通俗理解 LDA 主题模型

2015 年 1 月 10 日

1 前言

印象中，最开始听说“LDA”这个名词，是缘于 rickjin 在 2013 年 3 月写的一个 LDA 科普系列，叫 LDA 数学八卦，我当时一直想看来着，记得还打印过一次，但不知是因为这篇文档的前序铺垫太长（现在才意识到这些“铺垫”都是深刻理解 **LDA** 的基础，但如果没有人帮助初学者提纲挈领、把握主次、理清思路，则很容易陷入 **LDA** 的细枝末节之中），还是因为其中的数学推导细节太多，导致一直没有完整看完过。

2013 年 12 月，在我组织的 Machine Learning 读书会第 8 期上，@ 夏粉 _ 百度讲机器学习中排序学习的理论和算法研究，@ 沈醉 2011 则讲主题模型的理解。又一次碰到了主题模型，当时貌似只记得沈博讲了一个汪峰写歌词的例子，依然没有理解 LDA 到底是怎样一个东西（但理解了 LDA 之后，再看沈博主题模型的 PPT 会很赞）。

直到昨日下午，机器学习班第 12 次课上，邹博讲完 LDA 之后，才真正明白 LDA 原来是那么一个东东！上完课后，趁热打铁，再次看 LDA 数学八卦，发现以前看不下去的文档再看时竟然一路都比较顺畅，一口气看完大部。看完大部后，思路清晰了，知道理解 LDA，可以分为下述 5 个步骤：

1. 一个函数：gamma 分布
2. 四个分布：二项分布、多项分布、beta 分布、Dirichlet 分布
3. 一个概念和一个理念：共轭先验和贝叶斯框架
4. 两个模型：pLSA、LDA（在本文第 4 部分阐述）
5. 一个采样：Gibbs 采样

本文便按照上述 5 个步骤来阐述，希望读者看完本文后，能对 LDA 有个尽量清晰完整的了解。同时，本文基于邹博讲 LDA 的PPT、rickjin 的 LDA 数学八卦及其它参考资料写就，可以定义为一篇学习笔记或课程笔记，当然，后续不断加入了很多自己的理解。若有任何问题，欢迎随时于本文评论下指出，thanks。

2 gamma 函数

2.1 整体把握 LDA

关于 LDA 有两种含义，一种是线性判别分析（Linear Discriminant Analysis），一种是概率主题模型：隐含狄利克雷分布（Latent Dirichlet Allocation，简称 LDA），本文讲后者（前者会在后面的博客中阐述）。

另外，我先简单说下 LDA 的整体思想，不然我怕你看了半天，铺了太长的前奏，却依然因没见到 LDA 的影子而显得“心浮气躁”，导致不想再继续看下去。所以，先给你吃一颗定心丸，明白整体框架后，咱们再一步步抽丝剥茧，展开来论述。

按照 wiki 上的介绍，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。此外，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。

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，然后通过学习的方式，获取每个主题 Topic 对应的词语。如下图所示：

"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

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.

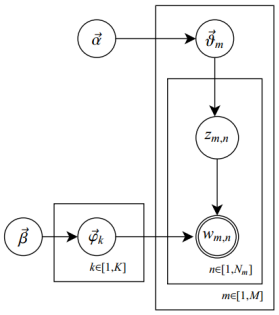
而当我们看到一篇文章后，往往喜欢推测这篇文章是如何生成的，我们可能会认为作者先确定这篇文章的几个主题，然后围绕这几个主题遣词造句，表达成文。LDA 就是要干这事：根据给定的一篇文档，推测其主题分布。

然，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 $\Phi_{zi,j}$
- 从词语的多项式分布 $\Phi_{zi,j}$ 中采样最终生成词语 $w_{i,j}$

其中，类似 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）：



恩，不错，短短 6 句话整体概括了整个 LDA 的主体思想！但也就是上面短短 6 句话，却接连不断或重复出现了二项分布、多项式分布、beta 分布、狄利克雷分布（Dirichlet 分布）、共轭先验概率分布、取样，那么请问，这些都是啥呢?

这里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。

- 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 $X\ b(n,p)$。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：

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

对于 $k=0,1,...,n$ 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 $C(n,k)$。回想起高中所学的那丁点概率知识了么：想必你当年一定死记过 $C(n,k)$ 这个二项式系数就是 $\frac{n!}{k!(n-k)!}$。

- 多项分布，是二项分布扩展到多维的情况

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：

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

- 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：

$$f(x;\alpha,\beta)=\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

其中：

$$\frac{1}{B(\alpha,\beta)}=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)},\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}\mathrm{d}t$$

注：$\Gamma(x)$ 便是所谓的 gamma 函数，下文会具体阐述。

- 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：

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

  其中：

$$B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)}, \sum x_i = 1$$

  至此，我们可以看到二项分布和多项分布很相似，**Beta** 分布和 **Dirichlet** 分布很相似，而至于“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”这点在下文中说明。

OK，接下来，咱们就按照本文开头所说的思路：“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，外加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（文档 -主题，主题 -词语），一个采样：Gibbs 采样”一步步详细阐述，争取给读者一个尽量清晰完整的 LDA。

（当然，如果你不想深究背后的细节原理，只想整体把握 LDA 的主体思想，可直接跳到本文第 4 部分，看完第 4 部分后，若还是想深究背后的细节原理，可再回到此处开始看）

2.2 gamma 函数

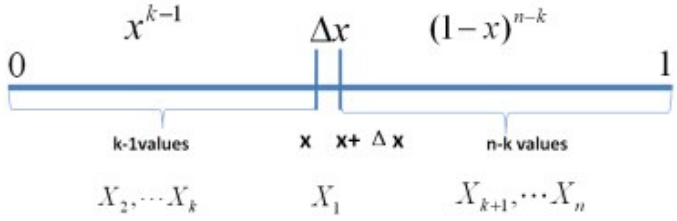
咱们先来考虑一个问题（此问题 1 包括下文的问题 2-问题 4 皆取材自 LDA 数学八卦）：

- 问题 **1** 随机变量 $X_1, X_2, ..., X_n \overset{iid}{\sim} Uniform(0, 1)$
- 把这 n 个随机变量排序后得到顺序统计量 $X_{(1)}, X_{(2)}, ..., X_{(n)}$
- 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 $X_{(k)}$ 落在区间 $[x, x + \Delta x]$ 的概率。即求下述式子的值：

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

首先，把 [0,1] 区间分成三段 $[0, x)$，$[x, x + \Delta x]$，$(x + \Delta x, 1]$，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 $[x, x + \Delta x]$ 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，$(x + \Delta x, 1]$ 这个区间中应该有 n−k 个数。如下图所示：



从而问题转换为下述事件 E：

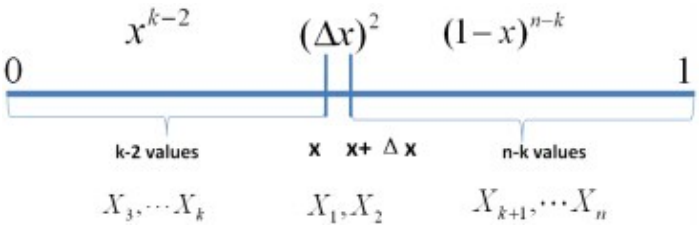
$$\begin{aligned} E \quad = \quad & \{X_1 \in [x, x + \Delta x], \\ & X_i \in [0, x) \quad (i = 2, \ldots, k), \\ & X_j \in (x + \Delta x, 1) \quad (j = k + 1, \ldots, n)\} \end{aligned}$$

