主题模型LDA

七月算法 邹博

2015年5月9日

主要内容和目标

- 共轭先验分布
- Dirichlet分布
- unigram model
- LDA
- Gibbs采样算法

共轭先验分布

- 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。
- In Bayesian probability theory, if the posterior distributions p(θ|x) are in the same family as the prior probability distribution p(θ), the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function.

共轭先验分布的提出

- 某观测数据服从概率分布P(θ)时，
- 当观测到新的X数据时，有如下问题：
  - 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
    - $\theta \leftarrow \theta + \Delta\theta$
  - 当重新估计θ的时候，给出新参数值θ的新概率分布。即：P(θ|x)

分析

- 根据贝叶斯法则

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

- 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。
- 方案：选取P(x| θ)的共轭先验作为P(θ)的分布，这样，P(x| θ)乘以P(θ)然后归一化结果后其形式和P(θ)的形式一样。

举例说明

- 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）。简单解释就是，这两个参数决定了θ参数。Beta分布形式为

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

先验概率和后验概率的关系

- 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，即：伯努利分布的共轭先验是Beta分布。

伪计数

- 可以发现，在后验概率的最终表达式中，参数α和β和x，1-x一起作为参数θ的指数。而这个指数的实践意义是：投币过程中，正面朝上的次数。因此，α和β常常被称作“伪计数”。

推广

- 二项分布→多项分布
- Beta分布→Dirichlet分布

Dirichlet分布

- Γ 函数是阶乘在实数上的推广

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

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

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

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

Dirichlet分布的定义

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

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

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

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

Dirichlet分布的分析

- α 是参数，共K个
- 定义在$x_1,x_2\ldots x_{K-1}$维上
  - $x_1+x_2+\ldots+x_{K-1}+x_K=1$
  - $x_1,x_2\ldots x_{K-1}>0$
  - 定义在(K-1)维的单纯形上，其他区域的概率密度为0
- α 的取值对$Dir(p|\alpha)$有什么影响？

Symmetric Dirichlet distribution

- A very common special case is the symmetric Dirichlet distribution, where all of the elements making up the parameter vector have the same value. Symmetric Dirichlet distributions are often used when a Dirichlet prior is called for, since there typically is *no prior knowledge* favoring one component over another. Since all elements of the parameter vector have the same value, the distribution alternatively can be parametrized by a single scalar value α, called the concentration parameter(聚集参数).

