**LDA变分推断**

武汉大学计算机学院 朱佳晖

zhujiahui@outlook.com

# LDA符号

：语料文本集合，即若干个文档

：文档（词汇的向量）

：潜在主题的个数

：词汇表中的单词个数

：某个文档中的词汇数

：文档-主题分布。

：某一个文档的主题分配序列。可以用一个三元组（三维矩阵）来表示。表示：第d个文档中的第n个词所对应的主题是否是（主题表中）第i个主题。取值只能为0或1

：文档-主题分布的Dirichlet先验参数，

：主题-词汇分布的先验参数，

：的变分参数，属于Dirichlet先验参数，

：的变分参数，属于多项式分布先验参数。可以用一个三元组（三维矩阵）来表示。表示第d个文档中的第n个词在第i个主题中出现的概率。

总体目标：根据文档以及先验参数、，估计和。

注意：先验参数在这个过程中也是不断被更新的。

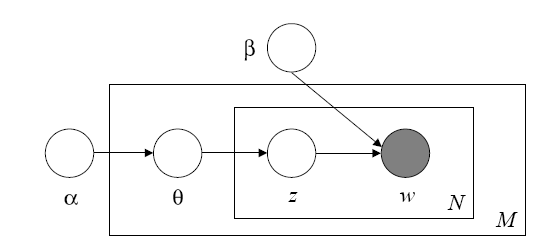


图1

除了文档的词汇向量，以及先验参数，，图中的其余的各个变量都是不可观测的。

# k维Dirichlet分布概率密度函数



主题分布服从参数为的Dirichlet分布。

# 重要的概率公式

* 主题分布，文档主题分配序列，以及某个文档的联合（生成）概率为：



结合图1，可知表示的生成概率，表示第n个词汇（对应的位置上）的主题分配的生成概率，而表示最终第n个词汇的生成概率。即的过程。

* 某个文档的生成概率为：



该公式推导自公式(2)，文档的生成概率即为(2)中的联合概率的关于的边缘概率(marginal distribution)。那么，只需将公式(2)依次关于和求积分即可。且由于主题分配序列是一个离散的量，因此积分便是求和。

* 则主题分布和文档主题分配序列的后验概率(posterior distribution)为：



本质就是贝叶斯公式：



求关于待估计参数的后验概率往往是进行参数估计的第一步，在EM等一系列算法中皆是如此。此类方法便是极大后验估计(MAP)法。因此，需要构造出主题分布和文档主题分配序列的后验概率。

# 参数估计哪家强

主题分布和文档主题分配序列的后验概率已经构造完毕，可以采用相应的参数估计方法估计和了。

在数理统计课本上，一般的参数估计方法主要有：点估计、矩估计、极大似然估计等等。在其他的一些资料上，或多或少还会接触到EM(Expectation Maximization)算法，其原理便是极大后验估计。即求解使得后验概率达到最大值（极大值）情况下的值作为待估计参数的最佳估计值。

但是，在这里，情况不是很妙。因为里面含有积分的形式，导致求解起来很麻烦，是不可手算的。由于无法直接用微积分的技巧求得精确解（解析解），因此不妨求个数值解。

求解此类难解问题的一般方法一般有那么几类：

* 随机算法
* 近似算法
* 有限元分析算法
* ……

M. H. Alsuwaiyel版的《算法设计技巧与分析》在第14章和第15章分别介绍了随机算法和近似算法。随机算法，顾名思义，就是随机采样，当精确地求解总体比较困难的时候，可以通过随机采样的方法，通过样本的相关数字特征求解总体的相关数字特征。

在LDA中，随机采样的算法主要有：

* 马尔可夫蒙特卡罗采样(MCMC)
* 吉布斯采样(Gibbs Sampling)

其中Gibbs Sampling是广为运用的一种，因为操作简便，可编程性强，小规模数据上具有较高的效率。因此，不少基于LDA的改进模型都是采用Gibbs Sampling方法求解。

当然，Gibbs Sampling也有一些缺点：

* 不太适合实时的在线的训练（当然这个问题也不突出）
* 为了保证准确率，采样的样本量往往比较巨大

在2003原版论文中，Blei他们是采用类似于近似的算法求解的。这便是变分推断。变分是泛函中的概念，所谓泛函，泛泛地来说是指函数的函数（积分变量是函数的函数）。变分法用来求解使得泛函达到极值情况的“函数”。

在此处的参数估计中，由于后验概率表达式十分复杂，因此可以考虑用一个近似的函数q去逼近它，但是这个函数q本身也是未知的。但是有一个要点，那就是q的形式要比原始的表达式简单地多，或者说是我们常见的概率分布函数（比如指数族函数、正态分布族函数等等），否则近似就没有太大的意义。“近似”肯定是为了简化，为了“好算”。

