LDA主题模型

信息学院-黄浩

2022年3月14日星期一

LDA知识图谱

- 一个函数：gamma函数
- 四个分布：二项分布、多项分布、beta分布、Dirichlet分布
- 一个概念和一个理念：共轭先验和贝叶斯框架
- 两个模型：pLSA、LDA
- 一个采样：Gibbs采样

LDA知识图谱

- LDA由Blei, David M.、Ng, Andrew Y.、Jordan于2003年提出，是一种主题模型，它可以将文档集 中每篇文档的主题以概率分布的形式给出，从而通过分析一些文档抽取出它们的主题（分布）出来后，便可以根据主题（分布）进行主题聚类或文本分类。同时，它是一种典型的词袋模型，即一篇文档是由一组词构成，词与词之间没有先后顺序的关系。
- 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。
- 人类是怎么生成文档的呢？LDA的这三位作者在原始论文中给了一个简单的例子。比如假设事先给定了这几个主题：Arts、Budgets、Children、Education，然后通过学习训练，获取每个主题Topic对应的词语。

LDA知识图谱

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

LDA知识图谱

- 然后以一定的概率选取上述某个主题，再以一定的概率选取那个主题下的某个单词，不断的重复这两步，最终生成如下图所示的一篇文章，其中不同颜色的词语分别对应上图中不同主题下的词。

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

LDA知识图谱

- 当看到一篇文章后，往往喜欢推测这篇文章是如何生成的，我们可能会认为作者先确定这篇文章的几个主题，然后围绕这几个主题遣词造句，表达成文。
- LDA就是要干这事：根据给定的一篇文档，反推其主题分布。
- 通俗来说，可以假定认为人类是根据上述文档生成过程写成了各种各样的文章，现在某小撮人想让计算机利用LDA干一件事：计算机推测分析网络上各篇文章分别都写了那些主题，且各篇文章中各个主题出现的概率大小（主题分布）是多少。

LDA知识图谱

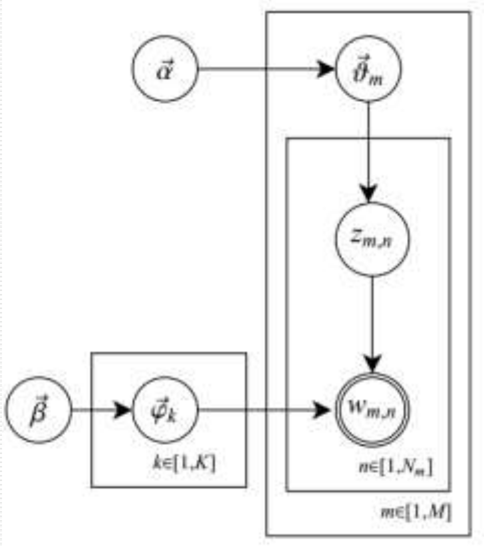
- 就是这么一个看似普通的LDA，一度吓退了不少想深入探究其内部原理的初学者。难在哪呢，难就难在LDA内部涉及到的数学知识点太多了。

在LDA模型中，一篇文档生成的方式如下：

- 从狄利克雷分布α中取样生成文档 i 的主题分布θ_i
- 从主题的多项式分布θ_i中取样生成文档i第 j 个词的主题$z_{i,j}$
- 从狄利克雷分布β中取样生成主题$z_{i,j}$对应的词语分布$\phi_{z_{i,j}}$
- 从词语的多项式分布$\phi_{z_{i,j}}$中采样最终生成词语$w_{i,j}$

LDA知识图谱

Beta分布是二项式分布的共轭先验概率分布，而狄利克雷分布（Dirichlet分布）是多项式分布的共轭先验概率分布。
此外，LDA的图模型结构如下图所示（类似贝叶斯网络结构）



分布与共轭

- 先简单解释下二项分布、多项分布、beta分布、Dirichlet 分布这4个分布，以及共轭。

- **二项分布（Binomial distribution）。**

二项分布是从伯努利分布推进的。伯努利分布，又称两点分布或0-1分布，是一个离散型的随机分布，其中的随机变量只有两类取值，非正即负{+，-}。而二项分布即重复n次的伯努利试验，记为 $X \sim b(n,p)$。简言之，只做一次实验，是伯努利分布，重复做了n次，是二项分布。二项分布的概率密度函数为：

$$P(K=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

分布与共轭

- 先简单解释下二项分布、多项分布、beta分布、Dirichlet 分布这4个分布，以及共轭。

对于k = 0, 1, 2, ..., n，其中的$\binom{n}{k} = \frac{n!}{k!(n-k)!}$是二项式系数（这就是二项分布的名称的由来），又记为$C(n,k)$。

分布与共轭

- 多项分布是指单次试验中的随机变量的取值不再是0-1的，而是有多种离散值可能（1,2,3...,k）。比如投掷6个面的骰子实验，N次实验结果服从K=6的多项分布。其中 $\sum_{i=1}^{k} p_i = 1, p_i > 0$
- 多项分布的概率密度函数为：

$$P(x_1, x_2, ..., x_k; n, p_1, p_2, ..., p_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

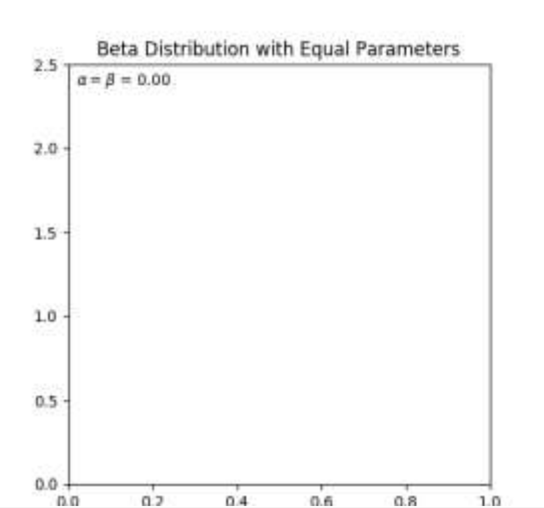
分布与共轭

- **Beta**分布，二项分布的共轭先验分布。

给定参数$\alpha > 0$和$\beta > 0$，取值范围为[0,1]的随机变量 x 的概率密度函数:

$$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \qquad \frac{1}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \qquad \Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$$

- 注：$\Gamma(x)$ 便是所谓的Gamma函数



分布与共轭

- **Dirichlet分布，是beta分布在高维度上的推广。**
- Dirichlet分布的的密度函数形式跟beta分布的密度函数如出一辙：

$$f(x_1,x_2,...,x_k;\alpha_1,\alpha_2,...,\alpha_k)=\frac{1}{B(\alpha)}\prod_{i=1}^{k}x_i^{\alpha^i-1} \qquad B(\alpha)=\frac{\prod_{i=1}^{k}\Gamma(\alpha^i)}{\Gamma(\sum_{i=1}^{k}\alpha^i)},\sum x_i=1$$

分布与共轭

- **Gamma函数**

1. **问题1** 随机变量 $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} Uniform(0,1)$

2. 把这n 个随机变量排序后得到顺序统计量 $X_{(1)}, X_{(2)} \cdots, X_{(n)}$

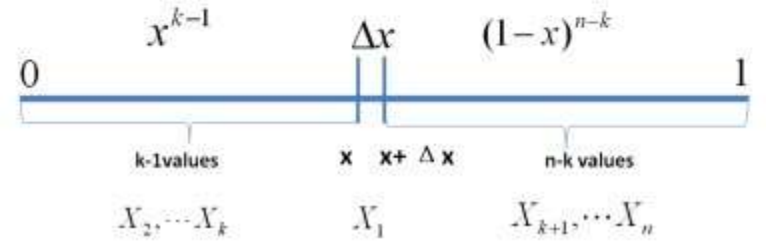
3. 然后请问 $X_{(k)}$ 的分布是什么。

- 为解决这个问题，可以尝试计算落在区间[x,x+Δx]的概率。即求下述式子的值：

$$P(x \le X_{(k)} \le x + \Delta x) = ?$$

分布与共轭

- 首先，把 **[0,1]** 区间分成三段 **[0,x)**，**[x,x+Δx]**，**(x+Δx,1]**，然后考虑下简单的情形：即假设**n** 个数中只有**1**个落在了区间 **[x,x+Δx]**内，由于这个区间内的数**X(k)**是第**k**大的，所以**[0,x)**中应该有 **k−1** 个数，**(x+Δx,1]** 这个区间中应该有**n−k** 个数。



分布与共轭

- 从而问题转换为下述事件E：

$$
\begin{aligned}
E = \{\ & X_1 \in [x, x+\Delta x], \\
& X_i \in [0, x) \quad (i = 2, \cdots, k), \\
& X_j \in (x+\Delta x, 1] \quad (j = k+1, \cdots, n)\}
\end{aligned}
$$

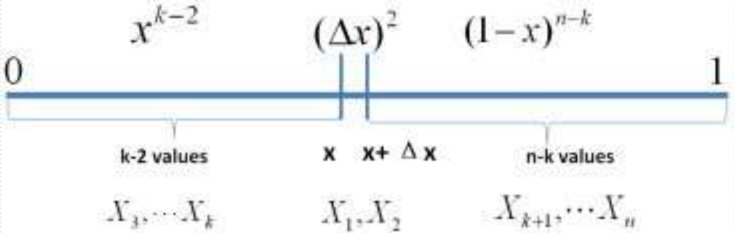
- 对于上述事件E，有：

$$
\begin{aligned}
P(E) &= \prod_{i=1}^{n} P(X_i) \\
&= x^{k-1}(1-x-\Delta x)^{n-k}\Delta x \\
&= x^{k-1}(1-x)^{n-k}\Delta x + o(\Delta x)
\end{aligned}
$$

分布与共轭

其中，o(Δx)表示Δx的高阶无穷小。显然，由于不同的排列组合，即n个数中有一个落在 [x,x+Δx]区间的有n种取法，余下n−1个数中有k−1个落在[0,x)的有$\binom{n-1}{k-1}$种组合，所以和事件E等价的事件一共有$n\binom{n-1}{k-1}$个。

如果有2个数落在区间[x,x+Δx]呢？如下图所示：



分布与共轭

- 类似于事件E，对于2个数落在区间[x,x+Δx]的事件E′：

$$
\begin{aligned}
E' = \{\, & X_1, X_2 \in [x, x+\Delta x], \\
& X_i \in [0, x) \quad (i = 3, \cdots, k), \\
& X_j \in (x+\Delta x, 1] \quad (j = k+1, \cdots, n)\}
\end{aligned}
$$

- 对于上述事件E′，有：

$$P(E') = x^{k-2}(1-x-\Delta x)^{n-k}(\Delta x)^2 = o(\Delta x)$$

分布与共轭

- 从上述的事件E、事件E'中，可以看出，只要落在[x,x+Δx]内的数字超过一个，则对应的事件的概率就是 o(Δx)。于是乎有：

$$
\begin{aligned}
&P(x \le X_{(k)} \le x + \Delta x) \\
&= n\binom{n-1}{k-1}P(E) + o(\Delta x) \\
&= n\binom{n-1}{k-1}x^{k-1}(1-x)^{n-k}\Delta x + o(\Delta x)
\end{aligned}
$$

分布与共轭

从而得到$X_{(k)}$的概率密度函数$f(x)$为：

$$
\begin{aligned}
f(x) &= \lim_{\Delta x \to 0} \frac{P(x \le X_{(k)} \le x + \Delta x)}{\Delta x} \\
&= n \binom{n-1}{k-1} x^{k-1} (1-x)^{n-k} \\
&= \frac{n!}{(k-1)!(n-k)!} x^{k-1} (1-x)^{n-k} \quad x \in [0,1]
\end{aligned}
$$

至此，本节开头提出的问题得到解决。然仔细观察$X_{(k)}$的概率密度函数，发现式子的最终结果有阶乘，联想到阶乘在实数上的推广$\Gamma(x)$函数：

$$
\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt
$$

分布与共轭

两者结合是否会产生奇妙的效果呢？考虑到$\Gamma(x)$具有如下性质：

$$\Gamma(n) = (n-1)!$$

故将代入到$X_{(k)}$的概率密度函数$f(x)$中，可得：

$$f(x) = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} x^{k-1}(1-x)^{n-k}$$

然后取$\alpha = k$，$\beta = n-k+1$，转换$f(x)$得到：

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

分布与共轭

- **beta分布**

在概率论中，beta是指一组定义在$(0,1)$区间的连续概率分布，有两个参数α和β，且$\alpha,\beta>0$。

beta分布的概率密度函数是：

$$f(x;\alpha,\beta)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du}$$
$$=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$
$$=\frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

$$\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt$$

随机变量X服从参数为的beta分布通常写作：$X\sim\mathrm{Be}(\alpha,\beta)$。

分布与共轭

- $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} Uniform(0,1)$，对应的顺序统计量是$X_{(1)}, X_{(2)} \cdots, X_{(n)}$，需要猜测$p = X_{(k)}$；
- $Y_1, Y_2, \cdots, Y_m \overset{\text{iid}}{\sim} Uniform(0,1)$，$Y_i$中有$m_1$个比p小，$m_2$个比$p$大；
- 那么，请问$P(p|Y_1, Y_2, \cdots, Y_m)$的分布是什么。