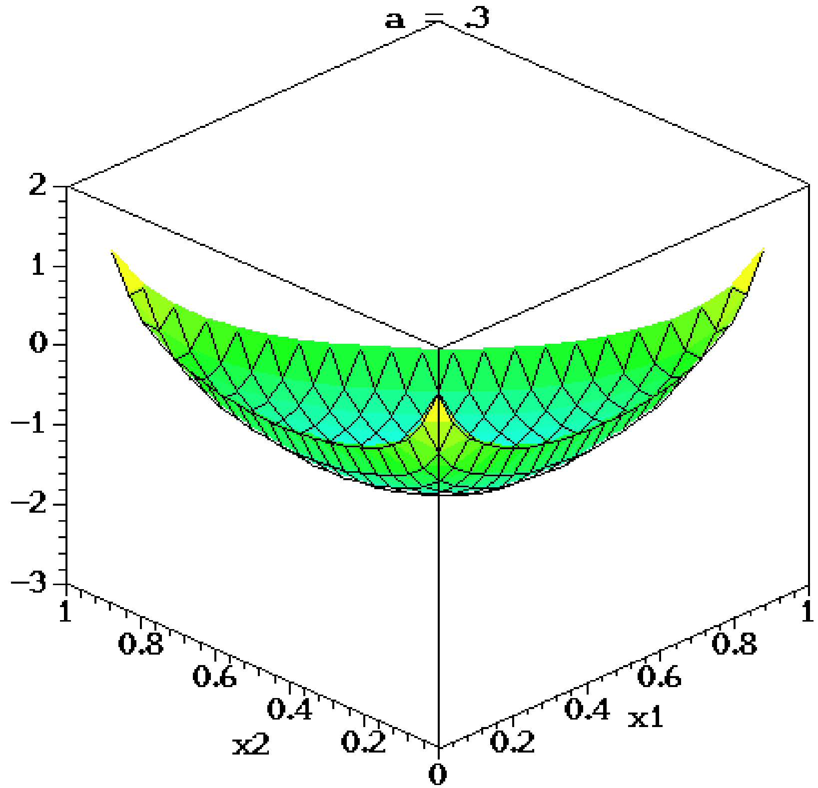
对称Dirichlet分布

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

$$
\Delta_K(\alpha) = \frac{\Gamma(\alpha)^K}{\Gamma(K\alpha)}
$$

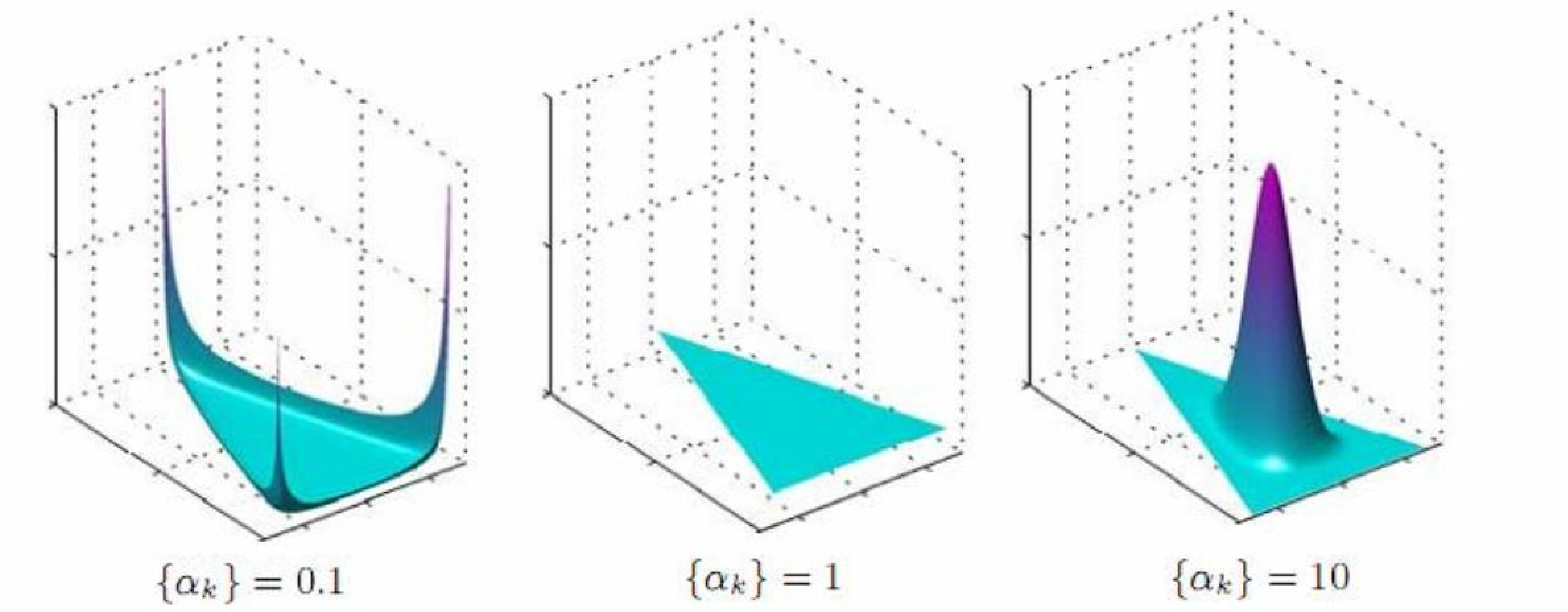
对称Dirichlet分布的参数分析

- α =1 时
  - 退化为均匀分布
- 当 α >1 时
  - p1=p2=…=pk的概率增大
- 当 α <1 时
  - p1=1，pi=0的概率增大



图像说明：将Dirichlet分布的概率密度函数取对数,绘制对称Dirichlet分布的图像，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α1=α2=α3=α，图中反映的是α从0.3变化到2.0的概率对数值的变化情况。