对上述事件 E，有：

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

其中，$o(\Delta x)$ 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 $[x, x+\Delta x]$ 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 $[x, x+\Delta x]$ 呢？如下图所示：



类似于事件 E，对于 2 个数落在区间 $[x, x+\Delta x]$ 的事件 E'：

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

有：

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

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

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

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：

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

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 $\Gamma(n) = (n-1)!$ 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：

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

如果熟悉 beta 分布的朋友，可能会惊呼：哇，竟然推出了 beta 分布！

3 beta 分布

3.1 beta 分布

在概率论中，beta 是指一组定义在区间 $(0,1)$ 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:

$$\begin{aligned} f(x;\alpha,\beta) &= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\mathrm{d}u} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \end{aligned}$$

其中的 Γ 便是 $\Gamma(x)$ 函数：

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

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

3.2 Beta-Binomial 共轭

回顾下 2.1 节开头所提出的问题：“问题 1 随机变量 $X_1, X_2, ..., X_n \overset{iid}{\sim} Uniform(0,1)$，把这 n 个随机变量排序后得到顺序统计量，然后请问的分布是什么。”如果，咱们要在这个问题的基础上增加一些观测数据，变成问题 2：

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

根据“Y_i 中有 m_1 个比 p 小，m_2 个比 p 大”，换言之，Y_i 中有 m_1 个比 $X_{(k)}$ 小，m_2 个比 $X_{(k)}$ 大，所以 $X_{(k)}$ 是 $X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_n \overset{iid}{\sim} Uniform(0,1)$ 中第 $k+m_1$ 大的数。

根据 2.1 节最终得到的结论“只要落在 $[x, x+\Delta x]$ 内的数字超过一个，则对应的事件的概率就是 $o(\Delta x)$”，继而推出事件服从 beta 分布，从而可知的概率密度函数为：

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

熟悉贝叶斯方法（不熟悉的没事，参见此文第一部分）的朋友心里估计又犯“嘀咕”了，这不就是贝叶斯式的思考过程么？

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”，针对是做了次贝努利实验，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 $\chi \Rightarrow$ 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 χ 后，人们对 θ 的认知为 $\pi(\theta|x)$。

类比到现在这个问题上，我们也可以试着写下：

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

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

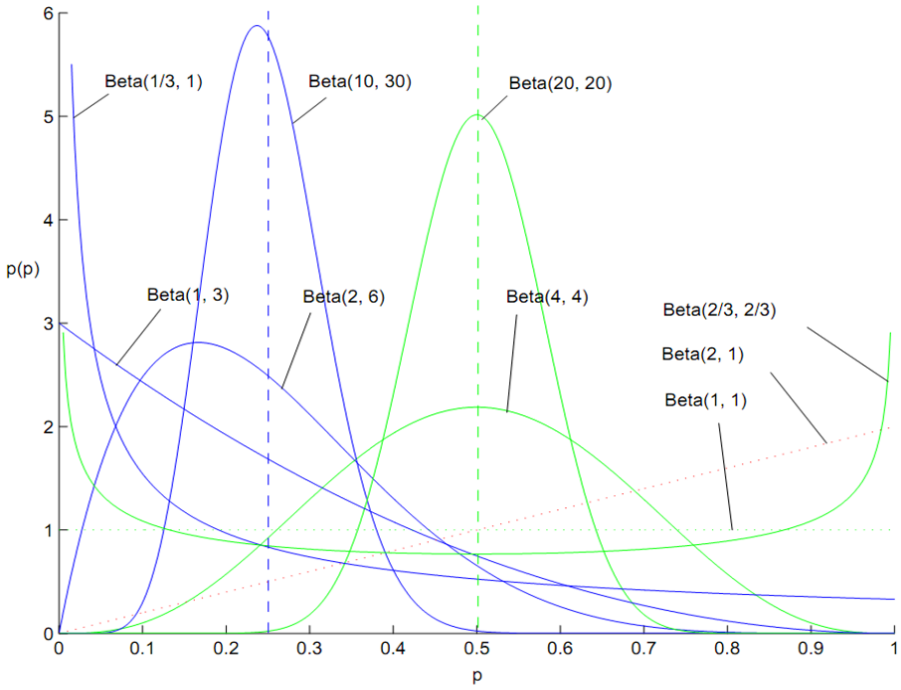
更一般的，对于非负实数和，我们有如下关系

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

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。

此外，如何理解参数 α 和 β 所表达的意义呢？α、β 可以认为形状参数，通俗但不严格的理解是，α 和 β 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：



3.3 共轭先验分布

什么又是共轭呢？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 $P(\theta|x)$ 和先验概率 $p(\theta)$ 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 $p(\theta)$ 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
  $\theta \leftarrow \theta + \Delta\theta$
- 当重新估计 θ 的时候，给出新参数值 θ 的新概率分布，即 $P(\theta|x)$。

事实上，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，$P(\theta|x)$ 表示以预估 θ 为参数的 x 概率分布，可以直接求得，$p(\theta)$ 是已有原始的 θ 概率分布。所以，如果我们选取 $P(\theta|x)$ 的共轭先验作为 $p(\theta)$ 的分布，那么 $P(\theta|x)$ 乘以 $p(\theta)$，然后归一化的结果 $P(\theta|x)$ 跟和 $p(\theta)$ 的形式一样。换句话说，先验分布是 $p(\theta)$，后验分布是 $P(\theta|x)$，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）。

举个例子。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，具有两个参数和，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

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

然后计算后验概率

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

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。

3.4 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：

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

注意到上式最后结果的右边积分

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

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

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

从而求得

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

的结果为

$$\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}$$

最后将此结果带入 $E(p)$ 的计算式，得到：

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

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）可以用来 $\frac{\alpha}{\alpha+\beta}$ 估计。此外，狄利克雷 Dirichlet 分布也有类似的结论，即如果 $\overrightarrow{p}\sim Dir(\overrightarrow{t}|\overrightarrow{\alpha})$，同样可以证明有下述结论成立：

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

那什么是 Dirichlet 分布呢？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。

4 Dirichlet 分布

4.1 Dirichlet 分布

根据 wikipedia 上的介绍，维度 $K\geq 2$（$x_1,x_2,...,x_{K-1}$ 维，共 K 个）的狄利克雷分布在参数 $\alpha_1,...,\alpha_K>0$ 上、基于欧几里得空间 R^{K-1} 里的勒贝格测度有个概率密度函数，定义为：

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

其中，$B(\alpha)$ 相当于是多项 beta 函数

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

且 $\alpha=(\alpha_1,...,\alpha_K)$

此外，$x_1+x_2+...+x_{K-1}+x_K=1$，$x_1,x_2,...,x_{K-1}>0$，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

$$\begin{aligned} p(\overrightarrow{p}|\overrightarrow{\alpha}) &= Dir(\overrightarrow{p}|\overrightarrow{\alpha}) \\ &\triangleq \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}p_k^{\alpha_k-1} \\ &\triangleq \frac{1}{\Delta(\overrightarrow{\alpha})}\prod_{k=1}^{K}p_k^{\alpha_k-1} \end{aligned}$$

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

$$\Delta(\overrightarrow{\alpha}) = \frac{\prod_{k=1}^{dim(\overrightarrow{\alpha})}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{dim(\overrightarrow{\alpha})}\alpha_k)}$$