# 变分推断



以上的后验概率表达式中，难求的主要是。因此，核心就是解决的近似求解问题。

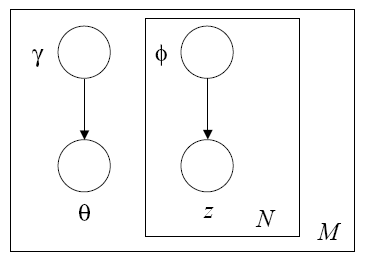


图2

图2是将图1简化了的模型，即变分推断下模型。图2将图1的节点删去，并将、“改写”（当然这里不仅仅是改写，也发生了本质的变化）为、，然后将依赖于。这样简化的好处是：消除了原先的和之间的耦合度，便于计算相关概率。此处的和成为变分参数。

于是乎，我们可以用以下变分分布去近似原表达式：



这里，需要转换一下固有的观念：和一样都是概率分布，只是换了一个字母而已。因为我们平时已经习惯了，所以有时候对有点不知所措。

此处，为三元组，表示第d个文档中的第n个词在第i个主题中出现的概率。

## 5.1 找待求概率的下界

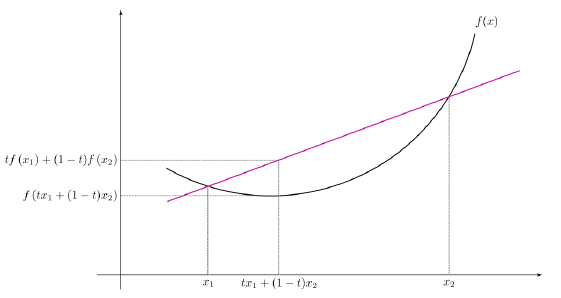
即求的一个下界。注意：这里求的是它的对数下界。因为概率表达式中含有指数的形式，所以一般可以求对数下界。

首先是詹森不等式(Jensens inequality)：对于随机变量X，和一个函数f。

如果f是凸函数，则有

如果f是凹函数，则有

画在图上是这样的：





其中



是的熵(entropy)，熵的概念详见信息论。

如此，我们便找到了的一个下界。将该下界记为，即



可以很容易地证明，和它的下界之间相差了。即



其中，表示两个概率分布之间的KL散度（相对熵）。KL散度一般用来度量两个概率分布之间的“距离”。一般地分布a和分布b之间的KL散度定义为



注意：KL散度是不对称的，即，跟一般意义上的“距离”不太一样。对于连续性随机变量及其分布，求和要改写为积分形式。

那下面就来证明公式(9)

 得证。□

注意：

* 第2行套KL散度的公式的时候，因为是连续的，所以是求积分，而是离散的，所以是求和。这里有2个随机变量。
* 第4行的结论运用了公式(4)。
* 第6行关于的展开运用了公式(2)。
* 第7行。公式(7)为了简便，中省略了2个变分参数作为条件。
* 第7行到第8行是因为有。这是因为单纯的与变分分布及其变分参数无关。当然，和与及其变分参数是相关的。

称为的证据下界(evidence lower bound)，简称为ELBO。因此，最大化就等同于最大化。

## 5.2 指数族函数

为了进一步求解，需要将其进行展开，表达成我们熟悉的可计算的式子。

在解决这个问题之前，先引入指数族(exponential family)分布一些知识，后面在求解过程中可能会用到。一般地，指数族分布的形式如下：



其中：

* 是自然参数(natural parameter)
* 基本量函数(underlying measure)
* 是充分统计量函数(sufficient statistic)
* 是对数正则化因子。



关于有一个性质，即



也就是说，对数正则化因子的一阶导数就是充分统计量的期望。

由公式(1)可得



将其和指数族函数对应，可得，对于，其充分统计量函数为，对数正则化因子。依据公式(11)，有



其中是对数函数的一阶导。

## 5.3 ELBO展开

已经有ELBO的表达式，下面就将其进行分解。



关于第一项



其中第3行到第四行运用了公式(12)。注：这里省略了参数，即没有写成条件概率的形式。

关于第二项



其中

* 表示当前文档中的第n个位置上的词的主题分配。是二值元素，表示当前第n个位置上是第i个主题。表示不是第i个主题。
* 第2行是分解展开。
* 第2行到第3行，是因为有。即当前第n个位置上是第i个主题的概率值体现在文档-主题分布的第i个位置上的概率值。
* 第4行到第5行，是因为有。表示第n个词在第i个主题中出现的概率。

