# 浅谈LDA

—— fangz_z

最近在看LDA相关的推导，大致掌握了来龙去脉，趁着脑子还清醒，做个笔记，以防过几天又忘了

## 从PLSA讲起

对于文本建模的主题模型，我们首先要从PLSA讲起，当然在此之前，还有一些十分粗浅的模型，我们就不多加赘述了。
PLSA将文本的生成分成两个步骤：doc-topic和topic-word。对于每一篇文档，在生成每个单词的时候，首先选择一个主题z，然后在该主题下抽取一个单词，依次生成所有单词。对于文档m，其doc-topic参数记为$\vec{\theta_m}$，$\theta_{mk}$即文档m中选择主题k的概率；对于主题k，其topic-word参数记为$\vec{\beta_k}$，$\beta_{kv}$即主题k中选择单词v的概率。于是整篇文档的生成概率为：

$$p(\vec{w}|d_m) = \prod_{i=1}^{n}\sum_{z} p(w_i|z)p(z|d_m) = \prod_{i=1}^{n}\sum_{k=1}^{K}\beta_{kw_i}\theta_{mk}$$

这个模型是没有解析解的，可以使用EM算法求得局部最优解，有空可以试试看

## 加入贝叶斯!

从贝叶斯学派的角度来看，上面的做法就太不地道了。参数怎么能是个确定的值呢？显然它们也是随机变量，给它们引入先验分布！
因为$\vec{\theta_m}, \vec{\beta_k}$都对应到多项分布，所以计算方便起见，将其先验分布设为狄利克雷分布。

这里准备一点数学基础，多项分布和狄利克雷分布是一对共轭分布，多项分布形式为：

$$Multinomial(\vec{m}|n,\vec{p}) = \frac{n!}{\prod_{i=1}^{M} m_i!}\prod_{i=1}^{M} p_i^{m_i}, n = \sum_{i=1}^{M} m_i$$

参数的先验分布：

$$Dirichlet(\vec{p}|\vec{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K} p_k^{\alpha_k - 1} = \frac{1}{\Delta(\alpha)}\prod_{k=1}^{K} p_k^{\alpha_k - 1}$$

那么根据贝叶斯定理，计算得到参数的后验分布仍然是狄利克雷分布，

$$p(\vec{p}|\vec{m},\vec{\alpha}) \propto p(\vec{p}|\vec{\alpha})p(\vec{m}|\vec{p}) = Dir(\vec{p}|\vec{\alpha}+\vec{m})$$

数据的作用就是改变了参数分布的计数，这很容易验证。并且狄利克雷分布的期望也是易得的，

$$E(Dir(\vec{p}|\vec{\alpha})) = (\frac{\alpha_1}{\sum_{k=1}^{K}\alpha_k}, ..., \frac{\alpha_K}{\sum_{k=1}^{K}\alpha_k})$$

于是，对于任一文档，其主题分布参数$\vec{\theta_m} \sim Dir(\vec{\alpha})$；对于任一主题，其词分布参数$\vec{\beta_k} \sim Dir(\vec{\eta})$。记$\vec{w},\vec{z}$分别为语料库的词分布和主题分布，具体来说，$w_i$的下标i是一个二维下标mn，指的是第m篇文档的第n个词，$\vec{w}$和$\vec{z}$是一一对应的，每个词背后都对应着一个主题，所以$\vec{w}$和$\vec{z}$的长度都等于语料库的总词数。每个$z_i$根据参数$\vec{\theta_m}$生成，而$w_i$则根据参数$\vec{\beta_{z_i}}$得到。
有了实际数据，我们就可以计算参数的后验分布，$\vec{n_k},\vec{n_m}$分别表示主题k中各个词的计数和文档m中各个主题的计数，$\vec{n_k} = (n_k^1, n_k^2, ..., n_k^V), \vec{n_m} = (n_m^1, n_m^2, ..., n_m^K)$，它们显然服从参数分别为$\vec{\beta_k},\vec{\theta_m}$的多项分布，根据前面的数学知识，后验分布仍然是狄利克雷分布，只不过要加上实际数据中的计数。所以，

$$p(\vec{\theta_m}|\vec{z}) \sim Dir(\vec{\alpha}+\vec{n_m})$$

$$p(\vec{\beta_k}|\vec{w}) \sim Dir(\vec{\eta}+\vec{n_k})$$

那么我们最终的问题是什么呢？看了很多博客和科普文章，都说是要求解$\vec{z}$，但这不是废话嘛-_-||。。。在我看来，因为$\vec{w}$是由语料库给定的，我们要做的其实就是最大$\vec{z}$的条件概率，即

