概率语言模型及其变形系列

LDA 及 Gibbs Sampling

yangliuyx@gmail.com

12/20/2012

本系列博文介绍常见概率语言模型及其变形模型，主要总结 PLSA、LDA 及 LDA 的变形模型及参数 Inference 方法.

概率语言模型及其变形系列-LDA 及 Gibbs Sampling

yangliuyx@gmail.com

December 20th 2012

本系列博文介绍常见概率语言模型及其变形模型，主要总结 PLSA、LDA 及 LDA 的变形模型及参数 Inference 方法。初步计划内容如下

第一篇：PLSA 及 EM 算法

第二篇：LDA 及 Gibbs Samping

第三篇：LDA 变形模型-Twitter LDA，TimeUserLDA，ATM，Labeled-LDA，MaxEnt-LDA 等

第四篇：基于变形 LDA 的 paper 分类总结

第二篇 LDA 及 Gibbs Sampling

1. LDA 概要

LDA 是由 Blei, Ng, Jordan 2002 年发表于 JMLR 的概率语言模型，应用到文本建模范畴，就是对文本进行"隐性语义分析"（LSA），目的是要以无指导学习的方法从文本中发现隐含的语义维度-即"Topic"或者"Concept"。隐性语义分析的实质是要利用文本中词项(term)的共现特征来发现文本的 Topic 结构，这种方法不需要任何关于文本的背景知识。文本的隐性语义表示可以对"一词多义"和"一义多词"的语言现象进行建模，这使得搜索引擎系统得到的搜索结果与用户的 query 在语义层次上 match，而不是仅仅只是在词汇层次上出现交集。

2. 概率基础

2.1 随机生成过程及共轭分布

要理解 LDA 首先要理解随机生成过程。用随机生成过程的观点来看，文本是一系列服从一定概率分布的词项的样本集合。最常用的分布就是 Multinomial 分布，即多项分布，这个分布是二项分布拓展到 K 维的情况，比如投掷骰子实验，N 次实验结果服从 K=6 的多项分布。相应的，二项分布的先验 Beta 分布也拓展到 K 维，称为 Dirichlet 分布。在概率语言模型中，通常为 Multinomial 分布选取的先验分布是 Dirichlet 分布，因为它们是共轭分布，可以带来计算上的方便性。什么是共轭分布呢？在文本语言模型的参数估计-最大似然估计、MAP 及贝叶斯估计一文中我们可以看到，当我们为二项分布的参数 p 选取的先验分布是 Beta 分布时，以 p 为参数的二项分布用贝叶斯估计得到的后验概率仍然服从 Beta 分布，由此我们说二项分布和 Beta 分布是共轭分布。这就是共轭分布要满足的性质。在 LDA 中，每个文档中词的 Topic 分布服从 Multinomial 分布，其先验选取共轭先验即 Dirichlet 分布；每个 Topic 下词的分布服从 Multinomial 分布，其先验也同样选取共轭先验即 Dirichlet 分布。

2.2 Multinomial 分布和 Dirichlet 分布

上面从二项分布和 Beta 分布出发引出了 Multinomial 分布和 Dirichlet 分布。这两个分布在概率语言模型中很常用，让我们深入理解这两个分布。Multinomial 分布的分布律如下

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

多项分布来自 N 次独立重复实验，每次实验结果可能有 K 种，式子中$\vec{n}$为实验结果向量，N 为实验次数，$\vec{p}$为出现每种实验结果的概率组成的向量，这个公式给出了出现所有实验结果的概率计算方法。当 K=2 时就是二项分布，K=6 时就是投掷骰子实验。很好理解，前面的系数其实是枚举实验结果的不同出现顺序，即

$$\frac{N!}{\prod_{i=1}^{K} n^{(k)}!}$$

后面表示第 K 种实验结果出现了$n^{(k)}$次，所以是概率的相应次幂再求乘积。但是如果我们不考虑文本中词出现的顺序性，这个系数就是 1。 本文后面的部分可以看出这一点。显然有$\vec{p}$各维之和为 1，所有$n^{(k)}$之和为 N。

