东北财经大学
DONGBEI UNIVERSITY OF FINANCE & ECONOMICS

Parameter estimation for text analysis
Gregor Heinrich

Xdy

DONGBEI University
Of finance and economics

April 15, 2018

Outline

1 Supplementary knowledge

- statistical distribution
  - Binomal and Multinomal
  - Gama
  - Beta
  - Dirichlet
- Bayes Estimation
  - 贝叶斯公式
  - 几个概念
- Statistical Sampling
  - Classical Sampling
  - MCMC Sampling
  - Gibbs Sampling

2 Text Model

- Unigram Model
- Topic Model and PLSA
- LDA(Latent Dirichlet Allocation)

Outline

1. Supplementary knowledge
   - statistical distribution
     - Binomal and Multinomal
     - Gama
     - Beta
     - Dirichlet
   - Bayes Estimation
     - 贝叶斯公式
     - 几个概念
   - Statistical Sampling
     - Classical Sampling
     - MCMC Sampling
     - Gibbs Sampling
2. Text Model
   - Unigram Model
   - Topic Model and PLSA
   - LDA(Latent Dirichlet Allocation)

Binomal and Multinomal

$$b(n,p): \quad p(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$Mult(\overrightarrow{n}\,|\,\overrightarrow{p}, N) = \binom{\overrightarrow{n}}{N} \prod_{k}^{K} p_k^{n_k}$$

Gama

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \qquad \Gamma(n+1) = n\Gamma(n) = n!$$

$$Gamma(\alpha, \lambda): \qquad P(x|\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(x)} x^{\alpha-1} e^{-\lambda x}, \qquad x \geqslant 0$$

Beta

$B(a,b) = \int_0^1 x^{\alpha-1}(1-x)^{b-1}dx, \qquad B(a,b) = B(b,a) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

$$\begin{aligned} Beta(a,b) P(x|a,b) &= \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \quad 0 < x < 1 \end{aligned}$$

Beta

- $Beta(1,1) \Longrightarrow Uniform(0,1)$
- 次序统计量的概率密度函数: $x_{(k)}$

$$
\begin{aligned}
P_k(x) &= \frac{n!}{(k-1)!(n-k)!} * F(x)^{k-1} * [1-F(x)]^{n-k} * p(x) \\
&= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} * F(x)^{k-1} * [1-F(x)]^{n-k} * p(x)
\end{aligned}
$$

$$
if\ \ x \sim Uniform(0,1), \quad then\ p(x) = 1 \quad let\ \alpha = k,\ \beta = n-k+1,\ then
$$

$$
P_k(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} * F(x)^{k-1} * [1-F(x)]^{n-k} \sim Beta(\alpha,\beta)
$$

Dirichlet

$$Dir(\overrightarrow{\alpha}): \quad P(\overrightarrow{x}|\overrightarrow{\alpha}) = \frac{1}{B(\overrightarrow{\alpha})}\prod_{i=1}^{K} x_i^{\alpha_i - 1} = \frac{\Gamma(\sum_{i=1}^{K})}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

$$\int Dir(\overrightarrow{\alpha})dx = \frac{1}{B(\overrightarrow{\alpha})}\int\prod_{i=1}^{K}\Gamma(\alpha_i) = \frac{1}{B(\overrightarrow{\alpha})}B(\overrightarrow{\alpha}) = 1$$

Dirichlet

$$Dir(\overrightarrow{\alpha}): \quad P(\overrightarrow{x}|\overrightarrow{\alpha}) = \frac{1}{B(\overrightarrow{\alpha})}\prod_{i=1}^{K} x_i^{\alpha_i - 1} = \frac{\Gamma(\sum_{i=1}^{K})}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

- when k=3, three dimensional, we get:

$$P(\overrightarrow{x}|\overrightarrow{\alpha}) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} x_3^{\alpha_3 - 1}$$

$$\Downarrow$$

$$P(x_1, x_2, x_3|\alpha_1, \alpha_2, \alpha_3)$$

Dirichlet

- 次序统计量的联合概率密度函数:$x_{(k)}, x_{(j)}$

$$P_{kj} = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} * F(x_k)^{k-1}[F(x_j)-F(x_k)]^{j-k-1}[1-F(x_j)]^{n-j} * p(x)$$

if $x \sim Uniform(0,1)$,then $p(x_k) = p(x_j) = 1$ let
$\alpha_1 = k, \alpha_2 = k - j, \alpha_3 = n - j + 1$ then
$P_{kj} = \frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} * F(x_k)^{\alpha_1} * [F(x_j) - F(x_k)]^{\alpha_2 - 1} * [1 - F(x_j)]^{\alpha_3 - 1}$

