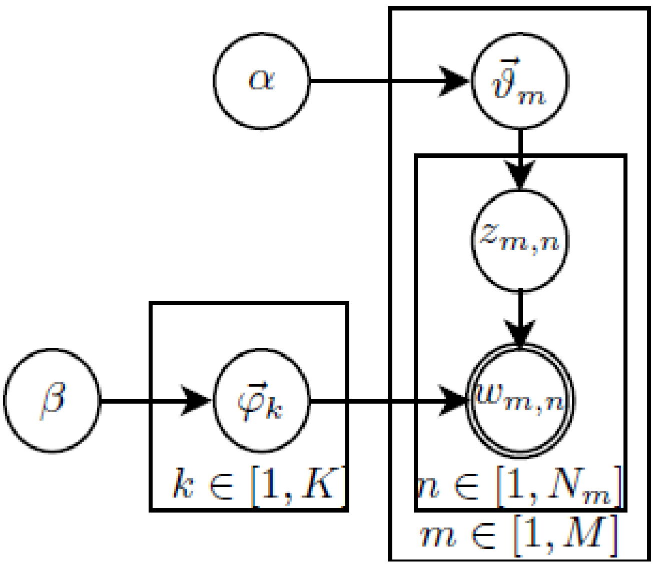
Crescent

心怀畏惧

LDA学习笔记---来自《Parameter estimation for text analysis》

LDA的概率图如下图1所示：



(b)

参数的意思如图2所示：

M number of documents to generate (const scalar).
K number of topics / mixture components (const scalar).
V number of terms t in vocabulary (const scalar).
$\vec{\alpha}$ hyperparameter on the mixing proportions (K-vector or scalar if symmetric).
$\vec{\beta}$ hyperparameter on the mixture components (V-vector or scalar if symmetric).
$\vec{\vartheta}_m$ parameter notation for $p(z|d{=}m)$, the topic mixture proportion for document m. One proportion for each document, $\underline{\Theta} = \{\vec{\vartheta}_m\}_{m=1}^M$ ($M \times K$ matrix).
$\vec{\varphi}_k$ parameter notation for $p(t|z{=}k)$, the mixture component of topic k. One component for each topic, $\underline{\Phi} = \{\vec{\varphi}_k\}_{k=1}^K$ ($K \times V$ matrix).
N_m document length (document-specific), here modelled with a Poisson distribution [BNJ02] with constant parameter ξ.
$z_{m,n}$ mixture indicator that chooses the topic for the nth word in document m.
$w_{m,n}$ term indicator for the nth word in document m.

Fig. 7. Quantities in the model of latent Dirichlet allocation

根据模型，文章m的第n个词t是这样生成的：先从文章m的doc-topic分布中生成一个topic编号$z_{m,n}$，在根据编号第$z_{m,n}$个的topic-word分布中生成这个词，总够有K个topic，所以总的概率为：

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

如果我们写出这篇文章的complete-data的联合分布，那么式子就是这样的：

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

通过对$\vec{\vartheta}_m$（doc-topic分布）和$\underline{\Phi}$（topic-word分布）积分以及$z_{m,n}$求和，我们可以求得$\vec{w}_m$的边缘分布：

$$p(\vec{w}_m|\vec{\alpha},\vec{\beta}) = \iint p(\vec{\vartheta}_m|\vec{\alpha}) \cdot p(\underline{\Phi}|\vec{\beta}) \cdot \prod_{n=1}^{N_m} \sum_{z_{m,n}} p(w_{m,n}|\vec{\varphi}_{z_{m,n}})p(z_{m,n}|\vec{\vartheta}_m)\,\mathrm{d}\underline{\Phi}\,\mathrm{d}\vec{\vartheta}_m \tag{58}$$

$$= \iint p(\vec{\vartheta}_m|\vec{\alpha}) \cdot p(\underline{\Phi}|\vec{\beta}) \cdot \prod_{n=1}^{N_m} p(w_{m,n}|\vec{\vartheta}_m,\underline{\Phi})\,\mathrm{d}\underline{\Phi}\,\mathrm{d}\vec{\vartheta}_m \tag{59}$$

因为一个语料库有很多篇文章，而且文章之间都是相互独立的，所以整个语料库的似然为

$$p(\mathcal{W}|\vec{\alpha},\vec{\beta}) = \prod_{m=1}^{M} p(\overrightarrow{w_m}|\vec{\alpha},\vec{\beta})$$

虽然LDA（latent Dirichlet allocation)是个相对简单的模型，对它直接推断一般也是不可行的，所以我们要采用近似推断的方法，比如Gibbs sampling。