Dirichlet 分布可以看做是“分布之上的分布”，从 Dirichlet 分布上 Draw 出来的每个样本就是多项分布的参数向量$\vec{p}$。其分布律为

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

$\vec{\alpha}$为 Dirichlet 分布的参数，在概率语言模型中通常会根据经验给定，由于是参数向量$\vec{p}$服从分布的参数，因此称为“hyperparamer”。$\Delta(\vec{\alpha})$是 Dirichlet delta 函数，可以看做是 Beta 函数拓展到 K 的情况，但是在有的文献中也直接写成$B(\vec{\alpha})$。根据 Dirichlet 分布在$\vec{p}$上的积分为 1（概率的基本性质），我们可以得到一个重要的公式

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

这个公式在后面 LDA 的参数 Inference 中经常使用。下图给出了一个 Dirichlet 分布的实例

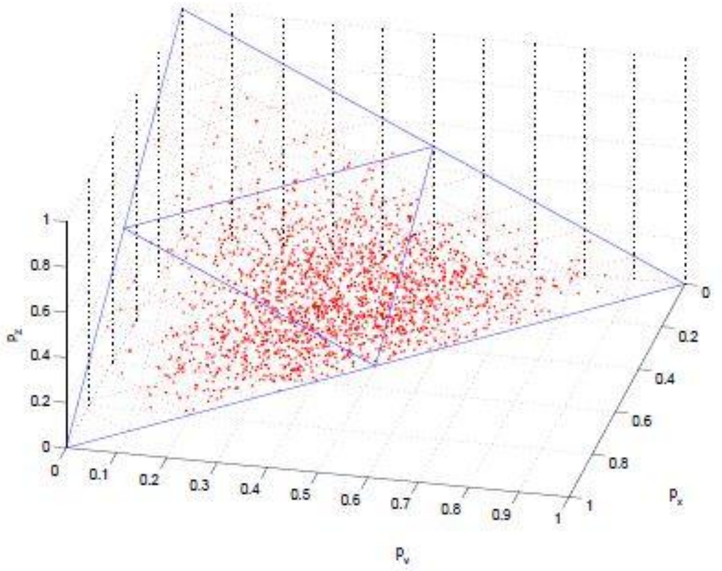
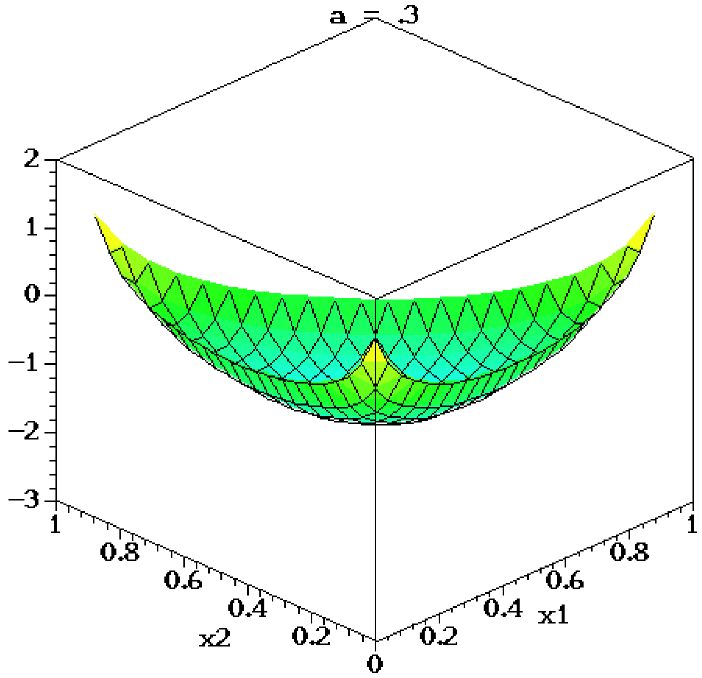


Fig. 3. 2000 samples from a Dirichlet distribution Dir(4, 4, 2). The plot shows that all samples are on a simplex embedded in the three-dimensional space, due to the constraint $\sum_k p_k = 1$.

