$移到一边，得到

$$
\begin{aligned}
P(w_j|z_k) &= \frac{\sum_i n\left(d_i, w_j\right) P\left(z_k|d_i, w_j\right)}{\tau_k} \\
&= \frac{\sum_i n\left(d_i, w_j\right) P\left(z_k|d_i, w_j\right)}{\sum_{j=1}^M \sum_i n(d_i, w_j) P(z_k|d_i, w_j)}
\end{aligned}
$$

同理可以得到

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

因此，M步的过程中更新$P(w_j|z_k)$和$P(z_k|d_i)$的公式就推导出来了。

之后就是不停的迭代即可，以上就完成了pLSA全部的推导。