$$\max p(\vec{z}|\vec{w})$$

根据贝叶斯定理，$p(\vec{z}|\vec{w}) \propto p(\vec{z},\vec{w})$，我们去求$\vec{z},\vec{w}$联合分布的最大就好了。那么，我们试着把他们联合分布的形式写出来：
先考虑某一篇文档m的主题分布，对所有可能的参数分布做积分，

$$
\begin{aligned}
p(\vec{z_m}|\vec{\alpha}) &= \int p(\vec{z_m}|\vec{\theta_m})p(\vec{\theta_m}|\vec{\alpha})d\vec{\theta_m} \\
&= \int \prod_{k=1}^{K} \theta_{mk}^{n_m^k} \frac{1}{\Delta(\alpha)} \prod_{k=1}^{K} \theta_{mk}^{\alpha_k - 1} d\vec{\theta_m} \\
&= \frac{1}{\Delta(\alpha)} \int \prod_{k=1}^{K} \theta_{mk}^{n_m^k + \alpha_k - 1} d\vec{\theta_m} \\
&= \frac{\Delta(\vec{n_m} + \vec{\alpha})}{\Delta(\vec{\alpha})}
\end{aligned}
$$

同理，主题k下的词分布也有表达式，

$$
p(\vec{w_k}|\vec{z}, \vec{\eta}) = \frac{\Delta(\vec{n_k} + \vec{\eta})}{\Delta(\vec{\eta})}
$$

这样，我们就能写出联合分布的形式：

$$
p(\vec{z}, \vec{w}|\vec{\alpha}, \vec{\eta}) = p(\vec{z}|\vec{\alpha})p(\vec{w}|\vec{z}, \vec{\eta}) = \prod_{m=1}^{M} \frac{\Delta(\vec{n_m} + \vec{\alpha})}{\Delta(\vec{\alpha})} \prod_{k=1}^{K} \frac{\Delta(\vec{n_k} + \vec{\eta})}{\Delta(\vec{\eta})}
$$

有了联合分布，我们就可以开始求解过程了！

# 绝妙的Gibbs采样

说到Gibbs采样，就又是可以单独写一篇文章的内容了（本来确实打算写一篇关于各种采样的文章的，但是看了一下，网上已经有很多总结得很好的了，所以暂时不想动笔了，有空再说）。受限于篇幅，在这里只能大概描述一下算法思想。
对于已知分布的采样，其实并不像人们想象中的那样轻松，只有对离散分布和少数常见的连续分布采样是容易的，其他的分布我们都不得不曲线救国，比如拒绝接受采样，重要性采样等等。但这些方法的问题就是，对于高维分布有些无能为力。于是就发明了著名的MCMC。
利用马尔可夫链的收敛性，当n次转移达到平稳分布后，再次经过转移矩阵仍然是平稳分布，这样我们可以认为从任意初始状态$x_0$出发，沿着马氏链转移，得到序列$x_0, x_1, ..., x_n, x_{n+1}, ...$，而序列$(x_n, x_{n+1}, ...)$就是来自于平稳分布的样本。那这里的关键就是，怎样为一个已知的分布去寻找对应的马尔科夫转移矩阵——MCMC和后来M-H算法借鉴了拒绝接受采样的思想，解决了这个问题。但是，MCMC面对维度特别高的情况，还是有些效率低下，最后，Gibbs采样闪亮登场。
Gibbs采样巧妙地证明了使用条件概率作为转移矩阵，能够满足细致平衡条件，于是对于高维分布，我们只需要在每个坐标轴轮换采样，利用它的条件概率作为转移概率，最后就能收敛到平稳分布，从而实

现了对已知分布的采样。
*如果以上内容看得一头雾水，建议先复习采样相关知识*
说回到我们的LDA，它和Gibbs采样有什么关系呢？很多科普博客都是这样一笔带过，

> 有了条件概率分布$p(z_i = k|\vec{w}, \vec{z_{\neg i}})$，我们就可以进行Gibbs采样，最终在Gibbs采样收敛后得到第i个词的主题。