且根据 Dirichlet 分布的积分为 1（概率的基本性质），可以得到：

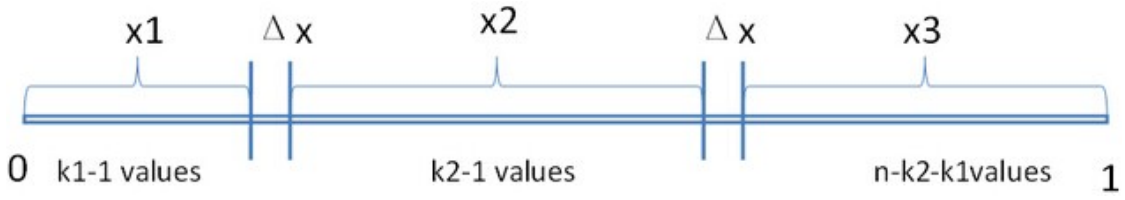
$$\int_{\overrightarrow{p}}\prod_{k=1}^{V}p_k^{\alpha_k-1}\mathrm{d}\overrightarrow{p} = \Delta(\overrightarrow{\alpha})$$

4.2 Dirichlet-Multinomial 共轭

下面，在 3.2 节问题 2 的基础上继续深入，引出问题 3。

- $X_1, X_2, ..., X_n \overset{iid}{\sim} Uniform(0,1)$,
- 排序后对应的顺序统计量 $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$,
- 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：



从而有：

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

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

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

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

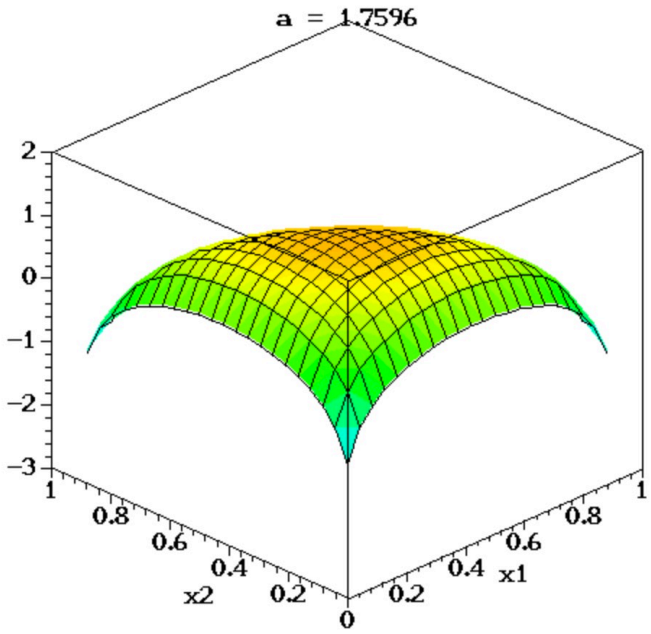
$$Dir(x_1, x_2, x_3|k_1, k_2, n-k_1-k_2+1)$$

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

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

这个就是一般形式的 3 维 Dirichlet 分布，即 $\overrightarrow{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ 便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，点击查看动态图）：



上图中，取 K=3，也就是有两个独立参数 x1,x2，分别对应图中的两个坐标轴，第三个参数始终满足 x3=1-x1-x2 且 $\alpha 1 = \alpha 2 = \alpha 3 = \alpha$，图中反映的是参数 α 从 $\alpha = (0.3, 0.3, 0.3)$ 变化到 $(2.0, 2.0, 2.0)$ 时的概率对数值的变化情况。

为了论证 Dirichlet 分布是多项式分布的共轭先验概率分布，下面咱们继续在上述问题 3 的基础上再进一步，提出问题 **4**。

1. 问题 4 ，$X_1, X_2, ..., X_n \overset{iid}{\sim} Uniform(0,1)$ 排序后对应的顺序统计量 $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$,
2. 令,$p_1 = X_{(k_1)}, p_2 = X_{(k_1+k_2)}, p_3 = 1-p_1-p_2$,(此处的 p3 非变量，只是为了表达方便)，现在要猜测; $\overrightarrow{p} = (p_1, p_2, p_3)$
3. $Y_1, Y_2, ..., Y_m \overset{iid}{\sim} Uniform(0,1)$, 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,m3, m=m1+m2+m3;
4. 问后验分布 $P(\overrightarrow{p}|Y_1, Y_2, ..., Y_m)$ 的分布是什么。

为了方便讨论，记 $\overrightarrow{m} = (m_1, m_2, m_3)$，及 $\overrightarrow{k} = (k_1, k_2, n-k_1-k_2+1)$，根据已知条件“$Y_1, Y_2, ..., Y_m \overset{iid}{\sim} Uniform(0,1)$, 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，后验分布应该为 $Dir(\overrightarrow{p}|k_1+m_1, k_1+m_2, n-k_1-k_2+1+m_3)$，即一般化的形式表示为: $Dir(\overrightarrow{p}|\overrightarrow{k}+\overrightarrow{m})$。

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 $\overrightarrow{p}=(p_1,p_2,p_3)$，其先验分布为；$Dir(\overrightarrow{p}|\overrightarrow{k})$

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 $\overrightarrow{m}=(m_1,m_2,m_3)$ 服从多项分布 $Mult(\overrightarrow{m}|\overrightarrow{p})$

3. 在给定了来自数据提供的知识 $\overrightarrow{m}$ 后，$\overrightarrow{p}$ 的后验分布变为 $Dir(\overrightarrow{p}|\overrightarrow{k}+\overrightarrow{m})$

上述贝叶斯分析过程的直观表述为：

$$Dir(\overrightarrow{p}|\overrightarrow{k})+MultCount(\overrightarrow{m})=ir(\overrightarrow{p}|\overrightarrow{k}+\overrightarrow{m})$$

令 $\overrightarrow{\alpha}=\overrightarrow{k}$，可把 $\overrightarrow{\alpha}$ 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：

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

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：

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

而对于给定的 $\overrightarrow{p}$ 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 $Dir(\overrightarrow{p}|\overrightarrow{\alpha})$ 和多项分布 $Mult(\overrightarrow{n}|\overrightarrow{p},N)$ 是共轭关系

5 主题模型 LDA

在开始下面的旅程之前，先来总结下我们目前所得到的最主要的几个收获：

- 通过上文的第 2.2 节，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。”

- 通过上文的 4.2 节，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 $\overrightarrow{\alpha}$ 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：

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

  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。”

- 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 $\chi \Rightarrow$ 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 $\pi(\theta)$ 看做是固定的未知常数，即概率虽然是未知的，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 $\pi(\theta)$ 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的分布。

OK，在杀到终极 boss——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：

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

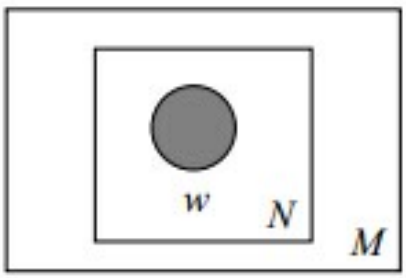
5.1 各个基础模型

5.1.1 Unigram model

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

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

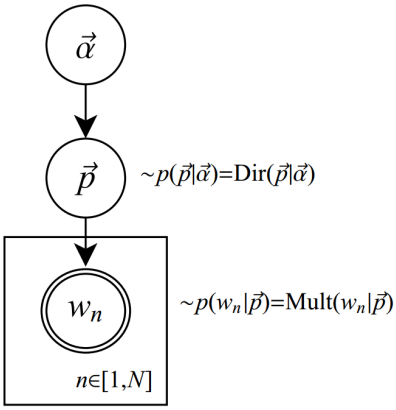
其图模型为（图中被涂色的 w 表示可观测变量，N 表示一篇文档中总共 N 个单词，M 表示 M 篇文档）：或为：