在许多应用场合，我们使用对称 Dirichlet 分布，其参数是两个标量：维数 K 和参数向量各维均值 $\alpha = \frac{\sum \alpha_k}{K}$. 其分布律如下

$$
\begin{aligned}
p(\vec{p}|\alpha, K) = \mathrm{Dir}(\vec{p}|\alpha, K) &\triangleq \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} p_k^{\alpha-1} \\
&\triangleq \frac{1}{\Delta_K(\alpha)} \prod_{k=1}^{K} p_k^{\alpha-1}, \quad \Delta_K(\alpha) = \frac{\Gamma(\alpha)^K}{\Gamma(K\alpha)}.
\end{aligned}
$$

关于 Dirichlet 分布，维基百科上有一张很有意思的图如下



这个图将 Dirichlet 分布的概率密度函数取对数,并且使用对称 Dirichlet 分布，取 K=3，也就是有两个独立参数 x_1, x_2，分别对应图中的两个坐标轴，第三个参数始终满足$x_3 = 1 - x_1 - x_2$ 且 $\alpha_1 = \alpha_2 = \alpha_3 = \alpha$，图中反映的是$\alpha$从 0.3 变化到 2.0 的概率对数值的变化情况。

3 unigram model

我们先介绍比较简单的 unigram model。其概率图模型图示如下

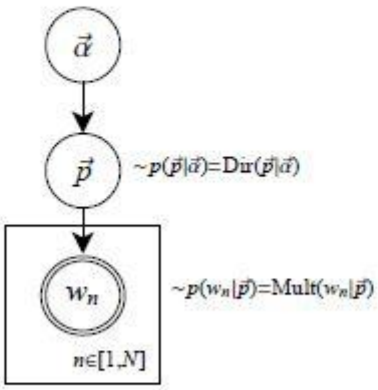


Fig. 4. Bayesian network of the Dirichlet–multinomial unigram model.

关于概率图模型尤其是贝叶斯网络的介绍可以参见 Stanford 概率图模型（Probabilistic Graphical Model）— 第一讲 贝叶斯网络基础一文。简单的说，贝叶斯网络是一个有向无环图，图中的结点是随机变量，图中的有向边代表了随机变量的条件依赖关系。unigram model 假设文本中的词服从 Multinomial 分布，而 Multinomial 分布的先验分布为 Dirichlet 分布。图中双线圆圈w_n表示我们在文本中观察到的第 n 个词，$n \in [1, N]$表示文本中一共有 N 个词。加上方框表示重复，就是说一共有 N 个这样的随机变量w_n。$\vec{p}$和$\vec{\alpha}$是隐含未知变量，分别是词服从的 Multinomial 分布的参数和该 Multinomial 分布的先验 Dirichlet 分布的参数。一般$\vec{\alpha}$由经验事先给定，$\vec{p}$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。

4 LDA

理解了 unigram model 之后，我们来看 LDA。我们可以假想有一位大作家，比如莫言，他现在要写 m 篇文章，一共涉及了 K 个 Topic，每个 Topic 下的词分布为一个从参数为$\vec{\beta}$的 Dirichlet 先验分布中 sample 出来的 Multinomial 分布（注意词典由 term 构成，每篇文章由 word 构成，前者不能重复，后者可以重复）。对于每篇文章，他首先会从一个泊松分布中 sample 一个值作为文章长度，再从一个参数为$\vec{\alpha}$的 Dirichlet 先验分布中 sample 出一个 Multinomial 分布作为该文章里面出现每个 Topic 下词的概率；当他想写某篇文章中的第 n 个词的时候，首先从该文章中出现每个 Topic 下词的 Multinomial 分布中 sample 一个 Topic，然后再在这个 Topic 对应的词的 Multinomial 分布中 sample 一个词作为他要写的词。不断重复这个随机生成过程，直到他把 m 篇文章全部写完。这就是 LDA 的一个形象通俗的解释。用数学的语言描述就是如下过程

