文献调研：隐含狄利克雷分布（Latent Dirichlet Allocation）

PB16111485 张劲暾

模型描述：

模型类型：

无监督的，抽取离散特征的生成模型

基本假设：

对于一篇文章，我们假设其是由k个潜在的主题和|V|个对应的词语随机联合采样生成的，其中k个主题对应的分布服从狄利克雷分布，而词典中的词语在每一个主题上服从多项式分布。对于一篇文章可以由类似于 Unigram Model 的思想不断地采样主题，再对得到的主题采样一个词语得到，实际上类似于加了Dirichlet先验的Unigram Model ，那么对于文章模型的描述，就是由主题分布的狄利克雷分布的参数和对于每一个主题的词语多项式分布生成的。

> LDA是一种典型的词袋模型，即它认为一篇文档是由一组词构成的一个集合，词与词之间没有顺序以及先后的关系。一篇文档可以包含多个主题，文档中每一个词都由其中的一个主题生成。

准确定义：

对于一篇有N个词的文章：

$$\vec{\omega} =< \omega_1, \ldots, \omega_N >$$

由如下过程采样生成：

第一步：

由狄利克雷分布：$Dirichlet(\alpha_1, \ldots, \alpha_k)$采样得到主题分布$\vec{\theta}$,

第二步：

对于主题$z_n \in \{1, \ldots, k\}, P(z_n = i|\vec{\theta}) = \theta_i$，然后由多项式分布$Mult(\vec{\theta})$采样选定一个主题$z_n$

第三步：

用服从概率分布$p(\vec{\omega}|z_n)$的多项式分布采样产生每个词ω_n

最终得到文章词语的联合分布：

$$p(\vec{\omega}) = \int_{\theta}(\prod_{n=1}^{N}\sum_{z_n=1}^{k}p(\omega_n|z_n;\beta)p(z_n|\vec{\theta}))p(\vec{\theta};\vec{\alpha})d\vec{\theta}$$

整个模型的参数包括狄利克雷分布的k维参数$\vec{\alpha}$和k个主题上的$|V|$维多项式分布参数β

模型的学习方法：

基本思想1：用对数最大似然的方法简化形式，然后用EM算法估计参数

基本思想2：Dirichlet先验分布 + 多项式分布 = Dirichlet后验分布

似然函数形式推导过程和EM算法：

用指示变量$\omega_n^j = 1$表示第j个词语在文本中，$z_n^i = 1$表示文章属于第i个主题，

$$\beta_{ij} = P(\omega^j = 1|z^i = 1)$$

则有：$p(\vec{\omega};\vec{\alpha},\beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\int_{\theta}(\prod_{n=1}^{N}\sum_{z_n=1}^{k}\prod_{j=1}^{|V|}(\theta_i\beta_{ij})^{\omega_n^j})(\prod_{i=1}^{k}\theta_i^{\alpha_i-1})d\vec{\theta}$

但是这种形式过于复杂，所以进一步推导其对数形式：

$$logP(\vec{\omega};\vec{\alpha},\beta) = log\int_{\theta}\sum_{\vec{z}}\frac{p(\vec{\omega}|\vec{z};\beta)p(\vec{z}|\vec{\theta})p(\vec{\theta}|\vec{\alpha})}{q(\vec{\theta},\vec{z};\gamma,\phi)}d\theta$$

$$\geq E_q[logP() + logP() + log() - log]$$

在EM算法中：

$$E步：\ logp(D) \geq \sum_{m=1}^{M}E_{q_m}[logP(\vec{\theta},\vec{z},\vec{w})] - E_{q_m}[logQ_m(\vec{\theta},\vec{z})],$$

$$D = \{\vec{w}_1, \ldots, \vec{w}_M\}$$

$$M步：\ \beta ij \propto \sum_{m=1}^{M}\sum_{n=1}^{\vec{w}_m}\phi_{mni}\omega_{mn}^j$$

$$\frac{\partial \ell}{\partial \alpha_i} = \sum_{m=1}^{M}(\Psi(\sum_{j=1}^{k}\alpha_j) - \Psi(\alpha_i)) + (\Psi(\gamma_{mi}) - \Psi(\sum_{j=1}^{k}\gamma_{mj}))$$

这样通过EM算法不断优化这个下限，从而按照最大似然的原理学习得到模型的参数。

模型要解决的问题：

模型的出发点是从离散的文本数据中寻找主题构建生成模型，但其实对于文本分类，协同过滤乃至图像处理都有一定的效果或者启发，这一点在LDA的非参数化改进模型HDP中体现更加明显。