Gibbs sampling

Gibbs sampling是MCMC(Markov-chain Monte Carlo)算法的一种特殊情况，经常用于处理高维模型的近似推断。MCMC方法可以通过马尔科夫链的平稳分布模拟高维的概率分布$p(\vec{x})$。当马尔科夫链经过了burn-in阶段，消除了初始参数的影响，进入平稳状态之后，它的每次转移都能生成一个$p(\vec{x})$的样本。Gibbs samppling 是MCMC的特殊情况，它每次固定一个维度的x_i,然后通过其他维度的数据（$\vec{x}_{\neg i}$)生成这个维度的样本。算法如下：

1. choose dimension i(random by permutation)。
2. sample x_i from $p(x_i|\vec{x}_{\neg i})$。

为了构造Gibbs抽样，我们必须知道条件概率$p(x_i|\vec{x}_{\neg i})$，这个概率可以通过以下公式获得:

$$p(x_i|\vec{x}_{\neg i}) = \frac{p(x_i,\vec{x}_{\neg i})}{p(\vec{x}_{\neg i})} = \frac{p(x_i,\vec{x}_{\neg i})}{\int p(\vec{x})dx_i}$$

对于那些含有隐藏变量$\vec{z}$的模型来说，通常需要求得他们的后验概率$p(\vec{z}|\vec{x})$，对于这样的模型，Gibbs sampler的式子如下：

$$p(z_i|\vec{z}_{\neg i},\vec{x}) = \frac{p(\vec{z},\vec{x})}{p(\vec{z}_{\neg i},\vec{x})} = \frac{p(\vec{z},\vec{x})}{\int_z p(\vec{z},\vec{x})dx_i}$$

当样本$\tilde{\overrightarrow{z_r}}, r \in [1,R]$的数量足够多时，隐藏变量的后验概率可以用以下式子来估计：

$$p(\vec{z}|\vec{x}) = \frac{1}{R}\sum_{r=1}^{R}\delta(\vec{z}-\tilde{\overrightarrow{z_r}})$$

其中Kronecker delta $\delta(\vec{u}) = \{1 \text{ if } \vec{u} = 0; 0 \text{ otherwise} \}$。

为了构造LDA的采样器，我们首先确定模型中的隐含变量为$z_{m,n}$ 。而参数$\underline{\Theta}$和$\underline{\Phi}$都可以用观察到的$w_{m,n}$ 和对应的$z_{m,n}$ 求积分得到。贝叶斯推断的目标是分布$p(\vec{z}|\vec{w})$，它与联合分布成正比:

$$p(\vec{z}|\vec{w}) = \frac{p(\vec{z},\vec{w})}{p(\vec{w})} = \frac{\prod_{i=1}^{W} p(z_i,w_i)}{\prod_{i=1}^{W}\sum_{k=1}^{K} p(z_i=k,w_i)}$$

这里忽略了超参数（hyperparameter）$\vec{\alpha}$和$\vec{\beta}$。可以看到分母部分十分难求，它包括了K^W个项的求和。所以我们使用Gibbs Sample方法，通过全部的条件分布$p(z_i|\vec{z}_{\neg i},\vec{w})$来模拟得到$p(\vec{z}|\vec{w})$。

LDA的联合分布

LDA的联合分布可以写成如下的式子:

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

因为式子中的第一部分与α独立，第二部分与β独立，所以两个式子可以分别处理。先看第一个分布$p(\vec{w}|\vec{z})$，可以从观察到的词以及其主题的多项分布中生成：

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

意思是，语料中的W个词是根据主题z_i观察到的独立多项分布。(我们把每个词看做独立的多项分布产生的结果，忽略顺序因素，所以没有多项分布的系数）。φ_{z_i,w_i}是一个$K*V$的矩阵，把词划分成主题和词汇表，公式如下：

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

$n_k^{(t)}$代表了主题k下词t出现的次数。目标分布$p(\vec{w}|\vec{z},\vec{\beta})$可以通过对$\underline{\Phi}$求狄利克雷积分得到:

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

类似地，主体分布$p(\vec{z}|\vec{a})$也可以通过这种方法产生，$\underline{\Theta}$为$D*K$的矩阵，公式如下：

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

$n_m^{(k)}$代表了文章m下主题k出现的次数。对$\underline{\Theta}$求积分，我们得到:

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

然后联合分布就变成了

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

完全条件分布(full conditional)

我们令$i=(m,n)$代表第m篇文章中的第n个词，$\neg i$代表除去这个词之后剩下的其他词，令$\vec{w}=\{w_i=t,\vec{w}_{\neg i}\}$，$\vec{z}=\{z_i=k,\vec{z}_{\neg i}\}$，我们求得

