**主题建模介绍**

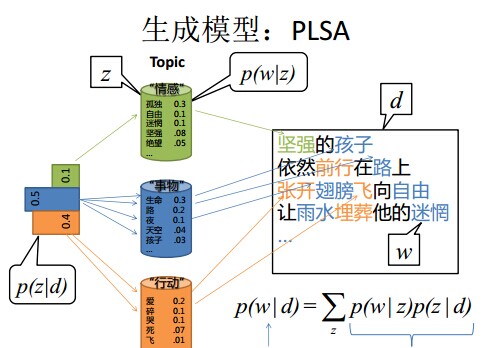
# 一 简述

* 定义：主题建模，就是将文档集中每篇文档的主题，以概率分布的形式给出。
* 实现：1.定义文档生成流程。2.根据文档生成流程，反推出文档的主题分布。
* 类型：
* pLSA (Probabilistic Latent Semantic Analysis)
* LDA (Latent Dirichlet Allocation)

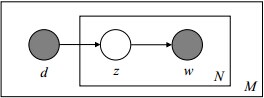
# 二 pLSA主题建模

## 1文档生成（pLSA）

1. pLSA文档生成过程：
2. 以一定概率选出主题z。（主题分布，是一个多项分布）
3. 在选定主题中，按一定的概率挑出词w。（词分布，也是一个多项分布）
4. 不停重复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。
5. pLSA文档生成过程示例：



1. pLSA文档生成过程的图模型：



注：

* 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。

所以，pLSA文档生成过程可定义为：

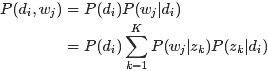
1. 以[http://latex.codecogs.com/gif.latex?P(d_i)](http://www.codecogs.com/eqnedit.php?latex=P(d_i))的概率选中文档[http://latex.codecogs.com/gif.latex?d_i](http://www.codecogs.com/eqnedit.php?latex=d_i)；
2. 以[http://latex.codecogs.com/gif.latex?P(z_k|d_i)](http://www.codecogs.com/eqnedit.php?latex=P(z_k|d_i))的概率选中主题[http://latex.codecogs.com/gif.latex?z_k](http://www.codecogs.com/eqnedit.php?latex=z_k)；
3. 以[http://latex.codecogs.com/gif.latex?P(w_j|z_k)](http://www.codecogs.com/eqnedit.php?latex=P(w_j|z_k))的概率产生一个单词。

而[http://latex.codecogs.com/gif.latex?P(z_k|d_i)](http://www.codecogs.com/eqnedit.php?latex=P(z_k|d_i))和[http://latex.codecogs.com/gif.latex?P(w_j|z_k)](http://www.codecogs.com/eqnedit.php?latex=P(w_j|z_k))分布对应了两组Multinomial 分布，**目的就是需要估计这两组分布的参数。**

## 2 根据文档生成流程，反推其主题分布（pLSA）

1. 似然函数

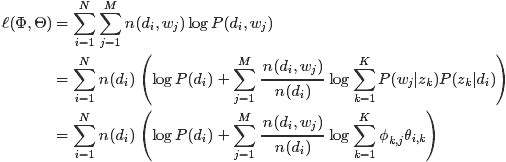
https://img-blog.csdn.net/20141124221914437（公式2.1）

（公式2.2）

https://img-blog.csdn.net/20141212232331064（公式2.3）

https://img-blog.csdn.net/20141212233703984（公式2.4）

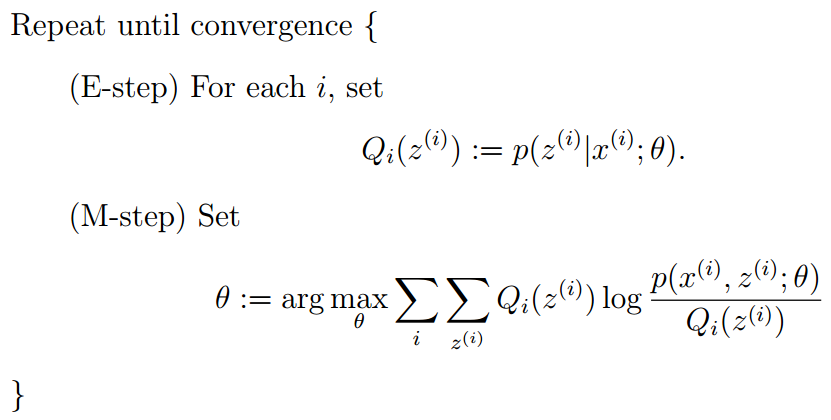
而后，用https://img-blog.csdn.net/20141119231428328表示词项https://img-blog.csdn.net/20141119231439125出现在主题https://img-blog.csdn.net/20141119231347433中的概率，即https://img-my.csdn.net/uploads/201411/22/1416585670_2090.jpg，用https://img-blog.csdn.net/20141119231757234表示主题https://img-blog.csdn.net/20141119231347433出现在文档https://img-blog.csdn.net/20141119231734648中的概率，即https://img-my.csdn.net/uploads/201411/22/1416585677_1570.jpg，从而把https://img-blog.csdn.net/20141119231257263转换成了“主题-词项”矩阵Φ（主题生成词），把https://img-blog.csdn.net/20141119231307918转换成了“文档-主题”矩阵Θ（文档生成主题）