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，$n \in [1, N]$ 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：

1. p 是词服从的 Multinomial 分布的参数



2. α 是 Dirichlet 分布（即 Multinomial 分布的先验分布）的参数。

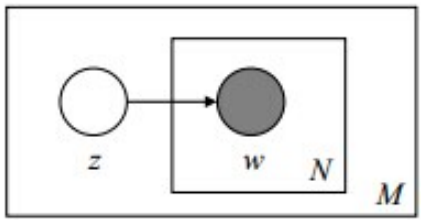
一般 α 由经验事先给定，p 由观察到的文本中出现的词学习得到，表示文本中出现每个词的概率

5.1.2 Mixture of unigrams model

一篇文档只由一个主题生成。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 $z_1, ..., z_k$，生成文档 $\boldsymbol{w}$ 的概率为：

$$p(\boldsymbol{w}) = p(z_1)\prod_{n=1}^{N} p(w_n|z_1) + \ldots + p(z_k)\prod_{n=1}^{N} p(w_n|z_k) = \sum_z p(z)\prod_{n=1}^{N} p(w_n|z)$$

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）：



5.2 PLSA 模型

啊哈，长征两万五，经过前面这么长的铺垫，终于快要接近 LDA 模型了！因为跟 LDA 模型最为接近的便是下面要阐述的这个 pLSA 模型，理解了 pLSA 模型后，到 LDA 模型也就一步之遥——给 pLSA 加上贝叶斯框架，便是 LDA。

5.2.1 什么是 pLSA 模型

OK，在上面的 Mixture of unigrams model 中，我们假定一篇文档只由一个主题生成，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。那么在 pLSA 中，文档是怎样被生成的呢？

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。

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。

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。

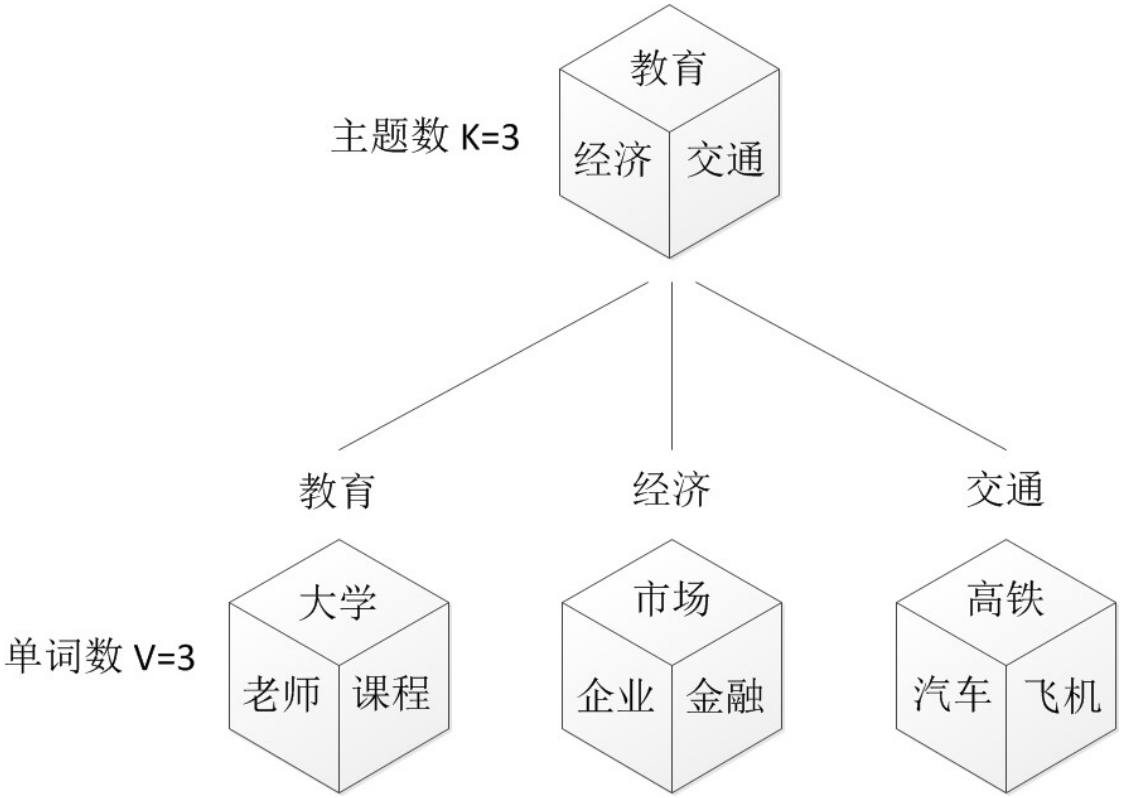
3. 最后，你不停的重复扔“文档 -主题”骰子和”主题 -词项“骰子，重复 N 次（产生 N 个词），完成一篇文档，重复这产生一篇文档的方法 M 次，则完成 M 篇文档。

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 $z_k \in \{z_1, ..., z_K\}$。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



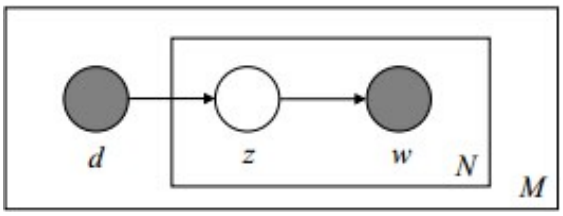
2. 选定文档 d_i 后，从主题分布中按照概率 $P(z_k|d_i)$ 选择一个隐含的主题类别 z_k

3. 选定后，从词分布中按照概率选择一个词

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。

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）。

文档 d 和单词 w 自然是可被观察到的，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）：



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，训练出文档 -主题和主题 -词项，如下公式所示：

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

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。

5.2.2 EM 算法的简单介绍

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

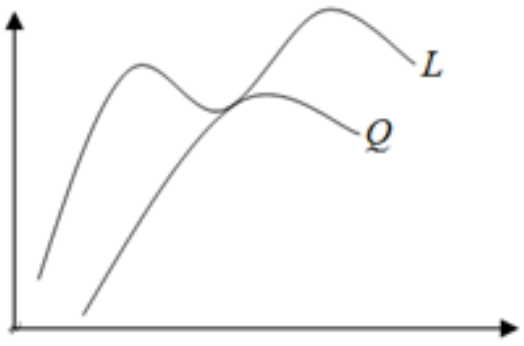
$$\theta^{(n+1)} = \max_{\theta} L(\theta) - L(\theta^{(n)})$$

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
   (a) 给出当前的参数估计 $\theta^{(n)}$，计算似然函数 $L(\theta)$ 的下界 $Q(\theta;\theta^n)$
   (b) 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。

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：



以下是详细介绍。

假定有训练集 $\{x^{(1)}, ..., x^{(m)}\}$，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：

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

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，$Q_i \geq 0$，且结合 Jensen 不等式，有：

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

为了寻找尽量紧的下界，我们可以让使上述等号成立，而若要让等号成立的条件则是：

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

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：

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

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

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

}

OK，EM 算法还会在本博客后面的博文中具体阐述。接下来，回到 pLSA 参数的估计问题上。

5.2.3 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 ϕ_k 用表示词表 V 在主题上的一个多项分布，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 出现在主题中的概率，即

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

- 用 θ_i 表示所有主题 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

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

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：

$$\begin{aligned} \Phi &= [\phi_1, \ldots, \phi_K], \quad z_k \in Z \\ \Theta &= [\theta_1, \ldots, \theta_M], \quad d_i \in D \end{aligned} \tag{2}$$