根据"Yi中有m_1个比p小，m_2个比p大"，换言之，Yi中有m_1个比$X_{(k)}$小，m_2个比$X_{(k)}$大，所以$X_{(k)}$是$X_1, X_2, \cdots, X_n, Y_1, Y_2, \cdots, Y_m \overset{\text{iid}}{\sim} Uniform(0,1)$中第$k + m_1$大的数。

根据得到的结论"只要落在[x,x+Δx]内的数字超过一个，则对应的事件的概率就是o(Δx)"，继而推出事件服从beta分布，从而可知$p = X_{(k)}$的概率密度函数为：

$$Beta(p|k + m_1, n - k + 1 + m_2)$$

分布与共轭

1. 为了猜测$p = X_{(k)}$，在获得一定的观测数据前，我们对P的认知是：
$f(p) = Beta(p|k, n-k+1)$，此称为P的先验分布；

2. 然后为了获得这个结果"Y_i中有m_1个比p小，m_2个比P大"，针对Y_i是做了m次贝努利实验，所以m_1服从二项分布$B(m, p)$；

3. 在给定了来自数据提供的(m_1, m_2)的知识后，P的后验分布变为
$f(p|m_1, m_2) = Beta(p|k+m_1, n-k+1+m_2)$。

回顾下贝叶斯派思考问题的固定模式：

- **先验分布**$\pi(\theta)$ **+ 样本信息**$X \Rightarrow$ **后验分布**$\pi(\theta|x)$

上述思考模式意味着，新观察到的样本信息将修正人们以前对事物的认知。换言之，在得到新的样本信息之前，人们对θ的认知是先验分布$\pi(\theta)$，在得到新的样本信息X后，人们对θ的认知为$\pi(\theta|x)$。

分布与共轭

- 类比到现在这个问题上，可以得到：

$$Beta(p|k, n-k+1) + Count(m_1, m_2) = Beta(p|k+m_1, n-k+1+m_2)$$

其中(m_1, m_2)对应的是二项分布$B(m_1+m_2, p)$的计数。

更一般的，对于非负实数α和β，我们有如下关系

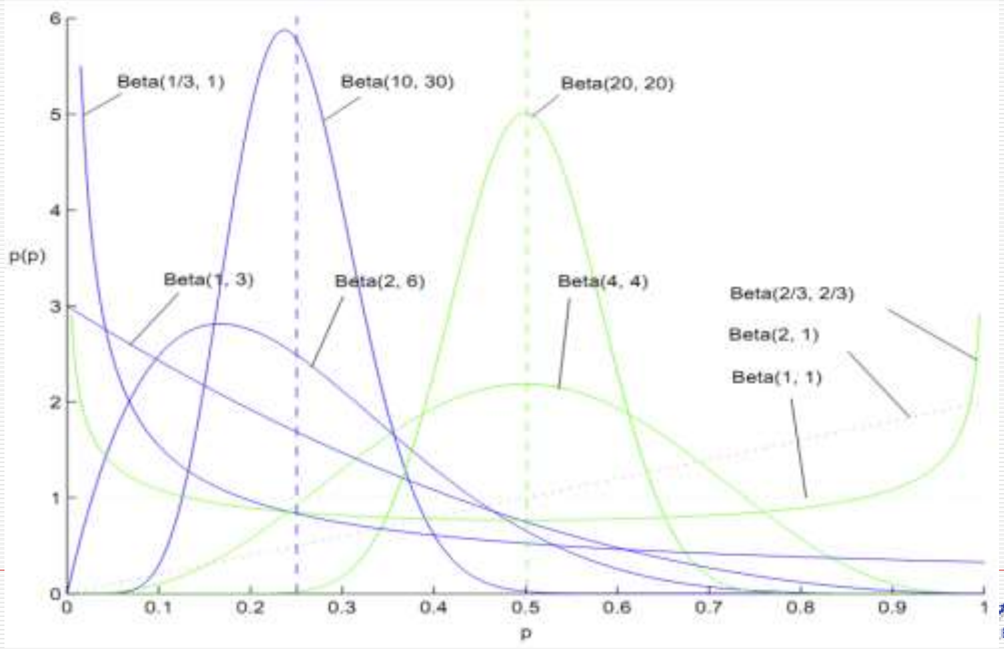
$$Beta(p|\alpha, \beta) + Count(m_1, m_2) = Beta(p|\alpha+m_1, \beta+m_2)$$

分布与共轭

- 针对于这种观测到的数据符合二项分布，参数的先验分布和后验分布都是Beta分布的情况，就是Beta-Binomial共轭。换言之，Beta分布是二项式分布的共轭先验概率分布。
- 二项分布和Beta分布是共轭分布意味着，如果我们为二项分布的参数p选取的先验分布是Beta分布，那么以p为参数的二项分布用贝叶斯估计得到的后验分布仍然服从Beta分布。

分布与共轭

- 此外，如何理解参数和所表达的意义呢？两个参数可以认为形状参数，通俗但不严格的理解是，和共同控制Beta分布的函数“长的样子”：形状千奇百怪，高低胖瘦，如下图所示：



分布与共轭

- 共轭先验分布
- 如何理解共轭：
- 轭的意思是束缚、控制，共轭从字面上理解，则是共同约束，或互相约束。
- 在贝叶斯概率理论中，如果后验概率**P(θ|x)**和先验概率**p(θ)**满足同样的分布律，那么，先验分布和后验分布被叫做共轭分布，同时，先验分布叫做似然函数的共轭先验分布。
- 比如，某观测数据服从概率分布**P(θ)**时，当观测到新的**X**数据时，我们一般会遇到如下问题：
- 可否根据新观测数据**X**，更新参数**θ**？
- 根据新观测数据可以在多大程度上改变参数**θ**，即

$$\theta \leftarrow \theta + \Delta\theta$$

分布与共轭

- 根据根据贝叶斯公式可知：

$$P(\theta|x)=\frac{P(x|\theta)\cdot P(\theta)}{P(x)}\propto P(x|\theta)\cdot P(\theta)$$

- 其中，**P(x|θ)**表示以预估**θ**为参数的**x**概率分布，可以直接求得，**P(θ)**是已有原始的**θ**概率分布。
- 如果选取**P(x|θ)**的共轭先验作为**P(θ)**的分布，那么**P(x|θ)**乘以**P(θ)**，然后归一化的结果**P(θ|x)**跟和**P(θ)**的形式一样。换句话说，先验分布是**P(θ)**，后验分布是**P(θ|x)**，先验分布跟后验分布同属于一个分布族，故称该分布族是**θ**的共轭先验分布（族）。

分布与共轭

- 投掷一个非均匀硬币，可以使用参数为θ的伯努利模型，θ为硬币为正面的概率，那么结果x的分布形式为：

$$P(x|\theta)=\theta^{x}\cdot(1-\theta)^{1-x}$$

其共轭先验为beta分布，具有两个参数α和β，称为超参数（hyperparameters）。且这两个参数决定了θ参数，其Beta分布形式为

$$P(\theta|\alpha,\beta)=\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta}$$

分布与共轭

- 然后计算后验概率：

$$
\begin{aligned}
& P(\theta|x) \\
& \propto P(x|\theta) \cdot P(\theta) \\
& \propto \left(\theta^{x}(1-\theta)^{x}\right)\left(\theta^{\alpha-1}(1-\theta)^{\beta-1}\right) \\
& = \theta^{x+a-1}(1-\theta)^{1-x+\beta-1}
\end{aligned}
$$

- 归一化这个等式后会得到另一个Beta分布，从而证明了Beta分布确实是伯努利分布的共轭先验分布。

分布与共轭

- 从**beta**分布推广到**Dirichlet** 分布

接下来，咱们来考察beta分布的一个性质。

如果$p \sim Beta(t|\alpha,\beta)$，则有:

$$
\begin{aligned}
E(p) &= \int_0^1 t * Beta(t|\alpha,\beta)dt \\
&= \int_0^1 t * \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha-1}(1-t)^{\beta-1}dt \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 t^{\alpha}(1-t)^{\beta-1}dt
\end{aligned}
$$

分布与共轭

- 注意到上式最后结果的右边积分：

$$\int_0^1 t^{\alpha}(1-t)^{\beta-1}dt$$

其类似于概率分布$Beta(t|\alpha+1,\beta)$，而对于这个分布有

$$\int_0^1 \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)} t^{\alpha}(1-t)^{\beta-1}dt = 1$$

- 从而求得

$$\int_0^1 t^{\alpha}(1-t)^{\beta-1}dt \Longrightarrow \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}$$

分布与共轭

最后将此结果带入$E(p)$的计算式，得到：

$$\begin{aligned}E(p) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}\\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+1)}\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)}\\ &= \frac{\alpha}{\alpha+\beta}\end{aligned}$$

最后的这个结果意味着对于Beta 分布的随机变量，其均值（期望）可以用$\frac{\alpha}{\alpha+\beta}$来估计。此外，狄利克雷Dirichlet 分布也有类似的结论，即如果$\vec{p} \sim Dir(\vec{t}\,|\vec{\alpha})$，同样可以证明有下述结论成立：

$$E(\vec{p}) = \left(\frac{\alpha_1}{\sum_{i=1}^K \alpha_i}, \frac{\alpha_2}{\sum_{i=1}^K \alpha_i}, \cdots, \frac{\alpha_K}{\sum_{i=1}^K \alpha_i}\right)$$

分布与共轭

- **Dirichlet 分布**
- **简单的理解Dirichlet 分布就是一组连续多变量概率分布，是多变量普遍化的beta分布。为了纪念德国数学家约翰·彼得·古斯塔夫·勒热纳·狄利克雷（Peter Gustav Lejeune Dirichlet）而命名。狄利克雷分布常作为贝叶斯统计的先验概率。**
- 根据wikipedia上的介绍，维度K ≥ 2（x1,x2…xK-1维，共K个）的狄利克雷分布在参数α1, ..., αK > 0上、基于欧几里得空间RK-1里的勒贝格测度存在概率密度函数，定义为：

$$f(x_1,\ldots,x_{K-1};\alpha_1,\ldots,\alpha_K) = \frac{1}{\mathrm{B}(\alpha)}\prod_{i=1}^{K} x_i^{\alpha_i-1}$$

分布与共轭

其中，$\mathrm{B}(\alpha)$相当于是多项beta函数

$$\mathrm{B}(\alpha) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}$$

且 $\alpha = (\alpha_1, \ldots, \alpha_K)$

此外，x1+x2+…+xK-1+xK=1，x1,x2…xK-1>0，且在(K-1)维的单纯形上，其他区域的概率密度为0。

当然，也可以如下定义Dirichlet 分布

$$p(\vec{p}|\vec{\alpha}) = \mathrm{Dir}(\vec{p}|\vec{\alpha})$$

$$\triangleq \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}p_k^{\alpha_k-1}$$

$$\triangleq \frac{1}{\Delta(\vec{\alpha})}\prod_{k=1}^{K}p_k^{\alpha_k-1},$$

分布与共轭

其中的$\Delta(\vec{\alpha})$称为Dirichlet 分布的归一化系数：

$$\Delta(\vec{\alpha}) = \frac{\prod_{k=1}^{\dim \vec{\alpha}} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{\dim \vec{\alpha}} \alpha_k\right)} = int \prod_{k=1}^{V} p_k^{\alpha_k - 1} \mathrm{d}\vec{p}$$

且根据Dirichlet分布的积分为1（概率的基本性质），可以得到：

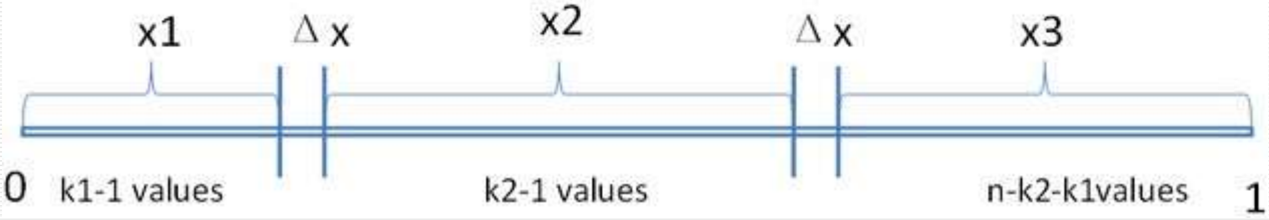
$$\int_{\vec{p}} \prod_{k=1}^{K} p_k^{\alpha_k - 1} d\vec{p} = \Delta(\vec{\alpha})$$

分布与共轭

- **Dirichlet-Multinomial** 共轭：

- $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} Uniform(0,1)$,
- 排序后对应的顺序统计量$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$,
- 问$(X_{(k_1)}, X_{(k_1+k_2)})$的联合分布是什么?