（公式2.5）

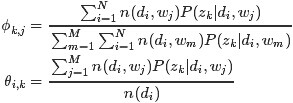
由于https://img-blog.csdn.net/20141119004850551和https://img-blog.csdn.net/20141119004915308未知，所以我们用EM算法去估计https://img-blog.csdn.net/20141119132914266这个参数的值。

1. EM做参数估计

EM算法，首先随机选取一个值去初始化待估计的值https://img-blog.csdn.net/20141119142305046，然后不断迭代寻找更优的https://img-blog.csdn.net/20141119142335046使得其似然函数likelihood https://img-blog.csdn.net/20141119142434921比原来的https://img-blog.csdn.net/20141119142450410要大。



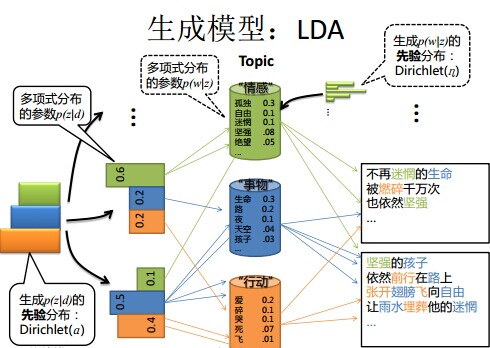
推导略。最终可求解出https://img-blog.csdn.net/20141119231428328、https://img-blog.csdn.net/20141119231757234：



# 三、LDA主题建模

## 1文档生成（LDA）

1. LDA文档生成过程：
2. 按照先验概率https://img-blog.csdn.net/20141119004830562选择一篇文档https://img-blog.csdn.net/20141119004838164
3. 从狄利克雷分布（即Dirichlet分布）https://img-blog.csdn.net/20141117160438989中取样生成文档 https://img-blog.csdn.net/20141119004838164的主题分布https://img-blog.csdn.net/20141117160452327，换言之，主题分布https://img-blog.csdn.net/20141117160452327由超参数为https://img-blog.csdn.net/20141117160438989的Dirichlet分布生成
4. 从主题的多项式分布https://img-blog.csdn.net/20141117160452327中取样生成文档https://img-blog.csdn.net/20141119004838164第 j 个词的主题https://img-blog.csdn.net/20141117160518098
5. 从狄利克雷分布（即Dirichlet分布）https://img-blog.csdn.net/20141117160531515中取样生成主题https://img-blog.csdn.net/20141117160518098对应的词语分布https://img-blog.csdn.net/20141117160613962，换言之，词语分布https://img-blog.csdn.net/20141117160613962由参数为https://img-blog.csdn.net/20141117160531515的Dirichlet分布生成
6. 从词语的多项式分布https://img-blog.csdn.net/20141117160613962中采样最终生成词语https://img-blog.csdn.net/20141117160656067
7. LDA文档生成过程示例：



1. LDA pLSA文档生成过程的图模型：



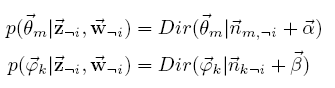
综上，对于一篇文档d中的每一个单词，LDA根据先验知识https://img-blog.csdn.net/20141117160438989确定某篇文档的主题分布θ（多项分布），然后从该文档所对应的主题分布 θ 中抽取一个主题z，接着根据先验知识https://img-blog.csdn.net/20141117160531515确定当前主题的词语分布ϕ（多项分布），然后从主题z所对应的词分布ϕ中抽取一个单词w。然后将这个过程重复N次，就产生了文档d。

## 2 根据文档生成流程，反推其主题分布（LDA参数估计）

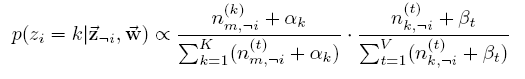
* 由于LDA把要估计的主题分布和词分布看作是其先验分布（Dirichlet）的随机变量:



* 主题分布和词分布的先验分布（Dirichlet）是事先给定的，所以LDA就是要去求对应的后验分布。



* Gibbs采样做参数估计：



得到词对应的主题的概率估计。

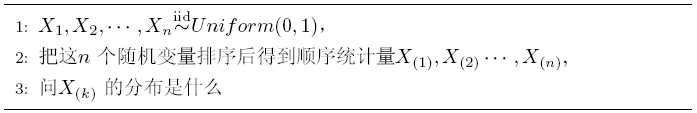
注：LDA原始论文使用变分-EM算法估计未知参数：

把w当做观察变量，θ和z当做隐藏变量，就可以通过EM算法学习出α和β，求解过程中遇到后验概率p(θ,z|w)无法直接求解，需要找一个似然函数下界来近似求解，原文使用基于分解（factorization）假设的变分法（varialtional inference）进行计算，用到了EM算法。每次E-step输入α和β，计算似然函数，M-step最大化这个似然函数，算出α和β，不断迭代直到收敛。