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：

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

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 d_i 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：

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

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。

- **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：

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

- **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：

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

这是一个多元函数求极值问题，并且已知有如下约束条件：

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

熟悉凸优化的朋友应该知道，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：

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

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：

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

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

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

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

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 $\phi_{k,j}$ 和 $\theta_{i,k}$。

5.3 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）。

5.3.1 pLSA 跟 LDA 的对比：生成文档与参数估计

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：

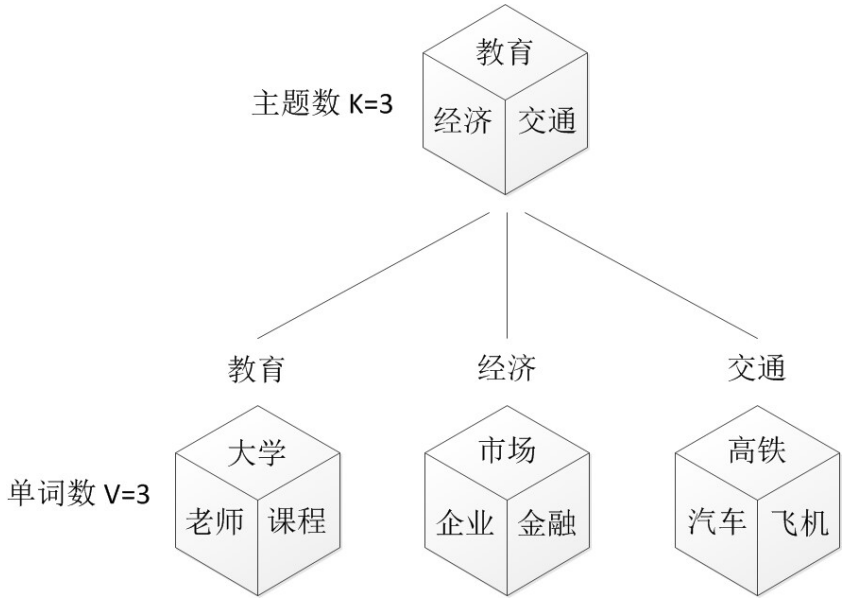
1. 按照概率 $P(d_i)$ 选择一篇文档 d_i
2. 选定文档 d_i 后，确定文章的主题分布
3. 从主题分布中按照概率 $P(z_k|d_i)$ 选择一个隐含的主题类别 z_k
4. 选定 z_k 后，确定主题下的词分布
5. 从词分布中按照概率 $P(w_j|z_k)$ 选择一个词”

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 α 为的 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。

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。



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。

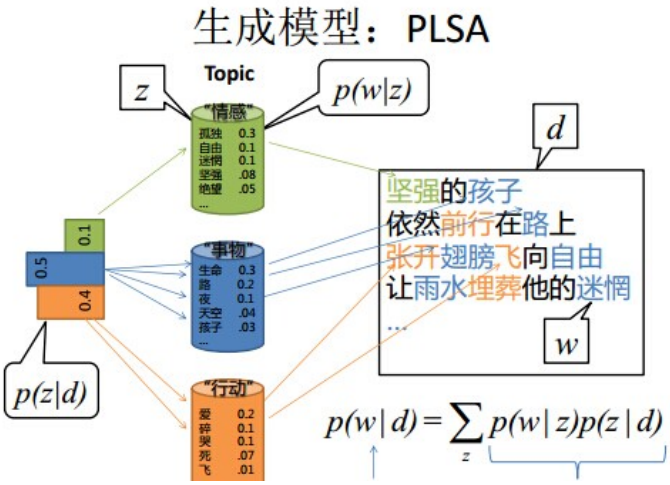
那 PLSA 跟 LDA 的区别在于什么地方呢？区别就在于：

- 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。

看到这，你可能凌乱了，你说面对多个主题或词，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。但现在你说主题分布和词分布本身也都是不确定的，这是怎么回事？没办法，谁叫 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。

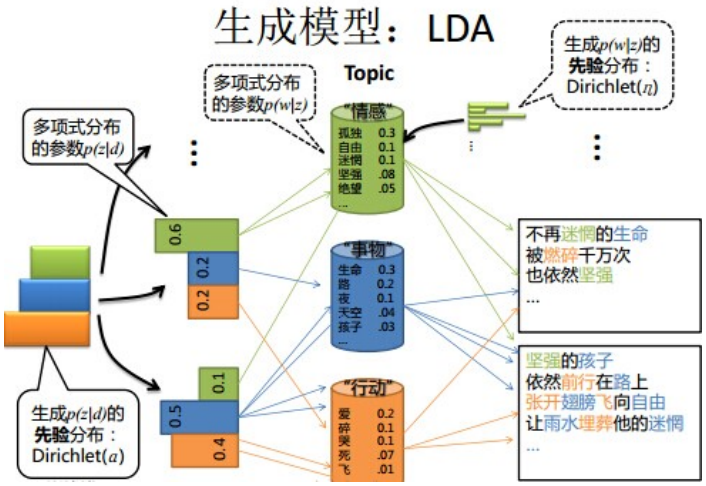
进一步，你会发现：

- 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（图截取自沈博 PPT 上）：



- 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。
    * 还是举文档 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（图截取自沈博 PPT 上）：



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），但至少明白哪些取值（0.6,0.7，0.8,0.9）更有可能，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

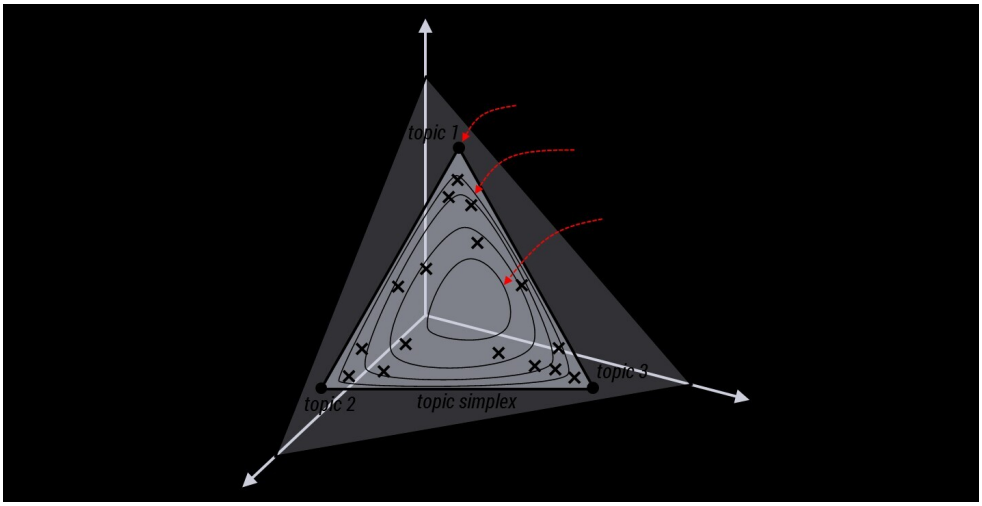
OK，相信已经解释清楚了。如果是在机器学习班上 face-to-face，更好解释和沟通