Dirichlet

- if $\overrightarrow{P} \sim Dir$ then:

$$E(\overrightarrow{P}) = E(p_1, p_2, \dots) = (\frac{\alpha_1}{\sum_{i=1}^{k} \alpha_i}, \frac{\alpha_2}{\sum_{i=1}^{k} \alpha_i}, \dots, \frac{\alpha_k}{\sum_{i=1}^{k} \alpha_i})$$

Same as:

$$E(P) = \frac{a}{a+b}, \qquad P \sim Beta(a, b)$$

Bayes Estimation

- 贝叶斯学派和古典学派对于参数的观点
- 贝叶斯学派的参数估计:

$$\left.\begin{array}{l}\text{同等无知}\\ \text{共轭先验}\\ \vdots\end{array}\right\}\text{先验分布}+\text{样本信息}=\text{后验分布}\Rightarrow\left\{\begin{array}{l}\text{后验分布最大}\\ \text{均值}\\ \vdots\end{array}\right.$$

贝叶斯公式

$$
\begin{aligned}
\pi(\theta|m_1) &= \frac{\pi(\theta, m_1)}{\pi(m_1)} \\
&= \frac{\pi(\theta)\pi(m_1|\theta)}{\int \pi(\theta, m_1)d\theta} \\
&= \frac{\pi(\theta)\pi(m_1|\theta)}{\int \pi(\theta)\pi(m_1|\theta)d\theta}
\end{aligned}
$$

贝叶斯公式

Example 1.不均匀硬币，正面概率θ，抛m次，m_1正，m_2反，求θ

- Classical: $\hat{\theta} = \frac{m_1}{m}$
- Bayes:Suppose $p(\theta) \sim Uniform(0,1), \quad p(m_1|\theta) \sim b(m,\theta)$ we have:

$$p(\theta) = 1, \quad p(m_1|\theta) = \binom{m_1}{m} \theta^{m_1}(1-\theta)^{m_2}$$

$$p(m_1,\theta) = p(\theta)p(m_1|\theta) = p(m_1|\theta)$$

$$p(m_1) = \int_0^1 p(m_1,\theta) = \binom{m_1}{m} \int \theta_{m_1}(1-\theta)^{m_2} = \binom{m_1}{m} B(m_1+1, m_2+1)$$

$$\therefore p(\theta|m_1) = \frac{1}{B(m_1+1, m_2+1)} \theta^{m_1}(1-\theta)^{m_2} \sim Beta(m_1+1, m_2+1)$$

Conjugate prior distribution

- Beta-Binomal Conjugate:

$$Beta(\theta|\alpha,\beta)+BinomCount(m_1,m_2)=Beta(\theta|\alpha+m_1,\beta+m_2)$$

see the Example 1
prior: $\theta \sim U(0,1)$
Sample: $p(m_1|\theta) \sim b(m,\theta)$
Posterior:$\pi(\theta|m_1) \sim Beta(m_1+1,m_2+1)$

Conjugate prior distribution

- Dirichlet-Multinominal Conjugate

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

Example 2: 3 dimensional

$$Dir(k_1, k_2, k_3)+MultCount(m_1, m_2, m_3) = Dir(k_1+m_1, k_2+m_2, k_3+m_3)$$

分布的核与满条件概率分布

- 分布的核

$$b(n,p) \propto p^x(1-p)^{n-x} \qquad Beta(a,b) \propto x^{a-1}(1-x)^{b-1}$$

$$Gamma(\alpha,\lambda) \propto x^{\alpha-1}e^{-\lambda x} \qquad Did(\overrightarrow{\alpha}) \propto \prod^{k} x_i^{\alpha_i-1}$$

- Full Conditional distribution

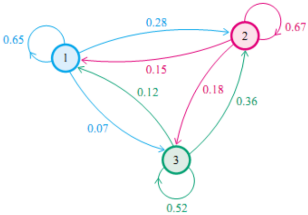
Classical Sampling

1. $U(0,1)$sampling
2. $N(\mu,\theta)$sampling
3. The low-dimensional and simple distribution sampling can use the similar method.But for the complicated or high-dimensional ,we can't get the sample by using classical method.