参数 α 对Dirichlet分布的影响



参数选择对对称Dirichlet分布的影响

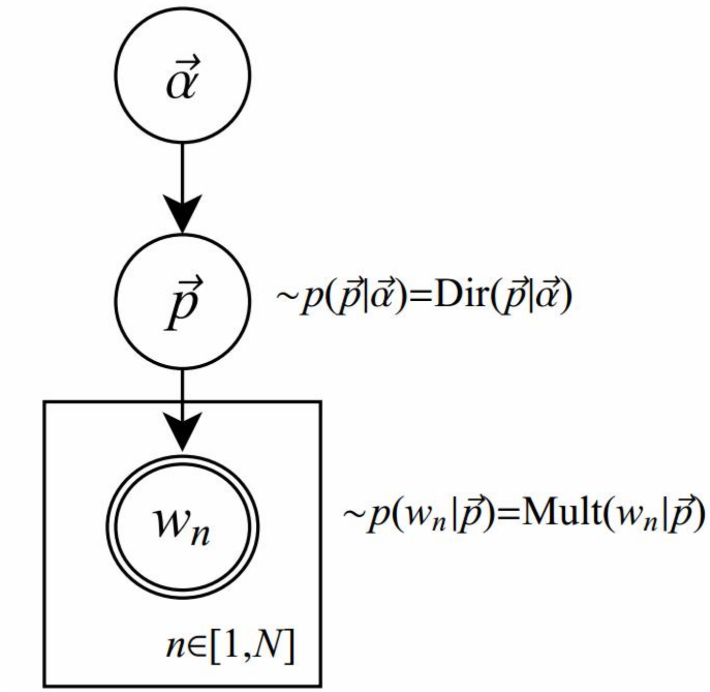
- When α =1, the symmetric Dirichlet distribution is equivalent to a uniform distribution over the open standard (K−1)-simplex, i.e. it is uniform over all points in its support. Values of the concentration parameter above 1 prefer variants that are dense, evenly distributed distributions, i.e. all the values within a single sample are similar to each other. Values of the concentration parameter below 1 prefer sparse distributions, i.e. most of the values within a single sample will be close to 0, and the vast majority of the mass will be concentrated in a few of the values.

多项分布的共轭分布是Dirichlet分布

$$
\begin{aligned}
\boldsymbol{\alpha} &= (\alpha_1, \cdots, \alpha_K) &&= \text{concentration hyperparameter} \\
\mathbf{p} \mid \boldsymbol{\alpha} &= (p_1, \cdots, p_K) &&\sim \operatorname{Dir}(K, \boldsymbol{\alpha}) \\
\mathbb{X} \mid \mathbf{p} &= (\mathbf{x}_1, \cdots, \mathbf{x}_K) &&\sim \operatorname{Cat}(K, \mathbf{p}) \\
\mathbf{c} &= (c_1, \cdots, c_K) &&= \text{number of occurrences of category } i \\
\mathbf{p} \mid \mathbb{X}, \boldsymbol{\alpha} &\sim \operatorname{Dir}(K, \mathbf{c} + \boldsymbol{\alpha}) &&= \operatorname{Dir}(K, c_1 + \alpha_1, \cdots, c_K + \alpha_K)
\end{aligned}
$$

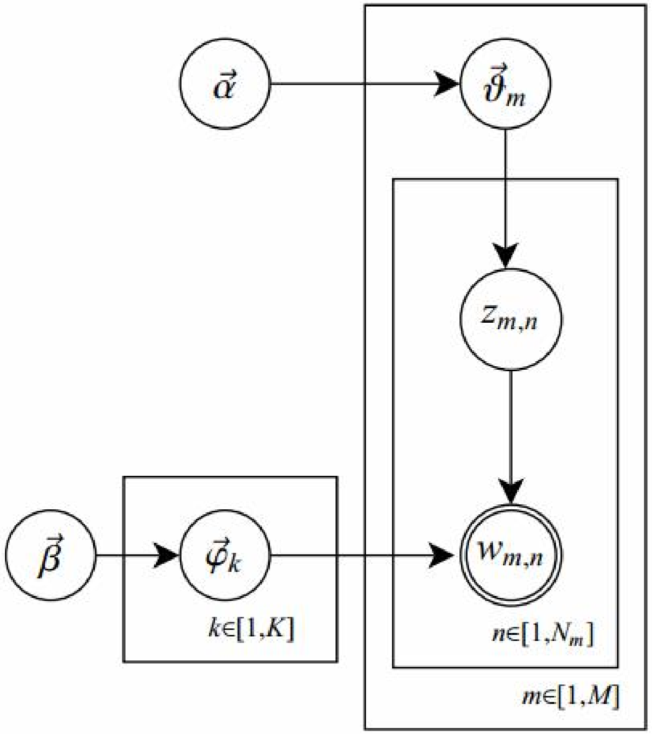
unigram model

- 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wn表示在文本中观察到的第n个词，n∈[1,N]表示文本中一共有N个词。加上方框表示重复，即一共有N个这样的随机变量wn。**p**和α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。



LDA的解释

- 共有m篇文章，一共涉及了K个主题；
- 每篇文章(长度为N_m)都有各自的主题分布，主题分布是多项分布，该多项分布的参数服从Dirichlet分布，该Dirichlet分布的参数为α；
- 每个主题都有各自的词分布，词分布为多项分布，该多项分布的参数服从Dirichlet分布，该Dirichlet分布的参数为β；
- 对于某篇文章中的第n个词，首先从该文章的主题分布中采样一个主题，然后在这个主题对应的词分布中采样一个词。不断重复这个随机生成过程，直到m篇文章全部完成上述过程。