5.3.2 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。

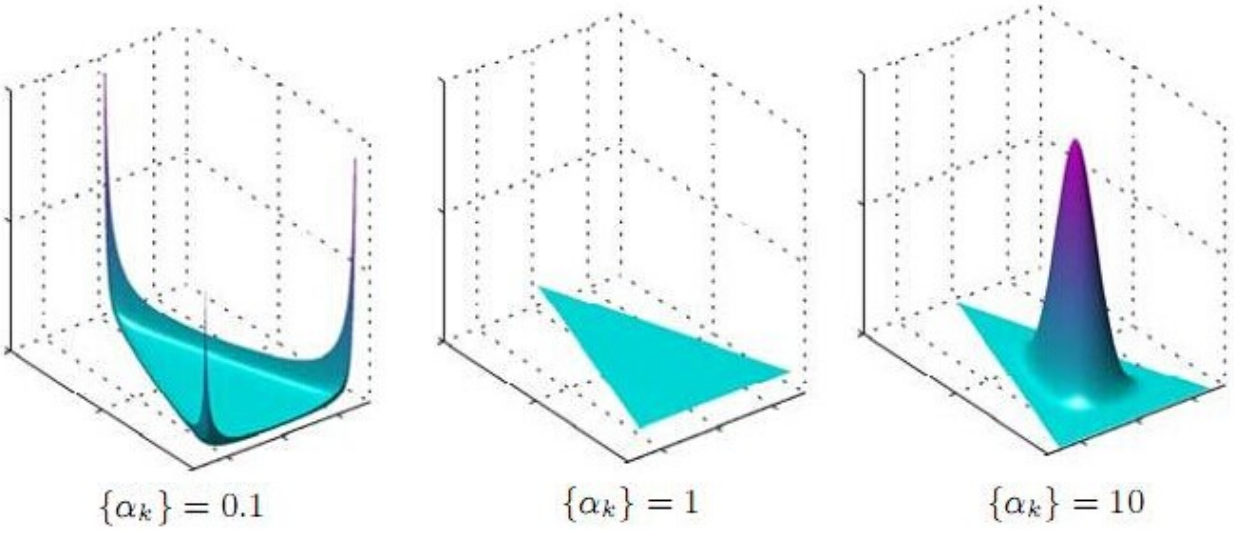
那么，dirichlet 先验到底是如何“随机”抽取主题分布的呢？

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：



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}。

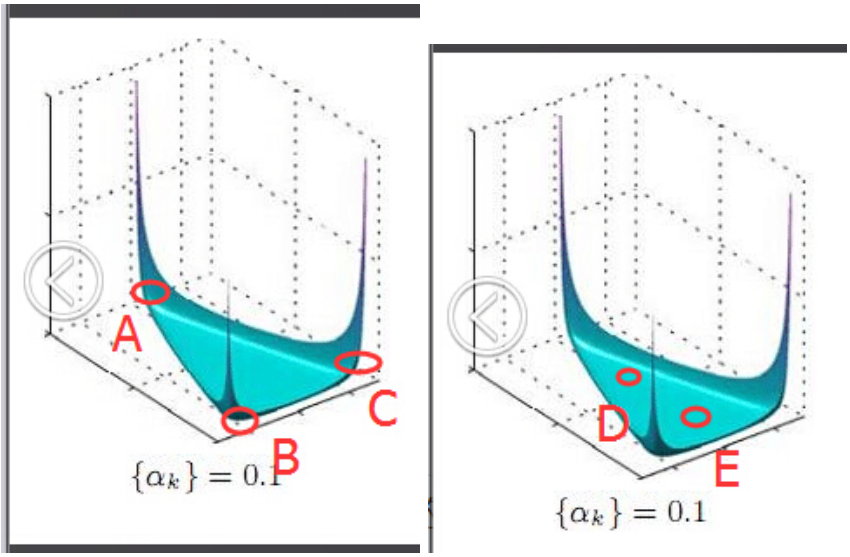
可以想象到，空间里有很多这样的点 (p1,p2,p3)，意味着有很多的主题分布可供选择，那 dirichlet 分布如何选择主题分布呢？把上面的斜三角形放倒，映射到底面的平面上，便得到如下所示的一些彩图（3 个彩图中，每一个点对应一个主题分布，高度代表某个主题分布被 dirichlet 分布选中的概率，且选不同的 α，**dirichlet** 分布会偏向不同的主题分布）：



我们来看下图中左边这个图，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

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。

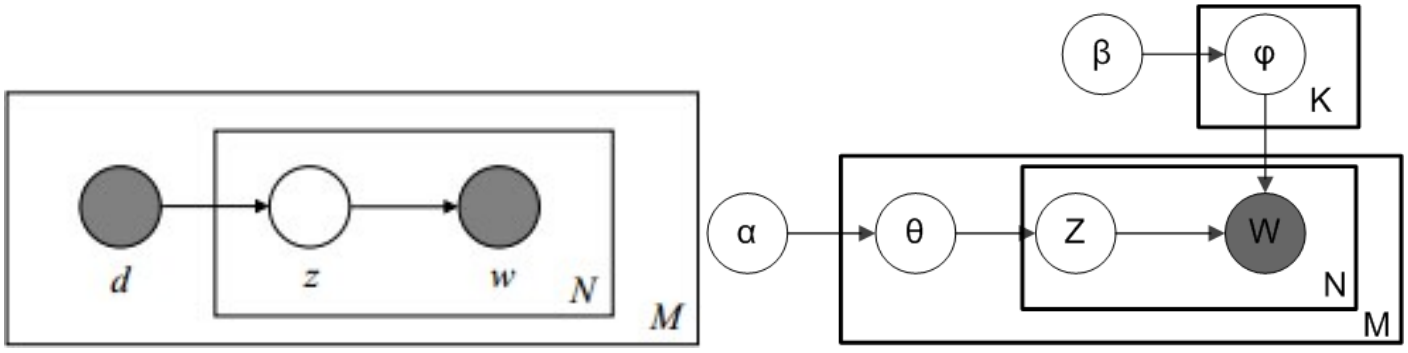
此外，就算说“随机”选主题也是根据主题分布来“随机”选取，这里的随机不是完全随机的意思，而是根据各个主题



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。

5.3.3 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）：



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。

- 其中，$\alpha \to \theta \to z$ 表示生成文档中的所有词对应的主题，显然 $\alpha \to \theta$ 对应的是 Dirichlet 分布，$\theta \to z$ 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：

$$\vec{\alpha} \underbrace{\longrightarrow}_{Dirichlet} \vec{\theta}_m \underbrace{\longrightarrow}_{Multinomial} \vec{z}_m$$

- 类似的，$\beta \to \phi \to w$，容易看出，此时 $\beta \to \phi$ 对应的是 Dirichlet 分布，$\phi \to w$ 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：

$$\vec{\beta} \underbrace{\longrightarrow}_{Dirichlet} \vec{\varphi}_k \underbrace{\longrightarrow}_{Multinomial} \vec{w}_{(k)}$$

5.3.4 pLSA 跟 LDA 参数估计方法的对比

上面对比了 pLSA 跟 LDA 生成文档的不同过程，下面，咱们反过来，假定文档已经产生，反推其主题分布。那么，它们估计未知参数所采用的方法又有什么不同呢？

- 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)}。

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 $\hat{\theta_{mk}}$、$\hat{\phi_{kt}}$）！

此外，不厌其烦的再插一句，在 LDA 中，主题分布和词分布本身都是多项分布，而由上文 3.2 节可知“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。

5.3.5 LDA 参数估计：Gibbs 采样

理清了 LDA 中的物理过程，下面咱们来看下如何学习估计。

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：

$$p(\vec{w_i}, \vec{z_i}, \vec{\theta_i}, \Phi|\vec{\alpha}, \vec{\beta}) = \prod_{n=1}^{N_i} p(w_{i,j}|\overrightarrow{\phi_{z_{i,j}}})p(z_{i,j}|\vec{\theta_i}) \cdot p(\vec{\theta_i}|\vec{\alpha}) \cdot p(\Phi|\vec{\beta})$$