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

这个式子需要注意的：

1. 因为忽略了$p(w_i)$这个常数，所以后来的式子是$\propto$成正比。
2. 对于第m篇文章中的第n个词，其主题为k。$n_k^{(t)} = n_{k,\neg i}^{(t)} + 1, n_m^{(k)} = n_{m,\neg i}^{(k)} + 1$，对于其他文档和其他主题都没有影响。

这个公式很漂亮，右边是$p(topic|doc) \cdot p(word|topic)$，这个概率其实就是$doc \rightarrow topic \rightarrow word$的路径概率，所以Gibbs Sampling 公式的物理意义就是在K条路径中采样。（图）

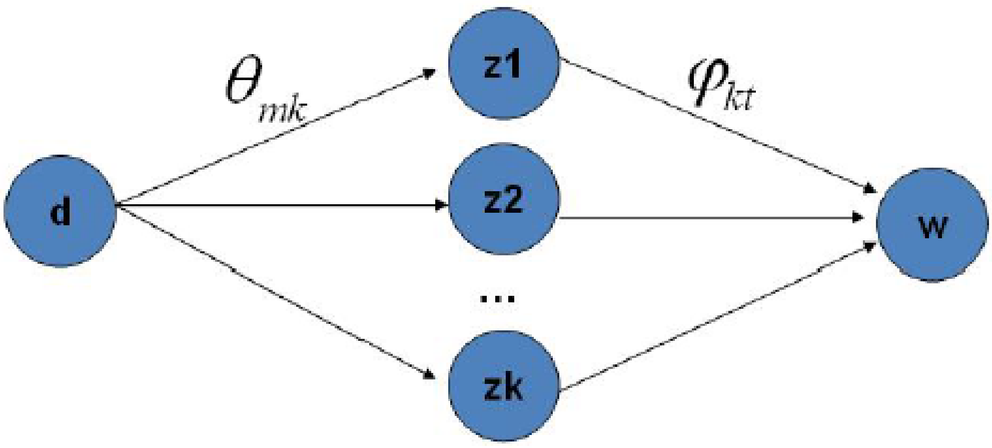


Figure 33: doc-topic-word 路径概率

多项分布参数

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

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

根据图3和图4的Dirichlet-Multinomial结构，我们知道$\overrightarrow{\theta_m}$和$\overrightarrow{\phi_k}$的后验概率为:(令$\mathcal{M} = \{\vec{w}, \vec{z}\}$)（备注1）：

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

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

最后，根据狄利克雷分布的期望$< Dir(\vec{a}) >= a_i / \sum_i a_i$（备注2），我们得到

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

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

最后，整个LDA算法的流程图为

⊡ 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: $n_k + 1$
  end for
end for
⊡ 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. 79 (decrements from previous step):
      sample topic index $\tilde{k} \sim p(z_i|\vec{z}_{\neg i}, \vec{w})$
      ⊡ use the new assignment of $z_{m,n}$ to the term t for word $w_{m,n}$ to:
      increment counts and sums: $n_m^{(\tilde{k})} + 1; n_m + 1; n_{\tilde{k}}^{(t)} + 1; n_{\tilde{k}} + 1$
    end for
  end for
  ⊡ 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. 82
    read out parameter set $\underline{\Theta}$ according to Eq. 83
  end if
end while

备注：

1.狄利克雷分布的后验概率公式：

$$\begin{aligned} p(\mathcal{W}|\vec{\alpha}) &= \int_{\vec{p}\in\mathcal{P}} \prod_{n=1}^{N} \mathrm{Mult}(W{=}w_n|\vec{p},1)\ \mathrm{Dir}(\vec{p}|\vec{\alpha})\,\mathrm{d}\vec{p} \\ &= \int_{\vec{p}\in\mathcal{P}} \prod_{v=1}^{V} p_v^{n^{(v)}} \frac{1}{\Delta(\vec{\alpha})} \prod_{v=1}^{V} p_v^{\alpha_v-1}\,\mathrm{d}^V\vec{p} \\ &= \frac{1}{\Delta(\vec{\alpha})} \int_{\vec{p}\in\mathcal{P}} \prod_{v=1}^{V} p_v^{n^{(v)}+\alpha_v-1}\,\mathrm{d}^V\vec{p} \qquad \Big|\ \text{Dirichlet}\int \\ &= \frac{\Delta(\vec{n}+\vec{\alpha})}{\Delta(\vec{\alpha})}, \qquad \vec{n}=\{n^{(v)}\}_{v=1}^{V} \end{aligned}$$

2.由于狄利克雷分布为：

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

对于$\vec{p}$中一项p_i的期望为：

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

参考文献：

1.主要来自《Parameter estimation for text analysis》
2.《LDA数学八卦》

本条目发布于 2013 年 3 月 12 日 [http://www.crescentmoon.info/?p=296] 。属于 学术 分类，被贴了 LDA、机器学习 标签。