为了简化计算，取x3满足x1+x2+x3=1,但只有x1,x2是变量，如下图所示:



分布与共轭

- 从而得到：

$$
\begin{aligned}
&P\Big(X_{(k_1)} \in (x_1, x_1+\Delta x), X_{(k_1+k_2)} \in (x_2, x_2+\Delta x)\Big) \\
&= n(n-1)\binom{n-2}{k_1-1, k_2-1} x_1^{k_1-1} x_2^{k_2-1} x_3^{n-k_1-k_2} (\Delta x)^2 \\
&= \frac{n!}{(k_1-1)!(k_2-1)!(n-k_1-k_2)!} x_1^{k_1-1} x_2^{k_2-1} x_3^{n-k_1-k_2} (\Delta x)^2
\end{aligned}
$$

继而得到于是我们得到$\big(X_{(k_1)}, X_{(k_1+k_2)}\big)$的联合分布为:

$$
\begin{aligned}
f(x_1, x_2, x_3) &= \frac{n!}{(k_1-1)!(k_2-1)!(n-k_1-k_2)!} x_1^{k_1-1} x_2^{k_2-1} x_3^{n-k_1-k_2} \\
&= \frac{\Gamma(n+1)}{\Gamma(k_1)\Gamma(k_2)\Gamma(n-k_1-k_2+1)} x_1^{k_1-1} x_2^{k_2-1} x_3^{n-k_1-k_2}
\end{aligned}
$$

分布与共轭

- 观察上述式子的最终结果，可以看出上面这个分布其实就是3维形式的 Dirichlet 分布

$$Dir(x_1,x_2,x_3|k_1,k_2,n-k_1-k_2+1)$$

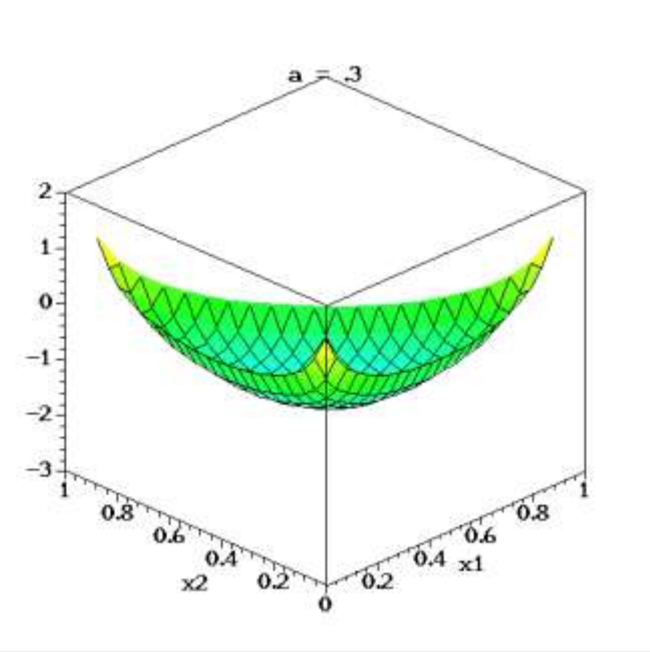
令$\alpha_1=k_1,\alpha_2=k_2,\alpha_3=n-k_1-k_2+1$，于是分布密度可以写为

$$f(x_1,x_2,x_3)=\frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}x_1^{\alpha_1-1}x_2^{\alpha_2-1}x_3^{\alpha_3-1}$$

分布与共轭

这个就是一般形式的3维 Dirichlet 分布，即便 $\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ 延拓到非负实数集合，以上概率分布也是良定义的。

将Dirichlet分布的概率密度函数取对数，绘制对称Dirichlet分布的图像如下图所示（截取自wikipedia上）：



分布与共轭

为了方便讨论，记 $\vec{m}=(m_1,m_2,m_3)$，及 $\vec{k}=(k_1,k_2,n-k_1-k_2+1)$，根据已知条件 $Y_1,Y_2,\cdots,Y_m \overset{iid}{\sim} Uniform(0,1)$，Yi中落到 $[0,p_1)$，$[p_1,p_2)$，$[p_2,1]$ 三个区间的个数分别为 m1,m2"，可得 p_1、p_2 分别是这m+n个数中第 k_1+m_1 大、第 k_2+m_2 大的数。于是，后验分布 $P(\vec{p}|Y_1,Y_2,\cdots,Y_m)$ 应该为 $Dir(\vec{p}\,|k_1+m_1,k_1+m_2,n-k_1-k_2+1+m_3)$，即一般化的形式表示为：$Dir(\vec{p}\,|\,\vec{k}+\vec{m})$。

同样的，按照贝叶斯推理的逻辑，可将上述过程整理如下：

1. 我们要猜测参数 $\vec{p}=(p_1,p_2,p_3)$，其先验分布为 $Dir(\vec{p}|\vec{k})$；
2. 数据Yi落到三个区间 $[0,p_1)$，$[p_1,p_2)$，$[p_2,1]$ 的个数分别为 m_1,m_2,m_3，所以 $\vec{m}=(m_1,m_2,m_3)$ 服从多项分布 $Mult(\vec{m}|\vec{p})$
3. 在给定了来自数据提供的知识 $\vec{m}$ 后，$\vec{p}$ 的后验分布变为 $Dir(\vec{p}|\vec{k}+\vec{m})$

上述贝叶斯分析过程的直观表述为：

$$Dir(\vec{p}\,|\,\vec{k})+MultCount(\vec{m})=Dir(\vec{p}\,|\,\vec{k}+\vec{m})$$

分布与共轭

令$\vec{\alpha}=\vec{k}$，可把$\vec{\alpha}$从整数集合延拓到实数集合，从而得到更一般的表达式如下：

$$Dir(\overrightarrow{p}\,|\,\overrightarrow{\alpha}) + MultCount(\overrightarrow{m}) = Dir(p|\,\overrightarrow{\alpha} + \overrightarrow{m})$$

- 针对于这种观测到的数据符合多项分布，参数的先验分布和后验分布都是**Dirichlet** 分布的情况，就是**Dirichlet-Multinomial** 共轭。换言之，至此已经证明了**Dirichlet**分布的确就是多项式分布的共轭先验概率分布。
- 意味着，如果我们为多项分布的参数**p**选取的先验分布是**Dirichlet**分布，那么以**p**为参数的多项分布用贝叶斯估计得到的后验分布仍然服从**Dirichlet**分布。
- 进一步，一般形式的**Dirichlet** 分布定义如下：

$$Dir(\overrightarrow{p}\,|\,\overrightarrow{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}p_k^{\alpha_k-1}$$

分布与共轭

而对于给定的$\vec{p}$和N，其多项分布为：

$$Mult(\overrightarrow{n}\,|\,\overrightarrow{p},N) = \binom{N}{\overrightarrow{n}} \prod_{k=1}^{K} p_k^{n_k}$$

结论是：Dirichlet分布$Dir(\vec{p}|\vec{\alpha})$和多项分布$Mult(\vec{n}|\vec{p},N)$是共轭关系。

主题模型LDA

总结下我们目前所得到的最主要的几个收获：

我们知道beta分布是二项式分布的共轭先验概率分布：

- “对于非负实数α和β，我们有如下关系

$$Beta(p|\alpha,\beta)+Count(m_1,m_2)=Beta(p|\alpha+m_1,\beta+m_2)$$

其中(m_1,m_2)对应的是二项分布$B(m_1+m_2,p)$的计数。针对于这种观测到的数据符合二项分布，参数的先验分布和后验分布都是Beta分布的情况，就是Beta-Binomial 共轭。”

主题模型LDA

我们知道狄利克雷分布（Dirichlet分布）是多项式分布的共轭先验概率分布：

- “把$\vec{\alpha}$从整数集合延拓到实数集合，从而得到更一般的表达式如下：

$$Dir(\vec{p}\mid\vec{\alpha}) + MultCount(\vec{m}) = Dir(p\mid\vec{\alpha}+\vec{m})$$

针对于这种观测到的数据符合多项分布，参数的先验分布和后验分布都是Dirichlet 分布的情况，就是 Dirichlet-Multinomial 共轭。”

- 以及贝叶斯派思考问题的固定模式：
  - **先验分布**$\pi(\theta)$ **+ 样本信息**$X \Rightarrow$ **后验分布**$\pi(\theta|x)$

主题模型LDA

上述思考模式意味着，新观察到的样本信息将修正人们以前对事物的认知。换言之，在得到新的样本信息之前，人们对θ的认知是先验分布$\pi(\theta)$，在得到新的样本信息X后，人们对θ的认知为$\pi(\theta|x)$。

- 顺便提下频率派与贝叶斯派各自不同的思考方式：
  - 频率派把需要推断的参数θ看做是固定的未知常数，即概率θ虽然是未知的，但最起码是确定的一个值，同时，样本X 是随机的，所以频率派重点研究样本空间，大部分的概率计算都是针对样本X 的分布；
  - 而**贝叶斯派**的观点则截然相反，他们认为待估计的参数θ是随机变量，服从一定的分布，而样本X 是固定的，由于样本是固定的，所以他们重点研究的是参数θ的分布。

主题模型LDA

- LDA模型之前，再循序渐进理解基础模型：Unigram model、mixture of unigrams model，以及跟LDA最为接近的pLSA模型。

为了方便描述，首先定义一些变量：

- w表示词，V表示所有单词的个数（固定值）
- z表示主题，k是主题的个数（预先给定，固定值）
- $D=(\mathbf{w}_1,\cdots,\mathbf{w}_M)$表示语料库，其中的$M$是语料库中的文档数（固定值）
- $\mathbf{w}=(w_1,w_2,\cdots,w_N)$表示文档，其中的$N$表示一个文档中的词数（随机变量）

主题模型LDA

各个基础模型

Unigram model

对于文档$\mathbf{w} = (w_1, w_2, \cdots, w_N)$，用$p(w_n)$表示词$w_n$的先验概率，生成文档$\mathbf{w}$的概率为：

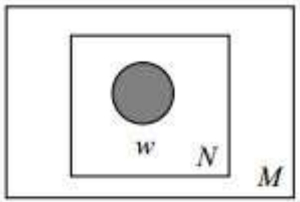
$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n)$$

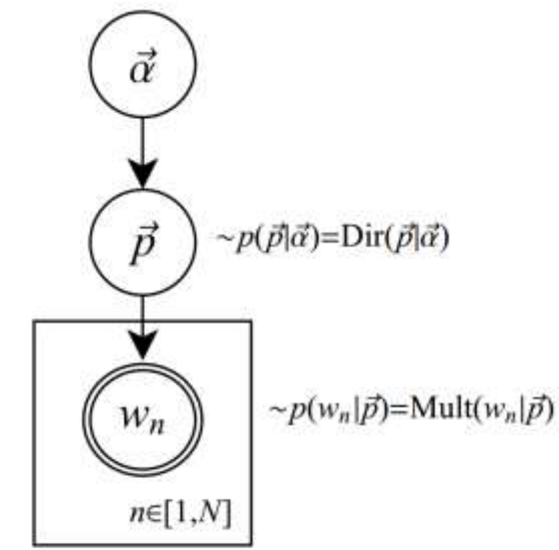
主题模型LDA

各个基础模型

Unigram model

其图模型为（图中被涂色的w表示可观测变量，N表示一篇文档中总共N个单词，M表示M篇文档）：





主题模型LDA

unigram model假设文本中的词服从Multinomial分布，而我们已经知道Multinomial分布的先验分布为Dirichlet分布。

上图中的w_n表示在文本中观察到的第n个词，n∈[1,N]表示该文本中一共有N个单词。加上方框表示重复，即一共有N个这样的随机变量w_n。其中，p和α是隐含未知变量：

- p是词服从的Multinomial分布的参数
- α是Dirichlet分布（即Multinomial分布的先验分布）的参数。

一般α由经验事先给定，p由观察到的文本中出现的词学习得到，表示文本中出现每个词的概率。

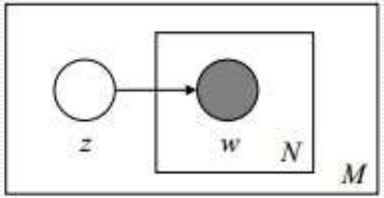
主题模型LDA

Mixture of unigrams model

该模型的生成过程是：给某个文档先选择一个主题 z，再根据该主题生成文档，该文档中的所有词都来自一个主题。假设主题有 $z_1, \ldots, z_k$，生成文档$\mathbf{w}$的概率为：

$$p(\mathbf{w}) = p(z_1)\prod_{n=1}^{N} p(w_n|z_1) + \cdots + p(z_k)\prod_{n=1}^{N} p(w_n|z_k) = \sum_{z} p(z)\prod_{n=1}^{N} p(w_n|z)$$