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 $p(\vec{w}, \vec{z})$：

$$p(\vec{w}, \vec{z}|\vec{\alpha}, \vec{\beta}) = p(\vec{w}|\vec{z}, \vec{\beta})p(\vec{z}|\vec{\alpha})$$

其中，第一项因子 $p(\vec{w}|\vec{z}, \vec{\beta})$ 表示的是根据确定的主题 $\vec{z}$ 和词分布的先验分布参数 β 采样词的过程，第二项因子 $p(\vec{z}|\vec{\alpha})$ 是根据主题分布的先验分布参数采样 α 主题的过程，这两项因子是需要计算的两个未知参数。

由于这两个过程是独立的，所以下面可以分别处理，各个击破。

第一项因子 $p(\vec{w}|\vec{z}, \vec{\beta})$，可以根据确定的主题 $\vec{z}$ 和先验分布 β 取样得到的词分布 Φ 产生：

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

由于样本中的词服从参数为主题的独立多项分布，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：

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

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 4.1 节定义的 Dirichlet 分布的归一化系数 $\Delta(\vec{\alpha})$ 的公式

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

可得:

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

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 $p(\overrightarrow{z}|\overrightarrow{\alpha})$。类似于 $p(\overrightarrow{w}|\overrightarrow{z},\overrightarrow{\beta})$ 的步骤，先写出条件分布，然后分解成两部分的乘积:

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

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。

对主题分布 Θ 积分可得:

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

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

$$p(\overrightarrow{z},\overrightarrow{w}|\overrightarrow{\alpha},\overrightarrow{\beta})=\prod_{z=1}^{K}\frac{\Delta(\overrightarrow{n_z}+\overrightarrow{\beta})}{\Delta(\overrightarrow{\beta})}\cdot\prod_{m=1}^{M}\frac{\Delta(\overrightarrow{n_m}+\overrightarrow{\alpha})}{\Delta(\overrightarrow{\alpha})}$$

接下来，有了联合分布 $p(\overrightarrow{w},\overrightarrow{z})$，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 $p(\overrightarrow{z}|\overrightarrow{w})$（后验分布）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，$\overrightarrow{w}=\{w_i=t,\overrightarrow{w_{\neg i}}\}$，$\overrightarrow{z}=\{z_i=k,\overrightarrow{z_{\neg i}}\}$。

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:

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

且有：

$$
\begin{aligned}
p(z_i = k|\overrightarrow{\boldsymbol{z}_{\neg i}}, \overrightarrow{\boldsymbol{w}}) &\propto p(z_i = k, w_i = t|\overrightarrow{\boldsymbol{z}_{\neg i}}, \overrightarrow{\boldsymbol{w}_{\neg i}}) \\
&= \int p(z_i = k, w_i = t, \overrightarrow{\theta_m}, \overrightarrow{\phi_k}|\overrightarrow{\boldsymbol{z}_{\neg i}}, \overrightarrow{\boldsymbol{w}_{\neg i}})\mathrm{d}\overrightarrow{\theta_m}\mathrm{d}\overrightarrow{\phi_k} \\
&= \int p(z_i = k, \overrightarrow{\theta_m}|\overrightarrow{\boldsymbol{z}_{\neg i}}, \overrightarrow{\boldsymbol{w}_{\neg i}}) \cdot p(w_i = t, \overrightarrow{\phi_k}|\overrightarrow{\boldsymbol{z}_{\neg i}}, \overrightarrow{\boldsymbol{w}_{\neg i}})\mathrm{d}\overrightarrow{\theta_m}\mathrm{d}\overrightarrow{\phi_k} \\
&= \int p(z_i = k|\overrightarrow{\theta_m})p(\overrightarrow{\theta_m}|\overrightarrow{\boldsymbol{z}_{\neg i}}, \overrightarrow{\boldsymbol{w}_{\neg i}}) \cdot p(w_i = t|\overrightarrow{\phi_k})p(\overrightarrow{\phi_k}|\overrightarrow{\boldsymbol{z}_{\neg i}}, \overrightarrow{\boldsymbol{w}_{\neg i}})\mathrm{d}\overrightarrow{\theta_m}\mathrm{d}\overrightarrow{\phi_k} \\
&= \int p(z_i = k|\overrightarrow{\theta_m})Dir(\overrightarrow{\theta_m}|\overrightarrow{n_{m,\neg i}} + \overrightarrow{\alpha})\mathrm{d}\overrightarrow{\theta_m} \\
&\cdot \int p(w_i = t|\overrightarrow{\phi_k})Dir(\overrightarrow{\phi_k}|\overrightarrow{n_{k,\neg i}} + \overrightarrow{\alpha})\mathrm{d}\overrightarrow{\phi_k} \\
&= \int \theta_{mk}Dir(\overrightarrow{\theta_m}|\overrightarrow{n_{m,\neg i}} + \overrightarrow{\alpha})\mathrm{d}\overrightarrow{\theta_m} \cdot \int \phi_{kt}Dir(\overrightarrow{\phi_k}|\overrightarrow{n_{k,\neg i}} + \overrightarrow{\beta})\mathrm{d}\overrightarrow{\phi_k} \\
&= E(\theta_{mk}) \cdot E(\phi_{kt}) \\
&= \hat{\theta_{mk}}\hat{\phi_{kt}}
\end{aligned}
$$

最后一步，便是根据 Markov 链的状态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）：

$$p(\overrightarrow{\theta_m}|\overrightarrow{z_m}, \overrightarrow{\alpha}) = \frac{1}{Z_{\theta_m}}\prod_{n=1}^{N_m} p(z_{m,n}|\overrightarrow{\theta_m}) \cdot p(\overrightarrow{\theta_m}|\overrightarrow{\alpha}) = Dir(\overrightarrow{\theta_m}|\overrightarrow{n_m} + \overrightarrow{\alpha})$$

$$p(\overrightarrow{\phi_k}|\overrightarrow{z}, \overrightarrow{w}, \overrightarrow{\beta}) = \frac{1}{Z_{\phi_k}}\prod_{\{i:z_i=k\}} p(w_i|\overrightarrow{\phi_k}) \cdot p(\overrightarrow{\phi_k}|\overrightarrow{\beta}) = Dir(\overrightarrow{\phi_k}|\overrightarrow{n_k} + \overrightarrow{\beta})$$

其中，$\overrightarrow{n_m}$ 是构成文档 m 的主题数向量，$\overrightarrow{n_k}$ 是构成主题 k 的词项数向量。

此外，别忘了上文中 3.4 节所述的 Dirichlet 的一个性质，如下：“如果 $\overrightarrow{p} \sim Dir(\overrightarrow{t}|\overrightarrow{\alpha})$，同样可以证明有下述结论成立：

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

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

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

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：

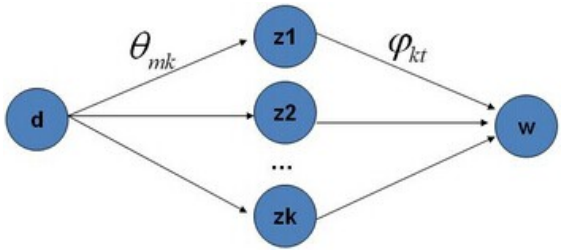
$$\phi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^V n_k^{(t)} + \beta_t}$$

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

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

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

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：



何等奇妙，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。

6 读者微评

本文发表后，部分热心的读者在微博上分享了他们自己理解 LDA 的心得，也欢迎更多朋友分享你的理解心得（比如评论在本文下，或评论在微博上），从而在分享、讨论的过程中让更多人可以更好的理解：

1. @SiNZeRo：lda 如果用 em 就是 map 估计了. lda 本意是要去找后验分布然后拿后验分布做 bayesian 分析. 比如 theta 的期望. 而不是把先验作为正则化引入。最后一点 gibbs sampling 其实不是求解的过程是去 explore 后验分布去采样用于求期望.