// topic plate
for all topics $k \in [1, K]$ **do**
  sample mixture components $\vec{\varphi}_k \sim \mathrm{Dir}(\vec{\beta})$
// document plate:
for all documents $m \in [1, M]$ **do**
  sample mixture proportion $\vec{\vartheta}_m \sim \mathrm{Dir}(\vec{\alpha})$
  sample document length $N_m \sim \mathrm{Poiss}(\xi)$
  // word plate:
  for all words $n \in [1, N_m]$ in document m **do**
    sample topic index $z_{m,n} \sim \mathrm{Mult}(\vec{\vartheta}_m)$
    sample term for word $w_{m,n} \sim \mathrm{Mult}(\vec{\varphi}_{z_{m,n}})$

Fig. 7. Generative model for latent Dirichlet allocation.

转化成概率图模型表示就是

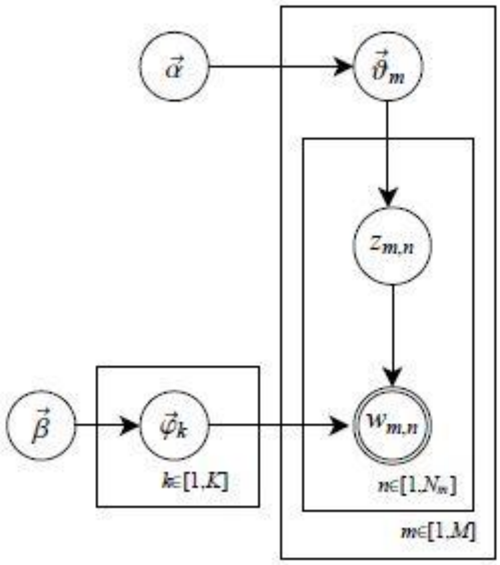


Fig. 6. Bayesian network of latent Dirichlet allocation.

图中 K 为主题个数，M 为文档总数，N_m是第 m 个文档的单词总数。$\vec{\beta}$ 是每个 Topic 下词的多项分布的 Dirichlet 先验参数，$\vec{\alpha}$ 是每个文档下 Topic 的多项分布的 Dirichlet 先验参数。$z_{m,n}$是第 m 个文档中第 n 个词的主题，$w_{m,n}$是 m 个文档中的第 n 个词。剩下来的两个隐含变量$\vec{\theta}_m$和$\vec{\phi}_k$分别表示第 m 个文档下的 Topic 分布和第 k 个 Topic 下词的分布，前者是 k 维(k 为 Topic 总数)向量，后者是 v 维向量（v 为词典中 term 总数）。

给定一个文档集合，$w_{m,n}$是可以观察到的已知变量，$\vec{\alpha}$和$\vec{\beta}$是根据经验给定的先验参数，其他的变量$z_{m,n}$，$\vec{\theta}_m$和$\vec{\phi}_k$都是未知的隐含变量，也是我们需要根据观察到的变量来学习估计的。根据 LDA 的图模型，我们可以写出所有变量的联合分布

$$p(\vec{w}_m, \vec{z}_m, \vec{\vartheta}_m, \underline{\Phi} \mid \vec{\alpha}, \vec{\beta}) = \overbrace{\underbrace{\prod_{n=1}^{N_m} p(w_{m,n} \mid \vec{\varphi}_{z_{m,n}}) p(z_{m,n} \mid \vec{\vartheta}_m)}_{\text{word plate}} \cdot p(\vec{\vartheta}_m \mid \vec{\alpha})}^{\text{document plate (1 document)}} \cdot \underbrace{p(\underline{\Phi} \mid \vec{\beta})}_{\text{topic plate}} .$$