其图模型为（图中被涂色的w表示可观测变量，未被涂色的z表示未知的隐变量，N表示一篇文档中总共N个单词，M表示M篇文档）：

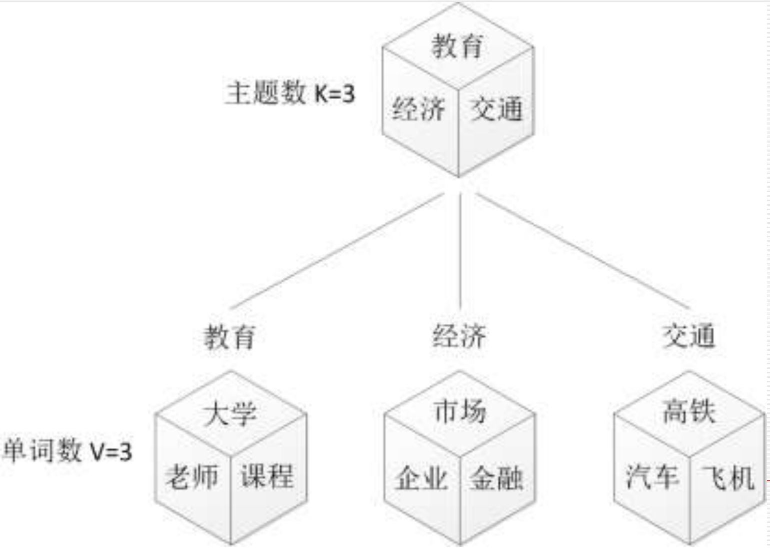


主题模型LDA

- **PLSA模型**
- pLSA模型下生成文档
- 在上面的Mixture of unigrams model中，我们假定一篇文档只有一个主题生成，可实际中，一篇文章往往有多个主题，只是这多个主题各自在文档中出现的概率大小不一样。比如介绍一个国家的文档中，往往会分别从教育、经济、交通等多个主题进行介绍。
- 假设你要写M篇文档，由于一篇文档由各个不同的词组成，所以你需要确定每篇文档里每个位置上的词。
- 再假定你一共有K个可选的主题，有V个可选的词，咱们来玩一个扔骰子的游戏。

主题模型LDA

- 1. 假设你每写一篇文档会制作一颗K面的“文档-主题”骰子（扔此骰子能得到K个主题中的任意一个），和K个V面的“主题-词项” 骰子（每个骰子对应一个主题，K个骰子对应之前的K个主题，且骰子的每一面对应要选择的词项，V个面对应着V个可选的词）。
  - 比如可令K=3，即制作1个含有3个主题的“文档-主题”骰子，这3个主题可以是：教育、经济、交通。然后令V = 3，制作3个有着3面的“主题-词项”骰子，其中，教育主题骰子的3个面上的词可以是：大学、老师、课程，经济主题骰子的3个面上的词可以是：市场、企业、金融，交通主题骰子的3个面上的词可以是：高铁、汽车、飞机。



主题模型LDA

- 2. 每写一个词，先扔该"文档-主题"骰子选择主题，得到主题的结果后，使用和主题结果对应的那颗"主题-词项"骰子，扔该骰子选择要写的词。
  - 先扔"文档-主题"的骰子，假设（以一定的概率）得到的主题是教育，所以下一步便是扔教育主题筛子，（以一定的概率）得到教育主题筛子对应的某个词：大学。
    - 上面这个投骰子产生词的过程简化下便是："先以一定的概率选取主题，再以一定的概率选取词"。事实上，一开始可供选择的主题有3个：教育、经济、交通，那为何偏偏选取教育这个主题呢？其实是随机选取的，只是这个随机遵循一定的概率分布。比如可能选取教育主题的概率是0.5，选取经济主题的概率是0.3，选取交通主题的概率是0.2，那么**这3个主题的概率分布便是{教育：0.5，经济：0.3，交通：0.2}**，我们把各个主题z在文档d中出现的概率分布称之为主题分布，且是一个多项分布。
    - 同样的，从主题分布中随机抽取出教育主题后，依然面对着3个词：大学、老师、课程，这3个词都可能被选中，但它们被选中的概率也是不一样的。比如大学这个词被选中的概率是0.5，老师这个词被选中的概率是0.3，课程被选中的概率是0.2，那么**这3个词的概率分布便是{大学：0.5，老师：0.3，课程：0.2}**，我们把各个词语w在主题z下出现的概率分布称之为词分布，这个词分布也是一个多项分布。
    - 所以，**选主题和选词都是两个随机的过程，先从主题分布{教育：0.5，经济：0.3，交通：0.2}中抽取出主题：教育，然后从该教育主题对应的词分布{大学：0.5，老师：0.3，课程：0.2}中抽取出词：大学。**
- 3. 最后，你不停的重复扔"文档-主题"骰子和"主题-词项"骰子，重复N次（产生N个词），完成一篇文档，重复这产生一篇文档的方法M次，则完成M篇文档。

主题模型LDA

上述过程抽象出来即是PLSA的文档生成模型。在这个过程中，我们并未关注词和词之间的出现顺序，所以pLSA是一种词袋方法。具体说来，该模型假设一组共现(co-occurrence)词项关联着一个隐含的主题类别$z_k \in \{z_1, \cdots, z_K\}$。同时定义：

- $P(d_i)$表示海量文档中某篇文档被选中的概率。
- $P(w_j|d_i)$表示词w_j在给定文档d_i中出现的概率。
    - 怎么计算得到呢？针对海量文档，对所有文档进行分词后，得到一个词汇列表，这样每篇文档就是一个词语的集合。对于每个词语，用它在文档中出现的次数除以文档中词语总的数目便是它在文档中出现的概率$P(w_j|d_i)$。
- $P(z_k|d_i)$表示具体某个主题z_k在给定文档d_i下出现的概率。
- $P(w_j|z_k)$表示具体某个词w_j在给定主题z_k下出现的概率，与主题关系越密切的词，其条件概率$P(w_j|z_k)$越大。

利用上述的第1、3、4个概率，我们便可以按照如下的步骤得到“文档-词项”的生成模型：

1. 按照概率$P(d_i)$选择一篇文档d_i
2. 选定文档d_i后，从主题分布中按照概率$P(z_k|d_i)$选择一个隐含的主题类别z_k
3. 选定z_k后，从词分布中按照概率$P(w_j|z_k)$选择一个词w_j

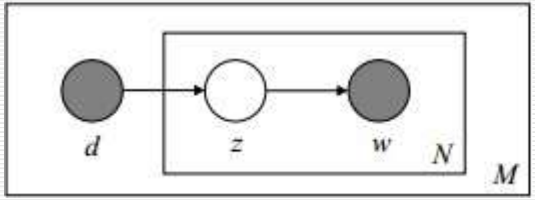
所以pLSA中生成文档的整个过程便是选定文档生成主题，确定主题生成词。

主题模型LDA

根据文档反推其主题分布

- 反过来，既然文档已经产生，那么如何根据已经产生好的文档反推其主题呢？这个利用看到的文档推断其隐藏的主题（分布）的过程（其实也就是产生文档的逆过程），便是主题建模的目的：自动地发现文档集中的主题（分布）。
- 换言之，人类根据文档生成模型写成了各类文章，然后丢给了计算机，相当于计算机看到的是一篇篇已经写好的文章。现在计算机需要根据一篇篇文章中看到的一系列词归纳出当篇文章的主题，进而得出各个主题各自不同的出现概率：主题分布。即文档**d**和单词**w**是可被观察到的，但主题**z**却是隐藏的。
- 如下图所示（图中被涂色的**d**、**w**表示可观测变量，未被涂色的**z**表示未知的隐变量，**N**表示一篇文档中总共**N**个单词，**M**表示**M**篇文档）：

主题模型LDA



上图中，文档d和词w是我们得到的样本（样本随机，参数虽未知但固定，所以pLSA属于频率派思想。区别于下文要介绍的LDA中：样本固定，参数未知但不固定，是个随机变量，服从一定的分布，所以LDA属于贝叶斯派思想），可观测得到，所以对于任意一篇文档，其$P(w_j|d_i)$是已知的。

从而可以根据大量已知的文档-词项信息$P(w_j|d_i)$，训练出文档-主题$P(z_k|d_i)$和主题-词项$P(w_j|z_k)$，如下公式所示:

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

主题模型LDA

故得到文档中每个词的生成概率为：

$$
\begin{aligned}
P(d_i, w_j) &= P(d_i)P(w_j|d_i) \\
&= P(d_i)\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)
\end{aligned}
$$

由于$P(d_i)$可事先计算求出，**而$P(w_j|z_k)$和$P(z_k|d_i)$未知，所以$\theta = (P(w_j|z_k), P(z_k|d_i))$就是我们要估计的参数（值）**，通俗点说，就是要最大化这个θ。

用什么方法进行估计呢，常用的参数估计方法有极大似然估计MLE、最大后验证估计MAP、贝叶斯估计等等。因为该待估计的参数中含有隐变量z，所以我们可以考虑EM算法。

主题模型LDA

EM算法的简介：

EM算法，全称为Expectation-maximization algorithm，为期望最大算法，其基本思想是：首先随机选取一个值去初始化待估计的值$\theta^{(0)}$，然后不断迭代寻找更优的$\theta^{(n+1)}$使得其似然函数likelihood $L(\theta^{(n+1)})$比原来的$L(\theta^{(n)})$要大。换言之，假定现在得到了$\theta^{(n)}$，想求$\theta^{(n+1)}$，使得

$$\theta^{(n+1)} = \max_{\theta} L(\theta) - L(\theta^{(n)})$$

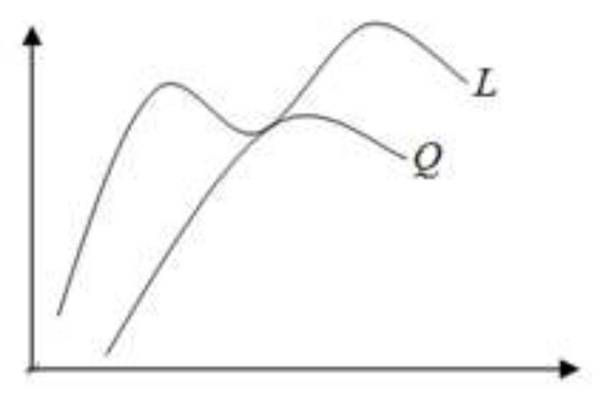
EM的关键便是要找到$L(\theta)$的一个下界$Q(\theta;\theta^{n})$（注：$L(\theta) = \log p(X|\theta)$，其中，X表示已经观察到的随机变量），然后不断最大化这个下界，通过不断求解下界$Q(\theta;\theta^{n})$的极大化，从而逼近要求解的似然函数$L(\theta)$。

所以EM算法的一般步骤为：

- 1. 随机选取或者根据先验知识初始化$\theta^{(0)}$；
- 2. 不断迭代下述两步
  - ①给出当前的参数估计$\theta^{(n)}$，计算似然函数$L(\theta)$的下界$Q(\theta;\theta^{n})$
  - ②重新估计参数θ，即求$\theta^{(n+1)}$，使得$\theta^{(n+1)} = \arg_{\theta}\max Q(\theta;\theta^{n})$
- 3. 上述第二步后，如果$L(\theta)$收敛（即$Q(\theta;\theta^{n})$收敛）则退出算法，否则继续回到第二步。

主题模型LDA

上述过程好比在二维平面上，有两条不相交的曲线，一条曲线在上（简称上曲线L），一条曲线在下（简称下曲线Q），下曲线为上曲线的下界。现在对上曲线未知，只已知下曲线，为了求解上曲线的最高点，我们试着不断增大下曲线，使得下曲线不断逼近上曲线，下曲线在某一个点达到局部最大值并与上曲线在这点的值相等，记录下这个值，然后继续增大下曲线，寻找下曲线上与上曲线上相等的值，迭代到$L(\theta)$收敛（即$Q(\theta;\theta^n)$收敛）停止，从而利用当前下曲线上的局部最大值当作上曲线的全局最大值（换言之，EM算法不保证一定能找到全局最优值）。如下图所示：



主题模型LDA

假定有训练集 $\{x^{(1)},\ldots,x^{(m)}\}$，包含m个独立样本，希望从中找到该组数据的模型p(x,z)的参数。

然后通过极大似然估计建立目标函数--对数似然函数：

$$
\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{m} \log p(x;\theta) \\
&= \sum_{i=1}^{m} \log \sum_{z} p(x,z;\theta)
\end{aligned}
$$

主题模型LDA

这里，z是隐随机变量，直接找到参数的估计是很困难的。我们的策略是建立$\ell(\theta)$的下界，并且求该下界的最大值；重复这个过程，直到收敛到局部最大值。

令Qi是z的某一个分布，Qi≥0，且结合Jensen不等式，有：

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

主题模型LDA

为了寻找尽量紧的下界，我们可以让上述等号成立，而若要让等号成立，则条件是：

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

换言之，有以下式子成立：$Q_i(z^{(i)}) \propto p(x^{(i)}, z^{(i)}; \theta)$，且由于有：$\sum_z Q_i(z^{(i)}) = 1$

所以可得：

$$\begin{aligned} Q_i(z^{(i)}) &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{\sum_z p(x^{(i)}, z; \theta)} \\ &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{p(x^{(i)}; \theta)} \\ &= p(z^{(i)} | x^{(i)}; \theta) \end{aligned}$$