2. @ 研究者 July：好问题好建议，这几天我陆续完善下！//@ 帅广应 s：LDA 这个东西该怎么用？可以用在哪些地方？还有就是 Gibbs 抽样的原理是什么？代码怎么实现？如果用 EM 来做，代码怎么实现？LDA 模型的变形和优化有哪些？LDA 不适用于解决哪类的问题？总之，不明白怎么用，参数怎么调优？

3. @xiangnanhe：写的很好，4.1.3 节中的那两个图很赞，非常直观的理解了 LDA 模型加了先验之后在学参数的时候要比 PLSI 更灵活；PLSI 在学参数的过程中比较容易陷入 local minimum 然后 overfitting。

4. @asker2：无论是 pLSA 中，还是 LDA 中，主题分布和词分布本身是固定的存在，但都未知。pLSA 跟 LDA 的区别在于，去探索这两个未知参数的方法或思想不一样。pLSA 是求到一个能拟合文本最好的参数（分布），这个值就认为是真实的参数。但 LDA 认为，其实我们没法去完全求解出主题分布、词分布到底是什么参数，我们只能把它们当成随机变量，通过缩小其方差（变化度）来尽量让这个随机变量变得更“确切”。换言之，我们不再求主题分布、词分布的具体值，而是通过这些分布生成的观测值（即实际文本）来反推分布的参数的范围，即在什么范围比较可能，在什么范围不太可能。所以，其实这就是一种贝叶斯分析的思想，虽然无法给出真实值具体是多少，但可以按照经验给一个相对合理的真实值服从的先验分布，然后从先验出发求解其后验分布。

7 参考文献与推荐阅读

1. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I. Latent Dirichlet allocation（LDA 原始论文）: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf。
2. Blei. Probabilistic Topic Models: http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf，一网友的翻译: http://www.cnblogs.com/siegfang/archive/2013/01/30/2882391.html;
3. 一堆 wikipedia，比如隐含狄利克雷分布 LDA 的 wiki: http://zh.wikipedia.org/wiki/%E9%9A%90%E5%90%AB%E7%8B%84%E5%88%A9%E5%85%8B%E9%9B%B7%E5%88%86%E5%B8%83，狄利克雷分布的 wiki: http://zh.wikipedia.org/wiki/%E7%8B%84%E5%88%A9%E5%85%8B%E9%9B%B7%E5%88%86%E5%B8%83;
4. 从贝叶斯方法谈到贝叶斯网络 ;
5. rickjin 的 LDA 数学八卦（力荐，本文部分图片和公式来自于此文档）网页版: http://www.flickering.cn/tag/lda/, PDF 版: http://emma.memect.com/t/9756da9a47744de993d8df13a26e04e38286c9bc1c5a0d2b259c4564c6613298/LDA;
6. Thomas Hofmann.Probabilistic Latent Semantic Indexing（pLSA 原始论文）: http://cs.brown.edu/~th/papers/Hofmann-SIGIR99.pdf;
7. Gregor Heinrich.Parameter estimation for text analysis（关于 Gibbs 采样最精准细致的论述）: http://www.arbylon.net/publications/text-est.pdf;
8. Probabilistic latent semantic analysis (pLSA): http://blog.tomtung.com/2011/10/plsa/http://blog.tomtung.com/2011/10/plsa/。
9. 《概率论与数理统计教程第二版茆诗松等人著》，如果忘了相关统计分布，建议复习此书或此文第二部分;
10. 《支持向量机通俗导论：理解 SVM 的三层境界》，第二部分关于拉格朗日函数的讨论;
11. 机器学习班第 11 次课上，邹博讲 EM & GMM 的 PPT: http://pan.baidu.com/s/1i3zgmzF;
12. 机器学习班第 12 次课上，邹博讲主题模型 LDA 的 PPT: http://pan.baidu.com/s/1jGghtQm;
13. 主题模型之 pLSA: http://blog.jqian.net/post/plsa.html;
14. 主题模型之 LDA: http://blog.jqian.net/post/lda.html;
15. 搜索背后的奥秘——浅谈语义主题计算: http://www.semgle.com/search-engine-algorithms-mystery-behind-search
16. LDA 的 EM 推导: http://www.cnblogs.com/hebin/archive/2013/04/25/3043575.html;
17. Machine Learning 读书会第 8 期上，沈博讲主题模型的 PPT: http://vdisk.weibo.com/s/zrFL6OXKgKMAf;
18. Latent Dirichlet Allocation（LDA）- David M.Blei: http://www.xperseverance.net/blogs/2012/03/17/;
19. 用 GibbsLDA 做 Topic Modeling: http://weblab.com.cityu.edu.hk/blog/luheng/2011/06/24/%E7%94%A8gibbslda%E5%81%9Atopic-modeling/#comment-87;
20. 主题模型在文本挖掘中的应用: http://net.pku.edu.cn/~zhaoxin/Topic-model-xin-zhao-wayne.pdf;
21. 二项分布和多项分布，beta 分布的对比: http://www.cnblogs.com/wybang/p/3206719.html;

22. LDA 简介：http://cos.name/2010/10/lda_topic_model/；

23. LDA 的相关论文、工具库：http://site.douban.com/204776/widget/notes/12599608/note/287085506/；

24. 一个网友学习 LDA 的心得：http://www.xuwenhao.com/2011/03/20/suggestions-for-programmers-to-learn-lda/；

25. http://blog.csdn.net/hxxiaopei/article/details/7617838；

26. 主题模型 LDA 及其在微博推荐 & 广告算法中的应用：http://www.wbrecom.com/?p=136；

27. LDA 发明人之一 Blei 写的毕业论文：http://www.cs.princeton.edu/~blei/papers/Blei2004.pdf；

28. LDA 的一个 C 实现：http://www.cs.princeton.edu/~blei/lda-c/index.html；

29. LDA 的一些其它资料：http://www.xperseverance.net/blogs/2012/03/657/。

8 后记

这个 LDA 的笔记从 11 月 17 日下午开始动笔，到 21 日基本写完，25 日基本改完，前前后后，基本写完 + 基本改完，总共花了近 10 天的时间，后面还得不断完善。前 5 天就像在树林里中行走，要走的大方向非常明确，但在选取哪条小道上则颇费了一番周折，但当最后走出了树林，登上山顶，俯瞰整个森林时，奥，原来它就长这样，会有一种彻爽的感觉！而后 5 天，则慢慢开始接近 LDA 的本质：pLSA 的贝叶斯版本。

写作过程艰难但结果透彻，也希望读者能享受到其中。

最后，再次感谢本文最主要的参考：LDA 原始论文、pLSA 原始论文、LDA 数学八卦、机器学习班第 12 次课主题模型 PPT，和 Parameter estimation for text analysis 等等的作者们（本文中大部分的图片、公式截取自这些参考资料上），因为有他们的创造或分享，我才有机会理解和再加工 LDA，最终让本文得以成文。

后续几天会不断修改完善本文，若有任何问题，可在本文下评论，thanks。

July、二零一四年十一月二十一日。