那么一个词$w_{m,n}$初始化为一个 term t 的概率是

$$p(w_{m,n}=t \mid \vec{\vartheta}_m, \underline{\Phi}) = \sum_{k=1}^{K} p(w_{m,n}=t \mid \vec{\varphi}_k) p(z_{m,n}=k \mid \vec{\vartheta}_m) ,$$

也就是每个文档中出现 topic k 的概率乘以 topic k 下出现 term t 的概率，然后枚举所有 topic 求和得到。整个文档集合的似然函数就是

$$p(\mathcal{W}|\underline{\Theta}, \underline{\Phi}) = \prod_{m=1}^{M} p(\vec{w}_m|\vec{\vartheta}_m, \underline{\Phi}) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} p(w_{m,n}|\vec{\vartheta}_m, \underline{\Phi})$$

5 用 Gibbs Sampling 学习 LDA

5.1 Gibbs Sampling 的流程

从第 4 部分的分析我们知道，LDA 中的变量$z_{m,n}$，$\vec{\theta}_m$和$\vec{\phi}_k$都是未知的隐含变量，也是我们需要根据观察到的文档集合中的词来学习估计的，那么如何来学习估计呢？这就是概率图模型的 Inference 问题。主要的算法分为 exact inference 和 approximate inference 两类。尽管 LDA 是最简单的 Topic Model， 但是其用 exact inference 还是很困难的，一般我们采用 approximate inference 算法来学习 LDA 中的隐含变量。比如 LDA 原始论文 Blei02 中使用的 mean-field variational expectation maximisation 算法和 Griffiths02 中使用的 Gibbs Sampling，其中 Gibbs Sampling 更为简单易懂。

Gibbs Sampling 是 Markov-Chain Monte Carlo 算法的一个特例。这个算法的运行方式是每次选取概率向量的一个维度，给定其他维度的变量值 Sample 当前维度的值。不断迭代，直到收敛输出待估计的参数。可以图示如下

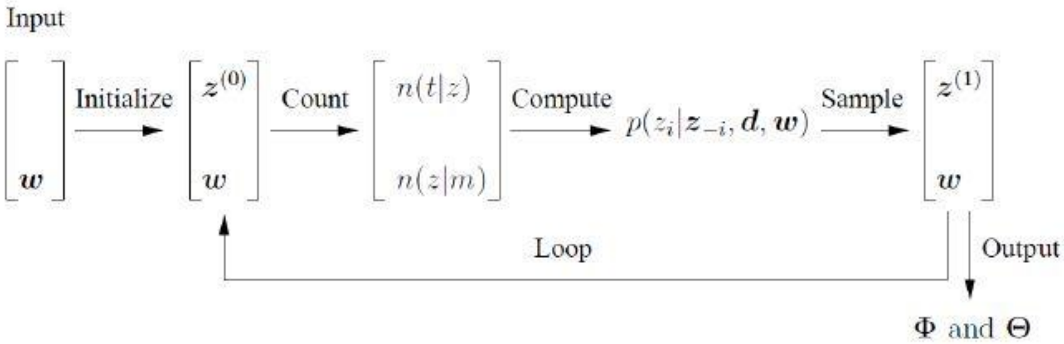


Figure 2.1: The procedure of learning LDA by Gibbs sampling.