主题模型LDA

- 最终得到EM算法的整体框架如下：

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

(M-step) Set

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

}

主题模型LDA

- EM算法估计pLSA的两未知参数

首先尝试**从矩阵的角度来描述待估计的两个未知变量**$P(w_j|z_k)$**和**$P(z_k|d_i)$。

- 假定用ϕ_k表示词表$\mathcal{V}$在主题z_k上的一个多项分布，则ϕ_k可以表示成一个向量，每个元素$\phi_{k,j}$**表示词项**w_j**出现在主题**z_k**中的概率**，即

$$P(w_j|z_k) = \phi_{k,j}, \quad \sum_{w_j \in \mathcal{V}} \phi_{k,j} = 1$$

- 用θ_i表示所有主题$\mathcal{Z}$在文档d_i上的一个多项分布，则θ_i可以表示成一个向量，每个元素$\theta_{i,k}$**表示主题**z_k**出现在文档**d_i**中的概率**，即

$$P(z_k|d_i) = \theta_{i,k}, \quad \sum_{z_k \in \mathcal{Z}} \theta_{i,k} = 1$$

这样，**巧妙的把**$P(w_j|z_k)$**和**$P(z_k|d_i)$**转换成了两个矩阵**。换言之，最终我们要求解的参数是这两个矩阵：

$$\Phi = [\phi_1, \cdots, \phi_K], \quad z_k \in \mathcal{Z} \qquad \Theta = [\theta_i, \cdots, \theta_M], d_i \epsilon D$$

主题模型LDA

- 由于词和词之间是相互独立的，所以整篇文档N个词的分布为：

$$P(W|d_i) = \prod_{j=1}^{N} P(d_i, w_j)^{n(d_i, w_j)}$$

- 再由于文档和文档之间也是相互独立的，所以整个语料库中词的分布为（整个语料库M篇文档，每篇文档N个词）：

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

其中，$n(d_i, w_j)$表示词项w_j在文档d_i中的词频，$n(d_i)$表示文档di中词的总数，显然有$n(d_i) = \sum_{w_j \in V} n(d_i, w_j)$。从而得到整个语料库的词分布的对数似然函数

主题模型LDA

$$
\begin{aligned}
\ell(\Phi,\Theta) &= \sum_{i-1}^{M}\sum_{j-1}^{N} n(d_i,w_j)\log P(d_i,w_j) \\
&= \sum_{i-1}^{M} n(d_i)\left(\log P(d_i) + \sum_{j-1}^{N}\frac{n(d_i,w_j)}{n(d_i)}\log\sum_{k-1}^{K}P(w_j|z_k)P(z_k|d_i)\right) \\
&= \sum_{i-1}^{M} n(d_i)\left(\log P(d_i) + \sum_{j-1}^{N}\frac{n(d_i,w_j)}{n(d_i)}\log\sum_{k-1}^{K}\phi_{k,j}\theta_{i,k}\right)
\end{aligned}
$$

现在，我们需要最大化上述这个对数似然函数来求解参数$\phi_{k,j}$和$\theta_{i,k}$。对于这种含有隐变量的最大似然估计，可以使用EM算法。EM算法，分为两个步骤：先E-step，后M-step。

主题模型LDA

- **E-step**：假定参数已知，计算此时隐变量的后验概率。

利用贝叶斯法则，可以得到：

$$
\begin{aligned}
P(z_k|d_i,w_j) &= \frac{P(z_k,d_i,w_j)}{\sum_{l=1}^{K} P(z_l,d_i,w_j)} \\
&= \frac{P(w_j|d_i,z_k)P(z_k|d_i)P(d_i)}{\sum_{l=1}^{K} \left(P(w_j|d_i,z_l)P(z_l|d_i)P(d_i)\right)} \\
&= \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{l=1}^{K} P(w_j|z_l)P(z_l|d_i)} \\
&= \frac{\phi_{k,j}\theta_{i,k}}{\sum_{l=1}^{K} \phi_{l,j}\theta_{i,l}}
\end{aligned}
$$

主题模型LDA

- **M-step**：带入隐变量的后验概率，最大化样本分布的对数似然函数，求解相应的参数。

观察之前得到的对数似然函数$\ell(\Phi,\Theta)$的结果，由于文档长度$P(d_i)\propto n(d_i)$可以单独计算，所以去掉它不影响最大化似然函数。此外，根据E-step的计算结果，把 $\phi_{k,j}\theta_{i,k}=P(z_k|d_i,w_j)\sum_{l=1}^{K}\phi_{l,j}\theta_{i,l}$代入$\ell(\Phi,\Theta)$，于是我们只要最大化下面这个函数 ℓ 即可

$$\ell=\sum_{i=1}^{M}\sum_{j=1}^{N}n(d_i,w_j)\sum_{k=1}^{K}P(z_k|d_i,w_j)\log\left[\phi_{k,j}\theta_{i,k}\right]$$

- 这是一个多元函数求极值问题，并且已知有如下约束条件：

$$\sum_{j=1}^{N}\phi_{k,j}=1$$

$$\sum_{k=1}^{K}\theta_{i,k}=1$$

主题模型LDA

熟悉凸优化的应该知道，一般处理这种带有约束条件的极值问题，常用的方法便是拉格朗日乘数法，即通过引入拉格朗日乘子将约束条件和多元（目标）函数融合到一起，转化为无约束条件的极值问题。

这里我们引入两个拉格朗日乘子τ和ρ，从而写出拉格朗日函数

$$\mathcal{H} = \mathcal{L}^c + \sum_{k=1}^{K} \tau_k \left(1 - \sum_{j=1}^{N} \phi_{k,j}\right) + \sum_{i=1}^{M} \rho_i \left(1 - \sum_{k=1}^{K} \theta_{i,k}\right)$$

因为我们要求解的参数是$\phi_{k,j}$和$\theta_{i,k}$，所以分别对$\phi_{k,j}$和$\theta_{i,k}$求偏导，然后令偏导结果等于0，得到

$$\sum_{i=1}^{M} n(d_i, w_j) P(z_k | d_i, w_j) - \tau_k \phi_{k,j} = 0, \quad 1 \le j \le N, 1 \le k \le K$$

$$\sum_{j=1}^{N} n(d_i, w_j) P(z_k | d_i, w_j) - \rho_i \theta_{i,k} = 0, \quad 1 \le i \le M, 1 \le k \le K$$

主题模型LDA

消去拉格朗日乘子，最终可估计出参数$\phi_{k,j}$和$\theta_{i,k}$

$$\phi_{k,j} = \frac{\sum_{i=1}^{M} n(d_i, w_j) P(z_k|d_i, w_j)}{\sum_{m=1}^{N} \sum_{i=1}^{M} n(d_i, w_m) P(z_k|d_i, w_m)}$$

$$\theta_{i,k} = \frac{\sum_{j=1}^{N} n(d_i, w_j) P(z_k|d_i, w_j)}{n(d_i)}$$

综上，在pLSA中：

1. 由于$P(w_j|z_k)$和$P(z_k|d_i)$未知，所以我们用EM算法去估计$\theta = (P(w_j|z_k), P(z_k|d_i))$这个参数的值。
2. 而后，用$\phi_{k,j}$表示词项w_j出现在主题z_k中的概率，即$P(w_j|z_k) = \phi_{k,j}$，用$\theta_{i,k}$表示主题z_k出现在文档d_i中的概率，即$P(z_k|d_i) = \theta_{i,k}$，从而把$P(w_j|z_k)$转换成了“主题-词项”矩阵Φ（主题生成词），把$P(z_k|d_i)$转换成了“文档-主题”矩阵Θ（文档生成主题）。
3. 最终求解出$\phi_{k,j}$、$\theta_{i,k}$。

主题模型LDA

LDA模型

事实上，理解了pLSA模型，也就差不多快理解了LDA模型，因为LDA就是在pLSA的基础上加层贝叶斯框架，即LDA就是pLSA的贝叶斯版本（正因为LDA被贝叶斯化了，所以才需要考虑历史先验知识，才加的两个先验参数）。

主题模型LDA

在pLSA模型中，我们按照如下的步骤得到“文档-词项”的生成模型：

1. 按照概率$P(d_i)$选择一篇文档d_i
2. 选定文档d_i后，确定文章的主题分布
3. 从主题分布中按照概率$P(z_k|d_i)$选择一个隐含的主题类别z_k
4. 选定z_k后，确定主题下的词分布
5. 从词分布中按照概率$P(w_j|z_k)$选择一个词w_j ”

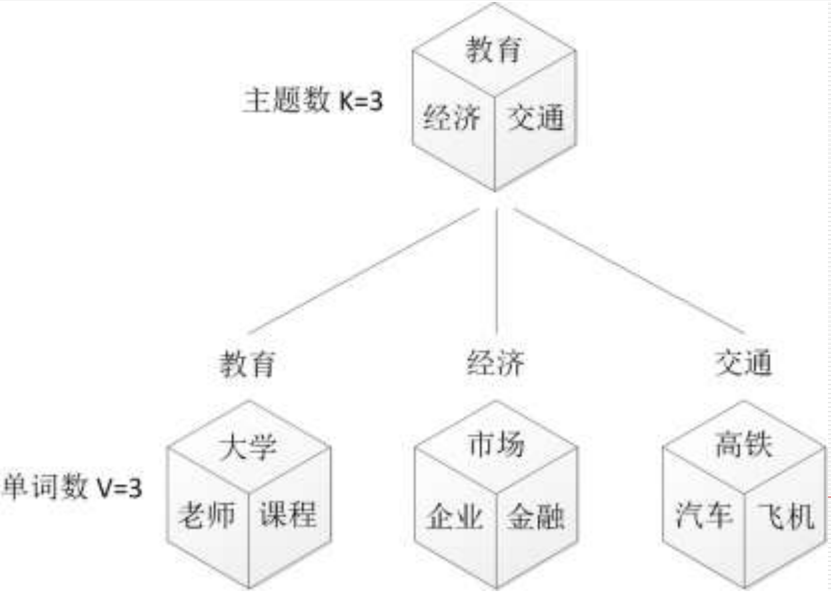
下面，咱们对比下本文开头所述的LDA模型中一篇文档生成的方式是怎样的：

1. 按照先验概率$P(d_i)$选择一篇文档d_i
2. 从狄利克雷分布（即Dirichlet分布）α中取样生成文档d_i的主题分布θ_i，换言之，主题分布θ_i由超参数为α的Dirichlet分布生成
3. 从主题的多项式分布θ_i中取样生成文档d_i第 j 个词的主题$z_{i,j}$
4. 从狄利克雷分布（即Dirichlet分布）β中取样生成主题$z_{i,j}$对应的词语分布$\phi_{z_{i,j}}$，换言之，词语分布$\phi_{z_{i,j}}$由参数为β的Dirichlet分布生成
5. 从词语的多项式分布$\phi_{z_{i,j}}$中采样最终生成词语$w_{i,j}$ ”

从上面两个过程可以看出，LDA在PLSA的基础上，为主题分布和词分布分别加了两个Dirichlet先验。

主题模型LDA

继续拿之前讲解PLSA的例子进行具体说明。如前所述，在PLSA中，选主题和选词都是两个随机的过程，先从主题分布{教育：0.5，经济：0.3，交通：0.2}中抽取出主题：教育，然后从该主题对应的词分布{大学：0.5，老师：0.3，课程：0.2}中抽取出词：大学。



主题模型LDA

而在LDA中，选主题和选词依然都是两个随机的过程，依然可能是先从主题分布{教育：0.5，经济：0.3，交通：0.2}中抽取出主题：教育，然后再从该主题对应的词分布{大学：0.5，老师：0.3，课程：0.2}中抽取出词：大学。

PLSA跟LDA的区别在于：

- PLSA中，主题分布和词分布是唯一确定的，能明确的指出主题分布可能就是{教育：0.5，经济：0.3，交通：0.2}，词分布可能就是{大学：0.5，老师：0.3，课程：0.2}。
- 但在LDA中，主题分布和词分布不再唯一确定不变，即无法确切给出。例如主题分布可能是{教育：0.5，经济：0.3，交通：0.2}，也可能是{教育：0.6，经济：0.2，交通：0.2}，到底是哪个我们不再确定（即不知道），因为它是随机的可变化的。但再怎么变化，也依然服从一定的分布，即主题分布跟词分布由Dirichlet先验随机确定。