关于第三项



其中第2行到第3行用到。表示第i个主题（主题-词汇分布）中词汇的概率值。当时，。

关于第四项



其中第4行到第5行用到了公式(13)和公式(15)的结论。

综合以上，对于ELBO，它的展开形式如下



## 5.4 变分参数的更新

将ELBO展开成可计算的式子之后即可进行参数的学习。这里有两类参数需要更新：变分参数和；模型先验参数和。

首先更新变分参数和。提取ELBO(公式17)中的与相关的式子。并加入拉格朗日乘子构造拉格朗日函数。



表示第n个词在第i个主题中出现的概率。那么该词在所有主题中出现的概率之和必然为1，即，因此有该等式约束。为拉格朗日乘子。

因此



令该导函数为0，可得



即



表示主要只与和相关。

然后提取ELBO(公式17)中的与相关的式子。并加入拉格朗日乘子构造拉格朗日函数。



注意，此处没有约束条件。

因此



注意：求导的过程中变为了。

令该导函数为0，可得



即



## 5.5先验参数的更新

同变分参数的处理形式类似。提取ELBO(公式17)中的与相关的式子。并加入拉格朗日乘子构造拉格朗日函数，这里需要注意的是，需要跟具体的文档结合起来，因此需要考虑文档这一维



为拉格朗日乘子。表示第d个文档中的第n个词在第i个主题中出现的概率。

因此



是一个符号函数，当满足括号里的条件时取1，不满足条件时取0。这里可忽略的影响，然后令导函数为0，可得



提取ELBO(公式17)中的与相关的式子。并加入拉格朗日乘子构造拉格朗日函数



因此



此导函数存在关于嵌套形式。即当前的依赖于另一系列，且。因此可以用迭代的方法求解极值点处的。一般可采用Newton-Paphson方法。

Hessian矩阵的求解如下：



迭代的形式如下：

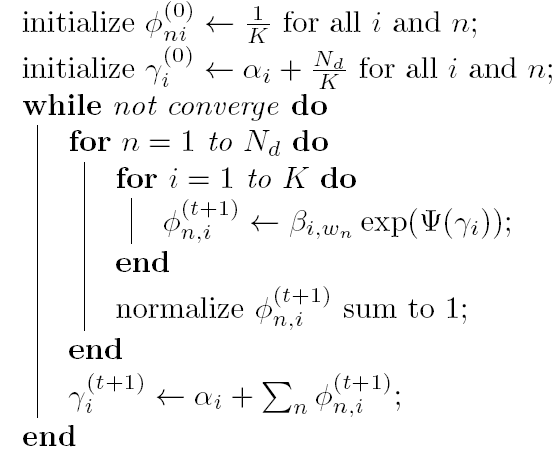


其中是Hessian矩阵，为处的梯度。

# LDA变分EM参数更新

5.4和5.5节已经给出了变分参数和的更新准则和先验参数和的更新准则。下面，可采用EM的更新迭代算法依次更新这一系列参数。

**E步**：根据先验参数和，极大化ELBO(公式17)，以找到较优的变分参数和：



**M步**：将E步得到的变分参数带入ELBO(公式17)，极大化ELBO，以此求得极值情况下的和，即更新和。

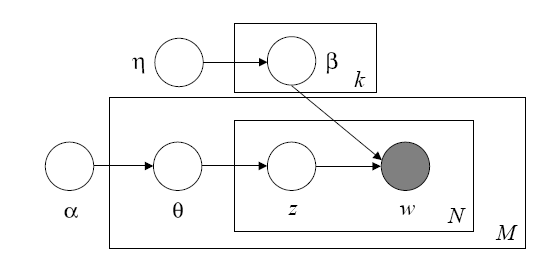
重复E步M步，直至收敛，将得到和，和的最终估计值。

# LDA平滑

由于文档的词汇列表往往很长，使得文档-词汇矩阵较为稀疏，当前新文档中的词汇可能在之前的训练文本中均没有出现过，导致估计出的文档的生成概率为0。因此需要进行相应的平滑处理，在LDA中一般采用拉普拉斯平滑(Laplace smoothing)。

在平滑过程中，我们将主题-词汇分布的多项式分布过程也向文档-主题的多项式分布一样，加入Dirichlet先验参数。即原先的集分布和参数为一体的进行功能拆分：分布还是原来的多项式分布，只不过变成了一个由Dirichlet先验参数决定的分布。这个Dirichlet分布的先验参数为。

由此，文档-主题，主题-词汇这两个分布均变成了带有Dirichlet先验参数的多项式分布，每一个可称之为Dirichlet-多项式共轭分布。



平滑之后的变分形式变成如下：



M为文档的总数，为的变分参数。注意，原来是，现在平滑之后为了与保持相同的地位，也变成了矩阵的粗体形式。

可以推断，的更新规则为：



表示第d个文档的第n个词是否属在第i个主题的第j个位置上出现。