# 四 相关数学知识

## Beta分布

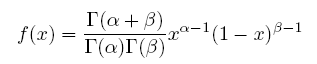
问题模型：



推导过程略。利用Gammar函数性质可推导出：

，其中取：



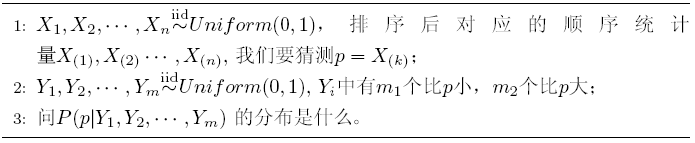


所以，第k个随机变量p对应的分布可表示为：



## Beta-Binomial共轭

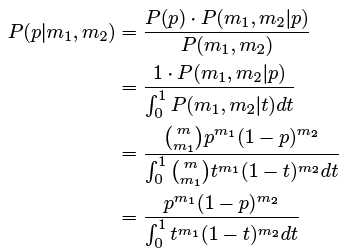
问题模型：



问题描述：

1. 为了猜测https://img-blog.csdn.net/20141117182441741，在获得一定的观测数据前，我们对https://img-blog.csdn.net/20141117184123125的认知是：https://img-blog.csdn.net/20141118100645570，此称为https://img-blog.csdn.net/20141117184123125的先验分布；
2. 然后为了获得这个结果“ https://img-blog.csdn.net/20141117183329656中有https://img-blog.csdn.net/20141117183159618个比p小，https://img-blog.csdn.net/20141117183207590个比https://img-blog.csdn.net/20141117184123125大”，针对https://img-blog.csdn.net/20141117183329656是做了https://img-blog.csdn.net/20141117184532817次贝努利实验，所以https://img-blog.csdn.net/20141117183159618服从二项分布https://img-blog.csdn.net/20141117184624234；
3. 在给定了来自数据提供的https://img-blog.csdn.net/20141117184713218的知识后，https://img-blog.csdn.net/20141117184123125的后验分布变为https://img-blog.csdn.net/20141117184810158。

推导出第k个随机变量p的后验概率：

正好是：



根据贝叶斯定理，符合共轭关系：

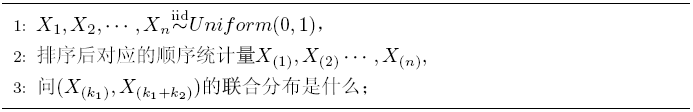
https://img-blog.csdn.net/20141117185216670

更一般的，对于非负实数https://img-blog.csdn.net/20141117181123550和https://img-blog.csdn.net/20141117181134937，我们有如下关系

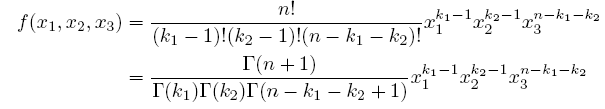
https://img-blog.csdn.net/20141117185325671,（式2.1）

## Dirichlet分布

问题模型：



求解过程略。最终推导出：



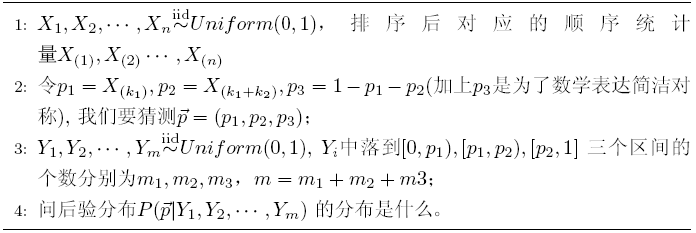
正好是三维形式的Dirichlet分布：





## Dirichlet-Multinomial共轭

问题模型：



问题描述：

1. 我们要猜测参数https://img-blog.csdn.net/20141118220412816，其先验分布为https://img-blog.csdn.net/20141118220433171；
2. 数据Yi落到三个区间https://img-blog.csdn.net/20141118180754809，https://img-blog.csdn.net/20141118180800191，https://img-blog.csdn.net/20141118180806244 的个数分别为https://img-blog.csdn.net/20141118220524935，所以https://img-blog.csdn.net/20141118220449991服从多项分布https://img-blog.csdn.net/20141118220553625
3. 在给定了来自数据提供的知识https://img-blog.csdn.net/20141118220601657后，https://img-blog.csdn.net/20141118220609473的后验分布变为https://img-blog.csdn.net/20141118220620315

推导过程略。

最终计算出目标向量的后验概率：

https://img-blog.csdn.net/20141118220632031

令https://img-blog.csdn.net/20141118221128562，可把https://img-blog.csdn.net/20141118221159725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：

https://img-blog.csdn.net/20141118220737504,（式2.2）