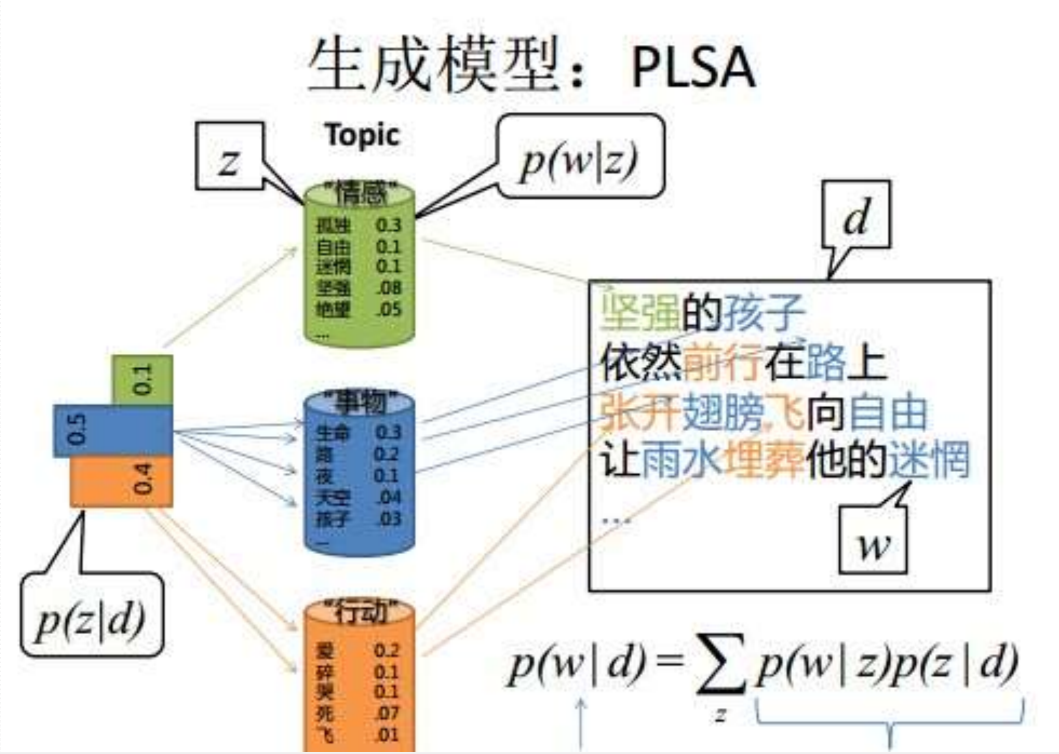
主题模型LDA

看到这，需要理解这样一个问题，面对多个主题或词，各个主题或词被抽中的概率不一样，所以抽取主题或词是随机抽取，还好理解。但现在主题分布和词分布本身也都是不确定的，这是如何理解？这是因为Blei等人“强行”给PLSA安了个贝叶斯框架呢，正因为LDA是PLSA的贝叶斯版本，所以主题分布跟词分布本身由先验知识随机给定。

进一步，会发现：

- **pLSA中**，主题分布和词分布确定后，以一定的概率（$P(z_k|d_i)$、$P(w_j|z_k)$）分别选取具体的主题和词项，生成好文档。而后根据生成好的文档反推其主题分布、词分布时，最终用EM算法（极大似然估计思想）求解出了两个未知但固定的参数的值：$\phi_{k,j}$（由$P(w_j|z_k)$转换而来）和$\theta_{i,k}$（由$P(z_k|d_i)$转换而来）。
  - 文档d产生主题z的概率，主题z产生单词w的概率都是两个固定的值。
    - 举个文档d产生主题z的例子。给定一篇文档d，主题分布是一定的，比如{ P(zi|d), i = 1,2,3 }可能就是{0.4,0.5,0.1}，表示z1、z2、z3，这3个主题被文档d选中的概率都是个固定的值：P(z1|d) = 0.4、P(z2|d) = 0.5、P(z3|d) = 0.1，如下图所示

主题模型LDA

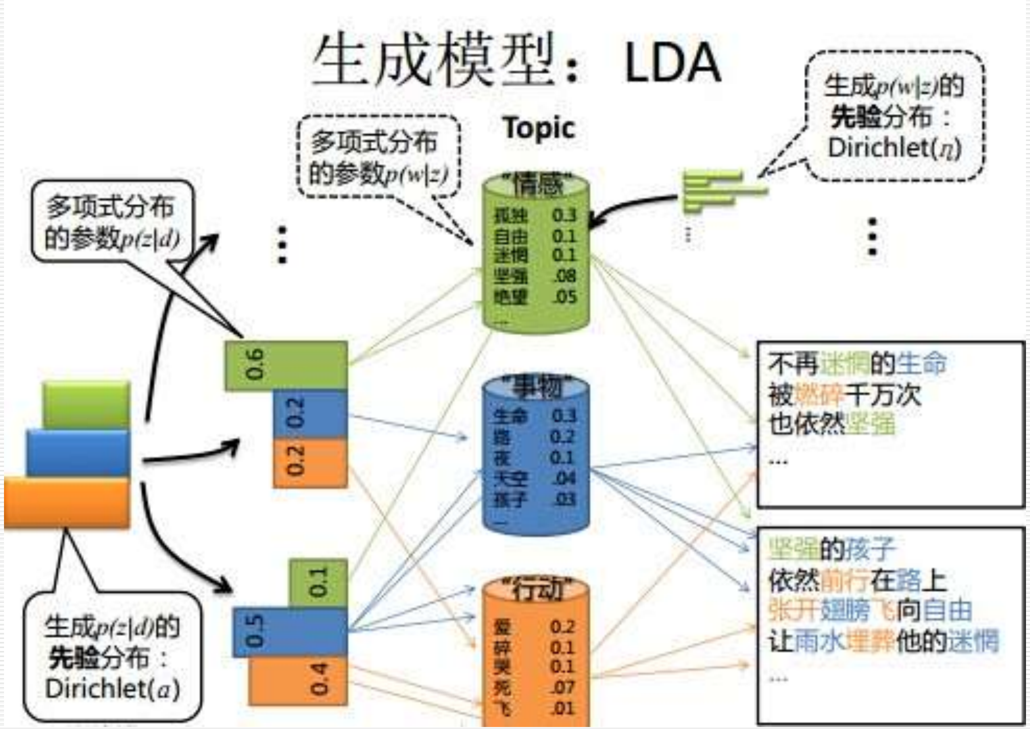


生成模型：PLSA
Topic
z
p(w|z)
"情感"
孤独 0.3
自由 0.1
迷惘 0.1
坚强 .08
绝望 .05
"事物"
生命 0.3
路 0.2
夜 0.1
天空 .04
孩子 .03
"行动"
爱 0.2
碎 0.1
哭 0.1
死 .07
飞 .01
0.1
0.5
0.4
p(z|d)
d
坚强的孩子
依然前行在路上
张开翅膀飞向自由
让雨水埋葬他的迷惘
...
w
$p(w|d)=\sum_z p(w|z)p(z|d)$

主题模型LDA

- 但在贝叶斯框架下的**LDA中**，我们**不再**认为主题分布（各个主题在文档中出现的概率分布）和词分布（各个词语在某个主题下出现的概率分布）是唯一确定的（而是随机变量），而是有很多种可能。但一篇文档总得对应一个主题分布和一个词分布吧，怎么办呢？LDA为它们弄了两个Dirichlet先验参数，这个Dirichlet先验为某篇文档随机抽取出某个主题分布和词分布。
  - 文档d产生主题z（准确的说，其实是Dirichlet先验为文档d生成主题分布Θ，然后根据主题分布Θ产生主题z）的概率，主题z产生单词w的概率都不再是某两个确定的值，而是随机变量。
    - 还是再次举下文档d具体产生主题z的例子。给定一篇文档d，现在有多个主题z1、z2、z3，它们的主题分布{ P(zi|d), i = 1,2,3 }可能是{0.4,0.5,0.1}，也可能是{0.2,0.2,0.6}，即这些主题被d选中的概率都不再认为是确定的值，可能是P(z1|d) = 0.4、P(z2|d) = 0.5、P(z3|d) = 0.1，也有可能是P(z1|d) = 0.2、P(z2|d) = 0.2、P(z3|d) = 0.6等等，而主题分布到底是哪个取值集合我们不确定（为什么？这就是贝叶斯派的核心思想，把未知参数当作是随机变量，不再认为是某一个确定的值），但其先验分布是dirichlet 分布，所以可以从无穷多个主题分布中按照dirichlet 先验随机抽取出某个主题分布出来。如下图所示

主题模型LDA



生成模型：LDA
Topic
多项式分布的参数p(w|z)
多项式分布的参数p(z|d)
生成p(w|z)的先验分布：Dirichlet(η)
生成p(z|d)的先验分布：Dirichlet(α)
"情感"
孤独 0.3
自由 0.1
迷惘 0.1
坚强 .08
绝望 .05
"事物"
生命 0.3
路 0.2
夜 0.1
天空 .04
孩子 .03
"行动"
爱 0.2
碎 0.1
哭 0.1
死 .07
飞 .01
0.6
0.2
0.2
0.1
0.5
0.4
不再迷惘的生命
被燃碎千万次
也依然坚强
...
坚强的孩子
依然前行在路上
张开翅膀飞向自由
让雨水埋葬他的迷惘
...

主题模型LDA

换言之，LDA在pLSA的基础上给这两参数（$P(z_k|d_i)$、$P(w_j|z_k)$）加了两个先验分布的参数（贝叶斯化）：一个主题分布的先验分布Dirichlet分布α，和一个词语分布的先验分布Dirichlet分布β。

综上，LDA真的只是pLSA的贝叶斯版本，文档生成后，两者都要根据文档去推断其主题分布和词语分布（即两者本质都是为了估计给定文档生成主题，给定主题生成词语的概率），只是用的参数推断方法不同，在pLSA中用极大似然估计的思想去推断两未知的固定参数，而LDA则把这两参数弄成随机变量，且加入dirichlet先验。

所以，pLSA跟LDA的本质区别就在于它们去估计未知参数所采用的思想不同，前者用的是频率派思想，后者用的是贝叶斯派思想。

好比，我去一朋友家：

- 按照频率派的思想，我估计他在家的概率是1/2，不在家的概率也是1/2，是个定值。
- 而按照贝叶斯派的思想，他在家不在家的概率不再认为是个定值1/2，而是随机变量。比如按照我们的经验（比如当天周末），猜测他在家的概率是0.6，但这个0.6不是说就是完全确定的，也有可能是0.7。如此，贝叶斯派没法确切给出参数的确定值（0.3,0.4，0.6,0.7，0.8,0.9都有可能），但至少明白在哪个范围或哪些取值（0.6,0.7，0.8,0.9）更有可能，哪个范围或哪些取值（0.3,0.4）不太可能。进一步，贝叶斯估计中，参数的多个估计值服从一定的先验分布，而后根据实践获得的数据（例如周末不断跑他家），不断修正之前的参数估计，从先验分布慢慢过渡到后验分布。

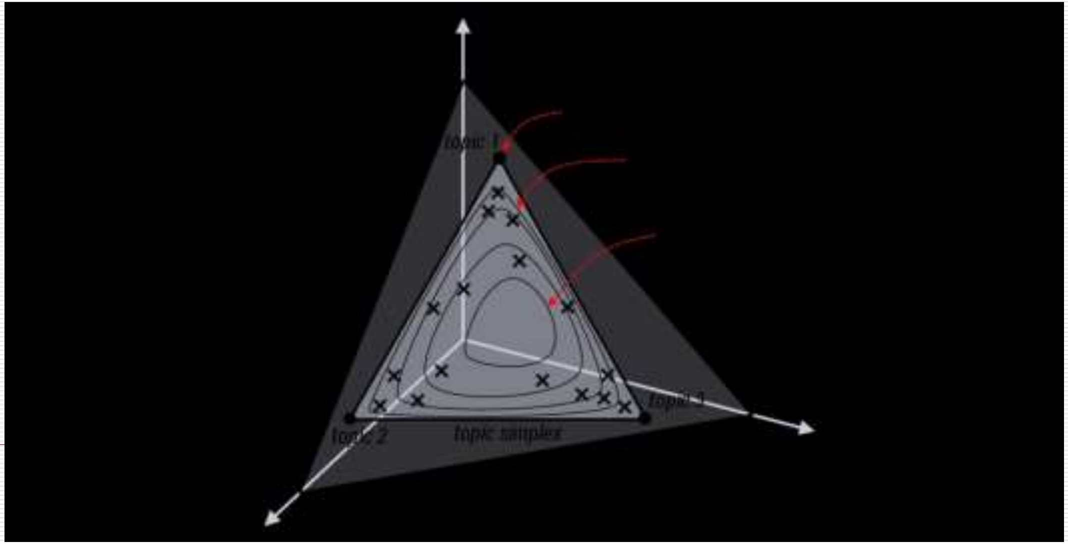
主题模型LDA

LDA生成文档过程的进一步理解

➢上面说，LDA中，主题分布 —— 比如{ P(zi), i =1,2,3 }等于{0.4,0.5,0.1}或{0.2,0.2,0.6} —— 是由dirichlet先验给定的，不是根据文档产生的。所以，**LDA生成文档的过程中，先从Dirichlet先验中“随机”抽取出主题分布**，然后从主题分布中“随机”抽取出主题，最后从确定后的主题对应的词分布中“随机”抽取出词。

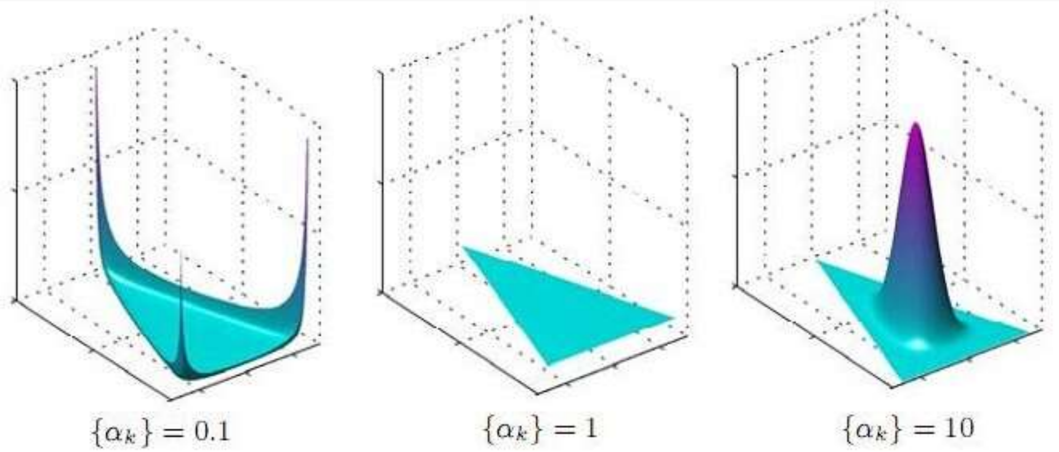
➢那么，dirichlet先验到底是如何“随机”抽取主题分布的呢？

➢事实上，从dirichlet分布中随机抽取主题分布，这个过程不是完全随机的。为了说清楚这个问题，咱们得回顾下dirichlet分布。事实上，如果我们取3个事件的话，可以建立一个三维坐标系，类似xyz三维坐标系，这里，我们把3个坐标轴弄为p1、p2、p3，如下图所示：