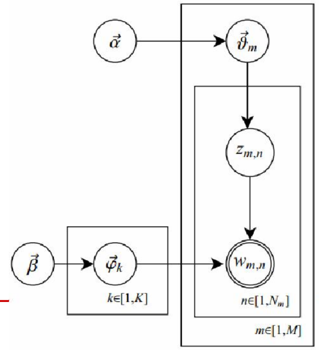
详细解释

- 字典中共有V个term，不可重复，这些term出现在具体的文章中，就是word
- 语料库中共有m篇文档d1,d2…dm
- 对于文档di，由Ni个word组成，可重复；
- 语料库中共有K个主题T1，T2…Tk；
- α，β为先验分布的参数，一般事先给定：如取0.1的对称Dirichlet分布
- θ是每篇文档的主题分布
  - 对于第i篇文档di，它的主题分布是$\theta_i=(\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK})$，是长度为K的向量
- 对于第i篇文档di，在主题分布θi下，可以确定一个具体的主题zij=j，j∈[1,K]，
- φk表示第k个主题的词分布
  - 对于第k个主题Tk，词分布$\phi_k=(\phi_{k1}, \phi_{k2} \ldots \phi_{kv})$，是长度为v的向量
- 由zij选择ϕ_{zij}，表示由词分布ϕ_{zij}确定word，从而得到w_{ix}

详细解释

- 图中K为主题个数，M为文档总数，Nm是第m个文档的单词总数。β是每个Topic下词的多项分布的Dirichlet先验参数，α是每个文档下Topic的多项分布的Dirichlet先验参数。zmn是第m个文档中第n个词的主题，wmn是m个文档中的第n个词。两个隐含变量θ和φ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）

参数的学习



- 给定一个文档集合，w_{mn}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_{mn}、θ、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：

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

似然概率

- 一个词wmn初始化为一个term t的概率是

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

- 每个文档中出现topick的概率乘以topick下出现termt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为：

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

Gibbs Sampling

- Gibbs Sampling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采样当前维度的值。不断迭代，直到收敛输出待估计的参数。
- 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采样一个新的主题。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θ和每个Topic下词的分布φ收敛，算法停止，输出待估计的参数θ和φ，最终每个单词的主题zmn也同时得出。
- 实际应用中会设置最大迭代次数。每一次计算$p(z_i|\mathbf{z}_{-i},\mathbf{d},\mathbf{w})$的公式称为Gibbs updating rule。

联合分布

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

- 第一项因子是给定主题采样词的过程
- 后面的因子计算，$n_z^{(t)}$表示term t被观察到分配topicz的次数， $n_m^{(t)}$ 表示topic k分配给文档m中的word的次数。

计算因子

$$p(\vec{w}|\vec{z},\vec{\beta}) = \int p(\vec{w}|\vec{z},\underline{\Phi})\ p(\underline{\Phi}|\vec{\beta})\ \mathrm{d}\underline{\Phi}$$

$$= \int \prod_{z=1}^{K} \frac{1}{\Delta(\vec{\beta})} \prod_{t=1}^{V} \varphi_{z,t}^{n_z^{(t)}+\beta_t-1}\ \mathrm{d}\vec{\varphi}_z$$

$$= \prod_{z=1}^{K} \frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{\beta})}, \qquad \vec{n}_z = \{n_z^{(t)}\}_{t=1}^{V}$$

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

计算因子

$$p(\vec{z}|\vec{\alpha}) = \int p(\vec{z}|\underline{\Theta})\ p(\underline{\Theta}|\vec{\alpha})\ \mathrm{d}\underline{\Theta}$$

$$= \int \prod_{m=1}^{M} \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^{K} \vartheta_{m,k}^{n_m^{(k)}+\alpha_k-1}\ \mathrm{d}\vec{\vartheta}_m$$

$$= \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})}, \quad \vec{n}_m = \{n_m^{(k)}\}_{k=1}^{K}$$

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

Gibbs updating rule

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

词分布和主题分布

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

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

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

参考文献

- Gregor Heinrich, Parameter estimation for text analysis
- David M. Blei, Andrew Y. Ng, Michael I. Jordan, Latent Dirichlet Allocation, 2003
- http://en.wikipedia.org/wiki/Dirichlet_distribution(Dirichlet分布)
- http://en.wikipedia.org/wiki/Conjugate_prior(共轭分布)

感谢大家！

恳请大家批评指正！