**???** why？关于Gibbs采样和LDA的关系，这里我先埋个坑，最后再说。
好了，不管怎么样，我们先按照Gibbs采样的步骤走，对$p(\vec{w}, \vec{z})$采样，我们需要的是各个分量坐标轴上的条件概率，注意$\vec{w}$是给定的，所以不需要进行采样，于是我们只需要$p(z_i|\vec{w}, \vec{z_{\neg i}})$，（$\vec{z_{\neg i}}$指的就是主题分布中不包含下标i的剩下部分，其余的同理）

$$
\begin{aligned}
p(z_i = k|\vec{w}, \vec{z_{\neg i}}) &\propto p(z_i = k, w_i = t|\vec{w_{\neg i}}, \vec{z_{\neg i}}) \\
&= p(z_i = k|\vec{w_{\neg i}}, \vec{z_{\neg i}}) p(w_i = t|z_i = k, \vec{w_{\neg i}}, \vec{z_{\neg i}}) \\
&= \int p(z_i = k|\vec{\theta_m}) p(\vec{\theta_m}|\vec{w_{\neg i}}, \vec{z_{\neg i}}) d\vec{\theta_m} \int p(w_i = t|\vec{\beta_k}) p(\vec{\beta_k}|\vec{w_{\neg i}}, \vec{z_{\neg i}}) d\vec{\beta_k} \\
&= \int \theta_{mk} Dir(\vec{\theta_m}|\vec{n_{d,\neg i}} + \vec{\alpha}) d\vec{\theta_m} \int \beta_{kt} Dir(\vec{\beta_k}|\vec{n_{k,\neg i}} + \vec{\eta}) d\vec{\beta_m} \\
&= E(Dir(\theta_{mk})) E(Dir(\beta_{kt})) \\
&= \frac{n^k_{m,\neg i} + \alpha^k}{\sum_{s=1}^K n^s_{m,\neg i} + \alpha^s} \cdot \frac{n^t_{k,\neg i} + \eta^t}{\sum_{f=1}^V n^f_{m,\neg i} + \eta^f}
\end{aligned}
$$

可以看到，条件概率由两个部分相乘得到，也就是doc-topic结构和topic-word结构。
至此，LDA的Gibbs采样环节理论推导部分就已经结束了，与繁复的理论推导相比，LDA的实验环节简单得令人惊讶

```
1. 给语料库中每个单词w随机分配一个主题z
2. 扫描语料库，给每个单词，根据前面的推导的条件概率，重新采样分配一个主题z'
3. 重复过程2直至Gibbs采样收敛
4. 统计语料库，得到各文档的主题分布以及各主题下的词分布
```

上面是训练算法，通过训练我们可以得到topic-word分布，这个分布可以为我们处理新文档提供帮助，当我们需要得到一篇新文档的主题分布时，仍然是先给每个单词随机分配主题，然后通过Gibbs采样更新主题，但此时我们并不更新topic-word部分，而是使用已有的分布，也就是说上面条件概率的第二项是固定的。

# LDA--Gibbs?

好了，现在让我们回过头来想想这个问题——Gibbs采样和LDA到底有什么关系？
（下面完全是个人理解，如有错误还请指正）
Gibbs方法只是一个采样手段，它确保了当达到收敛状态之后，我们采集的样本是来自于给定的分布的。那么思考一下，LDA里的给定分布是什么？没错，就是联合分布$p(\vec{w},\vec{z})$，当采样收敛后，根据收敛性，我们就能保证采样的结果是来自于分布$p(\vec{w},\vec{z})$的，但这似乎跟我们之前设定的目标有点风马牛不相及啊！还记得我们之前的优化目标是什么吗，$\max p(\vec{w},\vec{z})$才是我们的终极追求啊，怎么现在变成了从分布中随便采集一个样本了Σ( ° △ °|||)︴ 甚至可能就有人会问，那我岂不是随便挑一个$(\vec{w},\vec{z})$就行了吗，反正也是一个样本，搞这么麻烦干嘛！
个人理解，$p(\vec{w},\vec{z})$是一个非常高维空间中的概率分布，所以它可能是十分尖锐的，概率质量只集中在某些很小的区域，所以从这个分布中采样，大概率会收敛到局部最优解附近。如果是随便挑选的样本$(\vec{w},\vec{z})$，其概率密度极大概率会接近0，是不可能出现在Gibbs采样的结果中的。
从这个角度，我们也能理解，Gibbs采样的结果只会是LDA模型的一个近似解，甚至可以说是一个非常随意的解，但这样在理论上十分粗糙的结果，在应用中却取得了巨大的成功，不禁让人感叹——管他呢，能用就行……

关于LDA的求解，还有一个EM算法的版本，后面有空继续填坑

—— 2020.9.3