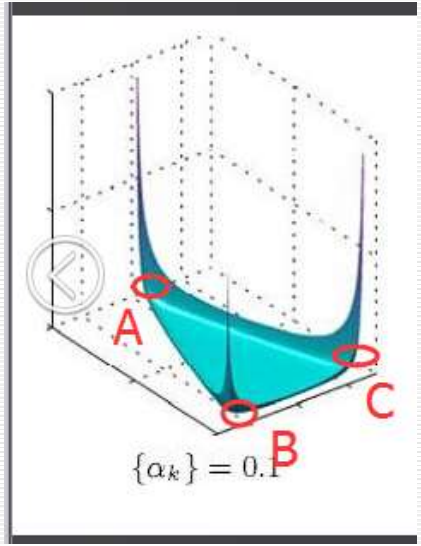
主题模型LDA

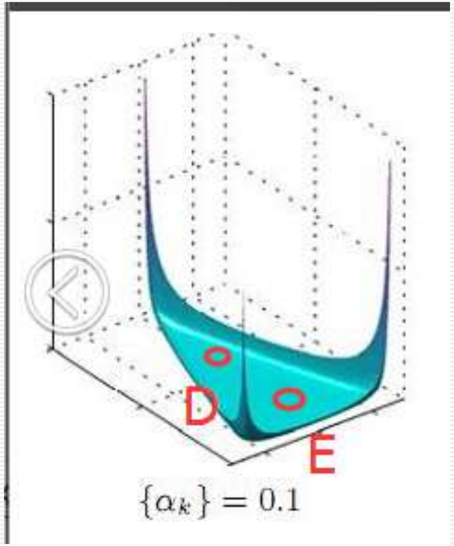
- 在这个三维坐标轴所划分的空间里，每一个坐标点(p1,p2,p3)就对应着一个主题分布，且某一个点(p1,p2,p3)的大小表示3个主题z1、z2、z3出现的概率大小（因为各个主题出现的概率和为1，所以p1+p2+p3 = 1，且p1、p2、p3这3个点最大取值为1）。比如(p1,p2,p3) = (0.4,0.5,0.1)便对应着主题分布{ P(zi), i =1,2,3 } = {0.4,0.5,0.1}。
- 可以想象到，空间里有很多这样的点(p1,p2,p3)，意味着有很多的主题分布可供选择，那dirichlet分布如何选择主题分布呢？把上面的斜三角形放倒，映射到底面的平面上，便得到如下所示的一些彩图（3个彩图中，每一个点对应一个主题分布，高度代表某个主题分布被dirichlet分布选中的概率，且**选不同的，Dirichlet 分布会偏向不同的主题分布**）：



主题模型LDA

上图中左边这个图，高度就是代表Dirichlet分布选取某个坐标点(p1,p2,p3)（这个点就是一个主题分布）的概率大小。如下图所示，平面投影三角形上的三个顶点上的点：A=(0.9,0.05,0.05)、B=(0.05,0.9,0.05)、C=(0.05,0.05,0.9)各自对应的主题分布被dirichlet分布选中的概率值很大，而平面三角形内部的两个点：D、E对应的主题分布被dirichlet分布选中的概率值很小。





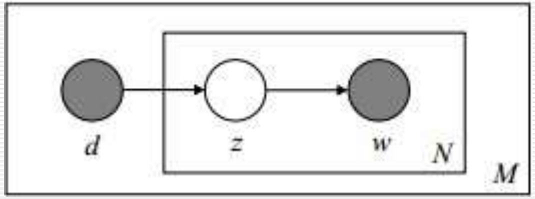
主题模型LDA

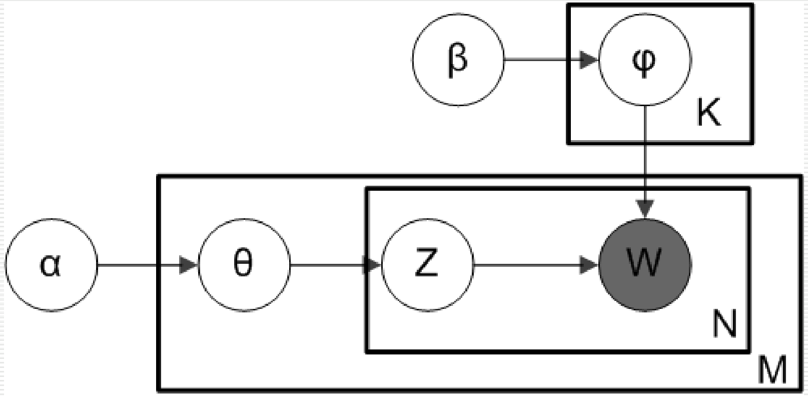
- 所以虽然说Dirichlet分布是随机选取任意一个主题分布的，但依然存在着P(A) = P(B) = P(C) >> P(D) = P(E)，即Dirichlet分布还是“偏爱”某些主题分布的。至于Dirichlet分布的参数是如何决定Dirichlet分布的形状的，可以从Dirichlet分布的定义和公式思考。
- 此外，就算说“随机”选主题也是根据主题分布来“随机”选取，这里的随机不是完全随机的意思，而是根据各个主题出现的概率值大小来抽取。比如当Dirichlet先验为文档d生成的主题分布{ P(zi), i =1,2,3 }是{0.4,0.5,0.1}时，那么主题z2在文档d中出现的概率便是0.5。所以，从主题分布中抽取主题，这个过程也不是完全随机的，而是按照各个主题出现的概率值大小进行抽取。

主题模型LDA

pLSA跟LDA的概率图对比

- 接下来，对比下LDA跟pLSA的概率模型图模型，左图是pLSA，右图是LDA（右图不太规范，z跟w都得是小写， 其中，阴影圆圈表示可观测的变量，非阴影圆圈表示隐变量，箭头表示两变量间的条件依赖性conditional dependency，方框表示重复抽样，方框右下角的数字代表重复抽样的次数）：





主题模型LDA

对应到上面右图的LDA，只有W / w是观察到的变量，其他都是隐变量或者参数，其中，Φ表示词分布，Θ表示主题分布，α 是主题分布Θ的先验分布（即Dirichlet 分布）的参数，β是词分布Φ的先验分布（即Dirichlet 分布）的参数，N表示文档的单词总数，M表示文档的总数。

所以，对于一篇文档d中的每一个单词，LDA根据先验知识α确定某篇文档的主题分布θ，然后从该文档所对应的多项分布（主题分布）θ中抽取一个主题z，接着根据先验知识β确定当前主题的词语分布ϕ，然后从主题z所对应的多项分布（词分布）ϕ中抽取一个单词w。然后将这个过程重复N次，就产生了文档d。

换言之：

1. 假定语料库中共有M篇文章，每篇文章下的Topic的**主题分布是一个从参数为α的Dirichlet先验分布中采样得到**的Multinomial分布，每个Topic下的**词分布是一个从参数为β的Dirichlet先验分布中采样得到**的Multinomial分布。
2. 对于某篇文章中的第n个词，首先从该文章中出现的每个主题的Multinomial分布**（主题分布）中选择或采样一个主题**，然后再在这个主题**对应的**词的Multinomial分布**（词分布）中选择或采样一个词。**不断重复这个随机生成过程，直到M篇文章全部生成完成。

综上，M 篇文档会对应于 M 个独立的 Dirichlet-Multinomial 共轭结构，K 个 topic 会对应于 K 个独立的 Dirichlet-Multinomial 共轭结构。

主题模型LDA

- 其中，α→θ→z 表示生成文档中的所有词对应的主题，显然 α→θ 对应的是Dirichlet 分布，θ→z 对应的是 Multinomial 分布，所以整体是一个Dirichlet-Multinomial 共轭结构，如下图所示：

$$\vec{\alpha} \underbrace{\xrightarrow{\hspace{3cm}}}_{Dirichlet} \vec{\theta}_m \underbrace{\xrightarrow{\hspace{3cm}}}_{Multinomial} \vec{z}_m$$

- 类似的，β→φ→w，容易看出， 此时β→φ对应的是 Dirichlet 分布， φ→w 对应的是 Multinomial 分布， 所以整体也是一个Dirichlet-Multinomial 共轭结构，如下图所示：

$$\vec{\beta} \underbrace{\xrightarrow{\hspace{3cm}}}_{Dirichlet} \vec{\varphi}_k \underbrace{\xrightarrow{\hspace{3cm}}}_{Multinomial} \vec{w}_{(k)}$$

主题模型LDA

pLSA跟LDA参数估计方法的对比

主题模型LDA

- 在pLSA中，我们使用EM算法去估计"主题-词项"矩阵Φ（由$P(w_j|z_k)$转换得到）和"文档-主题"矩阵Θ（由$P(z_k|d_i)$转换得到）这两个参数，而且这两参数都是个固定的值，只是未知，使用的思想其实就是极大似然估计MLE。
- **而在LDA中，估计Φ、Θ这两未知参数可以用变分(Variational inference)-EM算法，也可以用gibbs采样，前者的思想是最大后验估计MAP**（MAP与MLE类似，都把未知参数当作固定的值），**后者的思想是贝叶斯估计。**贝叶斯估计是对MAP的扩展，但它与MAP有着本质的不同，即贝叶斯估计把待估计的参数看作是服从某种先验分布的随机变量。
  - 关于贝叶斯估计再举个例子。假设中国的大学只有两种：理工科和文科，这两种学校数量的比例是1:1，其中，理工科男女比例7:1，文科男女比例1:7。某天你被外星人随机扔到一个校园，问你该学校可能的男女比例是多少？然后，你实际到该校园里逛了一圈，看到的5个人全是男的，这时候再次问你这个校园的男女比例是多少？

    1. 因为刚开始时，有先验知识，所以该学校的男女比例要么是7:1，要么是1:7，即P(比例为7:1) = 1/2，P(比例为1:7) = 1/2。
    2. 然后看到5个男生后重新估计男女比例，其实就是求P(比例7:1|5个男生）= ？，P(比例1:7|5个男生) = ？
    3. 用贝叶斯公式 $P(A|B)=\frac{P(B|A)\,P(A)}{P(B)}$，可得：P(比例7:1|5个男生) = P(比例7:1)*P(5个男生|比例7:1) / P(5个男生)，P(5个男生)是5个男生的先验概率，与学校无关，所以是个常数；类似的，P(比例1:7|5个男生) = P((比例1:7)*P(5个男生|比例1:7)/P(5个男生)。
    4. 最后将上述两个等式比一下，可得：P(比例7:1|5个男生)/P(比例1:7|5个男生) = {P((比例7:1)*P(5个男生|比例7:1)} / { P(比例1:7)*P(5个男生|比例1:7)}。

主题模型LDA

由于LDA把要估计的主题分布和词分布看作是其先验分布是Dirichlet分布的随机变量，所以，在LDA这个估计主题分布、词分布的过程中，它们的先验分布（即Dirichlet分布）事先由人为给定，那么LDA就是要去求它们的后验分布（LDA中可用gibbs采样去求解它们的后验分布，得到期望$\hat{\theta}_{mk}$、$\hat{\varphi}_{kt}$）！

在LDA中，主题分布和词分布本身都是多项分布，而由上文可知"Dirichlet分布是多项式分布的共轭先验概率分布"，因此选择Dirichlet 分布作为它们的共轭先验分布。意味着为多项分布的参数p选取的先验分布是Dirichlet分布，那么以p为参数的多项分布用贝叶斯估计得到的后验 分布仍然是Dirichlet分布。

主题模型LDA

LDA参数估计：Gibbs采样

类似于pLSA，LDA的原始论文中是用的变分-EM算法估计未知参数，后来发现另一种估计LDA未知参数的方法更好，这种方法就是：Gibbs Sampling，有时叫Gibbs采样或Gibbs抽样，都一个意思。Gibbs抽样是马尔可夫链蒙特卡尔理论（MCMC）中用来获取一系列近似等于指定多维概率分布（比如2个或者多个随机变量的联合概率分布）观察样本的算法。