相关模型以及它们相对于LDA的不足：

1. **Unigram Model**

（当上帝只有一颗骰子的时候，那么他只能按照一定的概率分布去产生词）

$$对于文档\vec{\omega} = (\omega_1, \omega_2, \cdots, \omega_n), 用p(\omega_n)表示词\omega_n的先验概率，生成文档\vec{\omega}的概率为$$

$$P(\vec{\omega}) = \prod_{n=1}^{N} P(\omega_n)$$

不足：完全没有考虑主题因素，而且粗糙地认为所有的文章词语分布相同，扩展性差

2. **PLSI(Probabilistic Latent Semantic Analysis)**

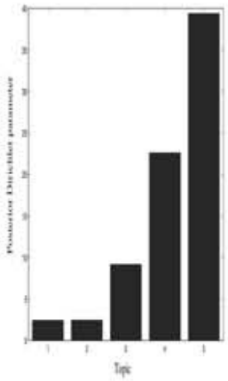
PLSI的思想已经非常接近于LDA了，即通过一定的主题分布采样得到一个主题，再根据这个采样得到的主题去采样得到词语，不同的关键点在于，PLSI认为每篇文章的概率分布是确定的，学习方法是确定这个分布，而对于LDA，这个分布也是随机的，通过Dirichlet分布采样产生的，直观地表现为PLSI的模型参数在主题分布上是一个文档-主题概率分布矩阵，而LDA与之对应的仅仅是狄利克雷分布的参数向量。

测试内容与效果（仅参照参考资料，没有亲自实验）：

测试集：

1. TREC AP corpus : 2500 articles, 37871 words
2. CRAN corpus: 1400 articles, 7747 words

实验结果：产生的文章主题分布以及对应主题的大概率产生词语基本符合常识



Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
SCHOOL	MILLION	SAID	SAID	SAID
SAID	YEAR	AIDS	NEW	NEW
STUDENTS	SAID	HEALTH	PRESIDENT	MUSIC
BOARD	SALES	DISEASE	CHIEF	YEAR
SCHOOLS	BILLION	VIRUS	CHAIRMAN	THEATER
STUDENT	TOTAL	CHILDREN	EXECUTIVE	MUSICAL
TEACHER	SHARE	BLOOD	VICE	BAND
POLICE	EARNINGS	PATIENTS	YEARS	PLAY
PROGRAM	PROFIT	TREATMENT	COMPANY	WON
TEACHERS	QUARTER	STUDY	YORK	TWO
MEMBERS	ORDERS	IMMUNE	SCHOOL	AVAILABLE
YEAROLD	LAST	CANCER	TWO	AWARD
GANG	DEC	PEOPLE	TODAY	OPERA
DEPARTMENT	REVENUE	PERCENT	COLUMBIA	BEST

Figure 3: The Dirichlet parameters where $\gamma_i > 1$ ($k = 100$), and the top 15 words from the corresponding topics, for the document discussed in the text.

模型试用范围：

这个模型有一定的适用范围，这是通过做课程实验得到的，如果我们用一条汽车评论的主题概率分布作为描述一条汽车评论的特征向量并用于文本分类，那么这个模型所体现出的信息量还不如TFIDF模型，但至于为什么效果不好？LDA适用于哪些问题和哪些数据？如何去做相应的提升，由于时间有限，并没有做过多的探索性阅读和实验，这是留下的问题和补充点。

一些改进提升：

1. **EM算法求解过于复杂**

除了[1]中提到的原始的用EM算法不断优化文档概率下界之外，[5]还提供了用Gibbs采样算法学习模型参数，得到更简单的计算步骤，大体思想如下：

所有文档联合起来形成的词向量$\vec{w}$是已知的数据，不知道的是语料库主题$\vec{z}$的分布。

假如我们可以先求出$\vec{w},\vec{z}$的联合分布$p(\vec{w},\vec{z})$，

进而可以求出某一个词w_i对应主题特征z_i的条件概率分布$p(z_i = k|\vec{w},\vec{z}_{\neg i})$。

其中，$\vec{z}_{\neg i}$代表去掉下标为i的词后的主题分布。

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。

如果我们通过采样得到了所有词的主题,那么通过统计所有词的主题计数，就可以得到各个主题的词分布。接着统计各个文档对应词的主题计数，就可以得到各个文档的主题分布。

2. **HDP(Hierarchical Dirichlet Process)(这一部分对于概率论和随机过程的要求较高，理解上可能有一定的偏差)**

HDP，也称“层次狄利克雷过程”，相当于非参的LDA模型，这里的非参是指不需要指定主题数目，主题数目由模型自主学习产生，关键在于将产生主题的狄利克雷分布变成狄利克雷过程，将主题分布本身也做为一个变量，从更高的层次上描述主题分布。

参考资料：

1. Latent Dirichlet Allocation, May 2003, Journal of Machine Learning Research 3(4-5):993-1022, DOI: 10.1162/jmlr.2003.3.4-5.993
2. Sharing Clusters Among Related Groups:Hierarchical Dirichlet Processes
3. LDA-math - 文本建模
4. 文本主题模型之LDA(一) LDA基础
5. 文本主题模型之LDA(二) LDA求解之Gibbs采样算法
6. 文本主题模型之LDA(三) LDA求解之变分推断EM算法
7. 主题模型TopicModel：Unigram、LSA、PLSA模型
8. 主题模型TopicModel：隐含狄利克雷分布LDA
9. Dirichlet Process 和 Hierarchical Dirichlet Process
10. 分层Dirichlet过程（HDP）的理解
11. 层次狄利克雷过程（Hierarchical Dirichlet Processes）