初始时随机给文本中的每个单词分配主题$z^{(0)}$,然后统计每个主题 z 下出现 term t 的数量以及每个文档 m 下出现主题 z 中的词的数量，每一轮计算$p(z_i|\mathbf{z}_{-i}, \mathbf{d}, \mathbf{w})$，即排除当前词的主题分配，根据其他所有词的主题分配估计当前词分配各个主题的概率。当得到当前词属于所有主题 z 的概率分布后，根据这个概率分布为该词 sample 一个新的主题$z^{(1)}$。然后用同样的方法不断更新下一个词的主题，直到发现每个文档下 Topic 分布$\vec{\theta}_m$和每个 Topic 下词的分布$\vec{\phi}_k$收敛，算法停止，输出待估计的参数$\vec{\theta}_m$和$\vec{\phi}_k$，最终每个单词的主题$z_{m,n}$也同时得出。实际应用中会设置最大迭代次数。

每一次计算 $p(z_i|\boldsymbol{z}_{-i},\boldsymbol{d},\boldsymbol{w})$ 的公式称为 Gibbs updating rule.下面我们来推导 LDA 的联合分布和 Gibbs updating rule。

5.2 LDA 的联合分布

由 LDA 的概率图模型，我们可以把 LDA 的联合分布写成

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

第一项和第二项因子分别可以写成

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

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

可以发现两个因子的展开形式很相似，第一项因子是给定主题 Sample 词的过程，可以拆分成从 Dirichlet 先验中 Sample Topic Z 下词的分布 $\vec{\phi}_z$ 和从参数为 $\vec{\phi}_z$ 的多元分布中 Sample 词这两个步骤，因此是 Dirichlet 分布和 Multinomial 分布的概率密度函数相乘，然后在 $\vec{\phi}_z$ 上积分。注意这里用到的多元分布没有考虑词的顺序性，因此没有前面的系数项。$n_z^{(t)}$ 表示 term t 被观察到分配 topic z 的次数，$n_m^{(k)}$ 表示 topic k 分配给文档 m 中的 word 的次数.此外这里面还用到了 2.2 部分中导出的一个公式

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

因此这些积分都可以转化成 Dirichlet delta 函数，并不需要算积分。第二个因子是给定文档，sample 当前词的主题的过程。由此 LDA 的联合分布就可以转化成全部由 Dirichlet delta 函数组成的表达式

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

这个式子在后面推导 Gibbs updating rule 时需要使用。

5.3 Gibbs updating rule

得到 LDA 的联合分布后，我们就可以推导 Gibbs updating rule，即排除当前词的主题分配，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

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

里面用到了伽马函数的性质

$$\Gamma(z+1) = z\,\Gamma(z).$$

同时需要注意到

$$[\textstyle\sum_{k=1}^{K} n_m^{(k)}+\alpha_k]-1$$

这一项与当前词的主题分配无关，因为无论分配那个主题，对所有 k 求和的结果都是一样的，区别只在于拿掉的是哪个主题下的一个词。因此可以当成常量，最后我们只需要得到一个成正比的计算式来作为 Gibbs updating rule 即可。

5.4 Gibbs sampling algorithm

当 Gibbs sampling 收敛后，我们需要根据最后文档集中所有单词的主题分配来计算$\vec{\theta}_m$和$\vec{\phi}_k$，作为我们估计出来的概率图模型中的隐含变量。每个文档上 Topic 的后验分布和每个 Topic 下的 term 后验分布如下

$$p(\vec{\vartheta}_m|\vec{z}_m,\vec{\alpha}) = \frac{1}{Z_{\vartheta_m}} \prod_{n=1}^{N_m} p(z_{m,n}|\vec{\vartheta}_m) \cdot p(\vec{\vartheta}_m|\vec{\alpha}) = \mathrm{Dir}(\vec{\vartheta}_m|\vec{n}_m + \vec{\alpha})$$
$$p(\vec{\varphi}_k|\vec{z},\vec{w},\vec{\beta}) = \frac{1}{Z_{\varphi_k}} \prod_{\{i:z_i=k\}} p(w_i|\vec{\varphi}_k) \cdot p(\vec{\varphi}_k|\vec{\beta}) = \mathrm{Dir}(\vec{\varphi}_k|\vec{n}_k + \vec{\beta})$$