OK，给定一个文档集合，w是可以观察到的已知变量，α和β是根据经验给定的先验参数，其他的变量z，θ和φ都是未知的隐含变量，需要根据观察到的变量来学习估计的。根据LDA的图模型，可以写出所有变量的联合分布：

$$p(\vec{w}_m, \vec{z}_m, \vec{\vartheta}_m, \underline{\Phi}|\vec{\alpha}, \vec{\beta}) = \prod_{n=1}^{N_m} p(w_{m,n}|\vec{\varphi}_{z_{m,n}}) p(z_{m,n}|\vec{\vartheta}_m) \cdot p(\vec{\vartheta}_m|\vec{\alpha}) \cdot p(\underline{\Phi}|\vec{\beta})$$

注：上述公式中及下文中，$z_{m,n}$等价上文中定义的$z_{i,j}$，$w_{m,n}$等价于上文中定义的$w_{i,j}$，$\vec{\varphi}_{z_{m,n}}$等价于上文中定义的$\phi_{z_{i,j}}$，等价于上文中定义的θ_i。

因为α产生主题分布θ，主题分布θ确定具体主题，且β产生词分布φ、词分布φ确定具体词，所以上述式子等价于下述式子所表达的**联合概率分布**$p(\vec{w}, \vec{z})$：

$$p(\vec{w}|\vec{z}, \Phi) = \prod_{i=1}^{W} p(w_i|z_i) = \prod_{i=1}^{W} \varphi_{z_i, w_i}$$

主题模型LDA

由于样本中的词服从参数为主题z_i 的独立多项分布，这意味着可以把上面对词的乘积分解成分别对主题和对词的两层乘积：

$$p(\vec{w}|\vec{z},\Phi)=\prod_{k=1}^{K}\prod_{\{i:z_i=k\}}p(w_i=t|z_i=k)=\prod_{k=1}^{K}\prod_{t=1}^{V}\varphi_{k,t}^{n_k^{(t)}}$$

其中，$n_k^{(t)}$ 是词 t 在主题 k 中出现的次数。

回到第一个因子上来。目标分布$p(\vec{w}|\vec{z},\vec{\beta})$需要对词分布Φ积分，且结合我们之前在3.1节定义的Dirichlet 分布的归一化系数$\Delta(\vec{\alpha})$的公式

$$\Delta(\vec{\alpha})=\int\prod_{k=1}^{V}p_k^{\alpha_k-1}d\vec{p}$$

主题模型LDA

$$
\begin{aligned}
p(\vec{w}|\vec{z},\vec{\beta}) &= \int p(\vec{w}|\vec{z},\Phi)p(\Phi|\vec{\beta})\mathrm{d}\Phi \\
&= \int \prod_{z=1}^{K} \frac{1}{\Delta(\vec{\beta})} \prod_{t=1}^{V} \varphi_{z,t}^{n_z^{(t)}+\beta_t-1}\mathrm{d}\vec{\varphi}_z \\
&= \prod_{z=1}^{K} \frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{\beta})}, \quad \vec{n}_z = \{n_z^{(t)}\}_{t=1}^{V}
\end{aligned}
$$

这个结果可以看作K个Dirichlet-Multinomial模型的乘积。

现在开始求第二个因子$p(\vec{z}|\vec{\alpha})$。类似于$p(\vec{w}|\vec{z},\vec{\beta})$的步骤，先写出条件分布，然后分解成两部分的乘积：

$$
p(\vec{z}|\Theta) = \prod_{i=1}^{W} p(z_i|d_i) = \prod_{m=1}^{M}\prod_{k=1}^{K} p(z_i=k|d_i=m) = \prod_{m=1}^{M}\prod_{k=1}^{K} \theta_{m,k}^{n_m^{(k)}}
$$

其中，d_i 表示的单词 i 所属的文档，$n_m^{(k)}$是主题 k 在文章 m 中出现的次数。

主题模型LDA

对主题分布Θ积分可得：

$$
\begin{aligned}
p(\vec{z}|\vec{\alpha}) &= \int p(\vec{z}|\Theta)p(\Theta|\vec{\alpha})\mathrm{d}\Theta \\
&= \int \prod_{m=1}^{M} \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^{K} \vartheta_{m,k}^{n_m^{(k)}+\alpha_k-1} \mathrm{d}\vec{\vartheta}_m \\
&= \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})}, \quad \vec{n}_m = \{n_m^{(k)}\}_{k=1}^{K}
\end{aligned}
$$

综合第一个因子和第二个因子的结果，**得到$p(\vec{w},\vec{z})$的联合分布结果为**：

$$
p(\vec{z},\vec{w}|\vec{\alpha},\vec{\beta}) = \prod_{z=1}^{K} \frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{\beta})} \cdot \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})}
$$

主题模型LDA

接下来，**有了联合分布**$p(\vec{w},\vec{z})$**，咱们便可以通过联合分布来计算在给定可观测变量 w 下的隐变量 z 的条件分布（后验分布）**$p(\vec{z}|\vec{w})$**来进行贝叶斯分析。**

换言之，有了这个联合分布后，要求解第m篇文档中的第n个词（下标为$i=(m,n)$的词）的全部条件概率就好求了。

先定义几个变量。$\neg i$ 表示除去i的词，$\vec{w}=\{w_i=t,\vec{w}_{\neg i}\}$，$\vec{z}=\{z_i=k,\vec{z}_{\neg i}\}$。

然后，排除当前词的主题分配，即根据其他词的主题分配和观察到的单词来计算当前词主题的概率公式为：

$$
\begin{aligned}
p(z_i=k|\vec{z}_{\neg i},\vec{w}) &= \frac{p(\vec{w},\vec{z})}{p(\vec{w},\vec{z}_{\neg i})} \\
&= \frac{p(\vec{w}|\vec{z})}{p(\vec{w}_{\neg i}|\vec{z}_{\neg i})p(w_i)} \cdot \frac{p(\vec{z})}{p(\vec{z}_{\neg i})} \\
&\propto \frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{n}_{z,\neg i}+\vec{\beta})} \cdot \frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{n}_{m,\neg i}+\vec{\alpha})} \\
&= \frac{\Gamma(n_k^{(t)}+\beta_t)\Gamma(\sum_{t=1}^V n_{k,\neg i}^{(t)}+\beta_t)}{\Gamma(n_{k,\neg i}^{(t)}+\beta_t)\Gamma(\sum_{t=1}^V n_k^{(t)}+\beta_t)} \cdot \frac{\Gamma(n_m^{(k)}+\alpha_k)\Gamma(\sum_{k=1}^K n_{m,\neg i}^{(k)}+\alpha_t)}{\Gamma(n_{m,\neg i}^{(k)}+\alpha_t)\Gamma(\sum_{k=1}^K n_m^{(k)}+\alpha_k)} \\
&= \frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^V n_{k,\neg i}^{(t)}+\beta_t} \cdot \frac{n_{m,\neg i}^{(k)}+\alpha_k}{[\sum_{k=1}^K n_m^{(k)}+\alpha_k]-1} \\
&\propto \frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^V n_{k,\neg i}^{(t)}+\beta_t}(n_{m,\neg i}^{(k)}+\alpha_k)
\end{aligned}
$$

主题模型LDA

$$
\begin{aligned}
p(z_i = k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}) &\propto p(z_i = k, w_i = t|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}) \\
&= \int p(z_i = k, w_i = t, \overrightarrow{\theta}_m, \overrightarrow{\varphi}_k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}) d\overrightarrow{\theta}_m d\overrightarrow{\varphi}_k \\
&= \int p(z_i = k, \overrightarrow{\theta}_m|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}) \cdot p(w_i = t, \overrightarrow{\varphi}_k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}) d\overrightarrow{\theta}_m d\overrightarrow{\varphi}_k \\
&= \int p(z_i = k|\overrightarrow{\theta}_m) p(\overrightarrow{\theta}_m|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}) \cdot p(w_i = t|\overrightarrow{\varphi}_k) p(\overrightarrow{\varphi}_k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}) d\overrightarrow{\theta}_m d\overrightarrow{\varphi}_k \\
&= \int p(z_i = k|\overrightarrow{\theta}_m) Dir(\overrightarrow{\theta}_m|\overrightarrow{n}_{m,\neg i} + \overrightarrow{\alpha}) d\overrightarrow{\theta}_m \\
&\quad \cdot \int p(w_i = t|\overrightarrow{\varphi}_k) Dir(\overrightarrow{\varphi}_k|\overrightarrow{n}_{k,\neg i} + \overrightarrow{\beta}) d\overrightarrow{\varphi}_k \\
&= \int \theta_{mk} Dir(\overrightarrow{\theta}_m|\overrightarrow{n}_{m,\neg i} + \overrightarrow{\alpha}) d\overrightarrow{\theta}_m \cdot \int \varphi_{kt} Dir(\overrightarrow{\varphi}_k|\overrightarrow{n}_{k,\neg i} + \overrightarrow{\beta}) d\overrightarrow{\varphi}_k \\
&= E(\theta_{mk}) \cdot E(\varphi_{kt}) \\
&= \hat{\theta}_{mk} \cdot \hat{\varphi}_{kt}
\end{aligned}
$$

主题模型LDA

最后一步，便是根据Markov链的状态z_i获取主题分布的参数Θ和词分布的参数Φ。

换言之根据贝叶斯法则和Dirichlet先验，以及上文中得到的$p(\vec{w}|\vec{z},\Phi)$和$p(\vec{z}|\Theta)$各自被分解成两部分乘积的结果，可以计算得到**每个文档上Topic的后验分布和每个Topic下的词的后验分布分别如下**（据上文可知：其后验分布跟它们的先验分布一样，也都是Dirichlet 分布）：

$$p(\vec{\vartheta}_m|\vec{z}_m,\vec{\alpha}) = \frac{1}{Z_{\vartheta_m}} \prod_{n=1}^{N_m} p(z_{m,n}|\vec{\vartheta}_m) \cdot p(\vec{\vartheta}_m|\vec{\alpha}) = \mathrm{Dir}(\vec{\vartheta}_m|\vec{n}_m + \vec{\alpha})$$

$$p(\vec{\varphi}_k|\vec{z},\vec{w},\vec{\beta}) = \frac{1}{Z_{\varphi_k}} \prod_{\{i:z_i=k\}} p(w_i|\vec{\varphi}_k) \cdot p(\vec{\varphi}_k|\vec{\beta}) = \mathrm{Dir}(\vec{\varphi}_k|\vec{n}_k + \vec{\beta})$$

其中，$\vec{n}_m$是构成文档m的主题数向量，$\vec{n}_k$是构成主题k的词项数向量。

此外，别忘了上文所述的Dirichlet的一个性质，如下：

“如果$\vec{p} \sim Dir(\vec{t}|\vec{\alpha})$，同样可以证明有下述结论成立：

$$E(\vec{p}) = \left(\frac{\alpha_1}{\sum_{i=1}^K \alpha_i}, \frac{\alpha_2}{\sum_{i=1}^K \alpha_i}, \cdots, \frac{\alpha_K}{\sum_{i=1}^K \alpha_i}\right)$$

主题模型LDA

即：如果$\vec{p} \sim \mathrm{Dir}(\vec{p}|\vec{\alpha})$，则$\vec{p}$中的任一元素$\vec{p}$的期望是：

$$
\begin{aligned}
E(p_i) &= \int_0^1 p_i \cdot \mathrm{Dir}(\vec{p}|\vec{\alpha}) \mathrm{d}p \\
&= \frac{\Gamma\left(\sum_{k-1}^{K} \alpha_k\right)}{\Gamma(\alpha_i)} \cdot \frac{\Gamma(\alpha_i + 1)}{\Gamma\left(\sum_{k-1}^{K} \alpha_k + 1\right)} \\
&= \frac{\alpha_i}{\sum_{k-1}^{K} \alpha_k}
\end{aligned}
$$

可以看出，超参数α_k的直观意义就是事件先验的伪计数(prior pseudo-count)。

所以，**最终求解的Dirichlet 分布期望为**：

$$
\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t-1}^{V} n_k^{(t)} + \beta_t}
$$

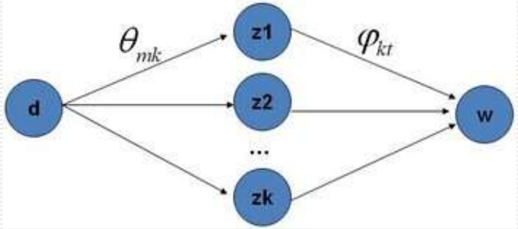
$$
\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k-1}^{K} n_m^{(k)} + \alpha_k}
$$

主题模型LDA

然后将 $\varphi_{k,t}$ 和 $\vartheta_{m,k}$ 的结果代入之前得到的 $p(z_i = k|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}) \propto p(z_i = k, w_i = t|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}_{\neg i})$ 的结果中，可得：

$$p(z_i = k|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}) \propto \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_{m,\neg i}^{(k)} + \alpha_k)} \cdot \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V}(n_{k,\neg i}^{(t)} + \beta_t)}$$

仔细观察上述结果，可以发现，式子的右半部分便是 $p(topic|doc) \cdot p(word|topic)$，这个概率的值对应着 $doc \rightarrow topic \rightarrow word$ 的路径概率。如此，K 个topic 对应着K条路径，Gibbs Sampling 便在这K 条路径中进行采样，如下图所示：



就这样，Gibbs Sampling通过求解出主题分布和词分布的后验分布，从而成功解决主题分布和词分布这两参数未知的问题。

谢谢大家