可以看出这两个后验分布对应的先验分布一样，仍然为 Dirichlet 分布，这也是共轭分布的性质决定的。
使用 Dirichlet 分布的的期望计算公式

$$\langle \mathrm{Dir}(\vec{a}) \rangle = a_i / \sum_i a_i.$$

我们可以得到两个 Multinomial 分布的参数$\vec{\theta}_m$和$\vec{\phi}_k$的计算公式如下

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} n_k^{(t)} + \beta_t},$$
$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k}.$$

综上所述，用 Gibbs Sampling 学习 LDA 参数的算法伪代码如下

Algorithm LdaGibbs($\{\vec{w}\}, \alpha, \beta, K$)
Input: word vectors $\{\vec{w}\}$, hyperparameters α, β, topic number K
Global data: count statistics $\{n_m^{(k)}\}, \{n_k^{(t)}\}$ and their sums $\{n_m\}, \{n_k\}$, memory for full conditional array $p(z_i|\cdot)$
Output: topic associations $\{\vec{z}\}$, multinomial parameters $\underline{\Phi}$ and $\underline{\Theta}$, hyperparameter estimates α, β

// initialisation
zero all count variables, $n_m^{(k)}, n_m, n_k^{(t)}, n_k$
for all documents $m \in [1, M]$ **do**
  for all words $n \in [1, N_m]$ in document m **do**
    sample topic index $z_{m,n}=k \sim \text{Mult}(1/K)$
    increment document–topic count: $n_m^{(k)}$ += 1
    increment document–topic sum: n_m += 1
    increment topic–term count: $n_k^{(t)}$ += 1
    increment topic–term sum: n_k += 1
// Gibbs sampling over burn-in period and sampling period
while not finished **do**
  for all documents $m \in [1, M]$ **do**
    for all words $n \in [1, N_m]$ in document m **do**
      // for the current assignment of k to a term t for word $w_{m,n}$:
      decrement counts and sums: $n_m^{(k)}$ −= 1; n_m −= 1; $n_k^{(t)}$ −= 1; n_k −= 1
      // multinomial sampling acc. to Eq. 78 (decrements from previous step):
      sample topic index $\tilde{k} \sim p(z_i|\vec{z}_{\neg i}, \vec{w})$
      // for the new assignment of $z_{m,n}$ to the term t for word $w_{m,n}$:
      increment counts and sums: $n_m^{(\tilde{k})}$ += 1; n_m += 1; $n_{\tilde{k}}^{(t)}$ += 1; $n_{\tilde{k}}$ += 1
  // check convergence and read out parameters
  if converged and L sampling iterations since last read out **then**
    // the different parameters read outs are averaged.
    read out parameter set $\underline{\Phi}$ according to Eq. 81
    read out parameter set $\underline{\Theta}$ according to Eq. 82

Fig. 9. Gibbs sampling algorithm for latent Dirichlet allocation

关于这个算法的代码实现可以参见

* Gregor Heinrich's LDA-J
* Yee Whye Teh's Gibbs LDA Matlab codes
* Mark Steyvers and Tom Griffiths's topic modeling matlab toolbox
* GibbsLDA++

6 参考文献及推荐 Notes

本文部分公式及图片来自 Parameter estimation for text analysis，感谢 Gregor Heinrich 详实细致的 Technical report。 看过的一些关于 LDA 和 Gibbs Sampling 的 Notes， 这个是最准确细致的，内容最为全面系统。下面几个 Notes 对 Topic Model 感兴趣的朋友也推荐看一看。

[1] Christopher M. Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[2] Gregor Heinrich. Parameter estimation for text analysis. Technical report, 2004.

[3] Wang Yi. Distributed Gibbs Sampling of Latent Topic Models: The Gritty Details Technical report, 2005.

[4] Wayne Xin Zhao, Note for pLSA and LDA, Technical report, 2011.

[5] Freddy Chong Tat Chua. Dimensionality reduction and clustering of text documents.Technical report, 2009.

[6] Wikipedia, Dirichlet distribution , http://en.wikipedia.org/wiki/Dirichlet_distribution