Markov Chain

- $P(X_{t+1} = x | X_t, X_{t-1}, \ldots) = P(X_{t+1} = x | X_t)$
- Example 3

		子代		
	State	1	2	3
父代	1	0.65	0.28	0.07
	2	0.15	0.67	0.18
	3	0.12	0.36	0.52



Markov Chain

使用矩阵的表示方式，转移概率矩阵记为

$$P = \begin{bmatrix} 0.65 & 0.28 & 0.07 \\ 0.15 & 0.67 & 0.18 \\ 0.12 & 0.36 & 0.52 \end{bmatrix}$$

假设初始概率分布为π_0=[0.21,0.68,0.11]，则我们可以计算前n代人的分布状况如下

第n代人	下层	中层	上层
0	0.210	0.680	0.110
1	0.252	0.554	0.194
2	0.270	0.512	0.218
3	0.278	0.497	0.225
4	0.282	0.490	0.226
5	0.285	0.489	0.225
6	0.286	0.489	0.225
7	0.286	0.489	0.225
8	0.289	0.488	0.225
9	0.286	0.489	0.225

Markov Chain

我们换一个初始概率分布π_0=[0.75,0.15,0.1] 试试看，继续计算前n 代人的分布状况如下

第n代人	下层	中层	上层
0	0.75	0.15	0.1
1	0.522	0.347	0.132
2	0.407	0.426	0.167
3	0.349	0.459	0.192
4	0.318	0.475	0.207
5	0.303	0.482	0.215
6	0.295	0.485	0.220
7	0.291	0.487	0.222
8	0.289	0.488	0.225
9	0.286	0.489	0.225
10	0.286	0.489	0.225
...	...	...	...

Markov Chain

$$P^{20} = P^{21} = \cdots = P^{100} = \cdots = \begin{bmatrix} 0.286 & 0.489 & 0.225 \\ 0.286 & 0.489 & 0.225 \\ 0.286 & 0.489 & 0.225 \end{bmatrix}$$

我们发现，当n 足够大的时候，这个P^n矩阵的每一行都是稳定地收敛到 π =[0.286,0.489,0.225] 这个概率分布。这个收敛现象并非是我们这个马氏链独有的，而是绝大多数马氏链的共同行为，关于马氏链的收敛我们有如下漂亮的定理：

Markov Chain

马氏链定理

如果一个非周期马氏链具有转移概率矩阵P, 且它的任何两个状态是连通的，那么$\lim_{n\to\infty} P_{ij}^n$存在且与$i$无关,记$\lim_{n\to\infty} P_{ij}^n = \pi(j)$，我们有:

1. $$\lim_{x\to\infty} P^n = \begin{bmatrix} \pi(1) & \pi(2) & \cdots & \pi(j) & \cdots \\ \pi(1) & \pi(2) & \cdots & \pi(j) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \pi(1) & \pi(2) & \cdots & \pi(j) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$
2. $$\pi(j) = \sum_{i=0}^{\infty} \pi(i) P_{ij}$$
3. π是方程$\pi P = \pi$的唯一非负解.

π称为马氏链的平稳分布。

MCMC

对于给定的概率分布$p(x)$, 我们希望能有便捷的方式生成它对应的样本。由于马氏链能收敛到平稳分布，于是一个很的漂亮想法是：如果我们能构造一个转移矩阵为P 的马氏链，使得该马氏链的平稳分布恰好是$p(x)$，那么我们从任何一个初始状态x_0出发沿着马氏链转移, 得到一个转移序列$x_0, x_1, x_2, x_n, x_{n+1}, \ldots$,如果马氏链在第n步已经收敛了，于是我们就得到了$\pi$ (x) 的样本$x_n, x_{n+1} \ldots$

MCMC

细细致平稳条件

如果非周期马氏链的转移矩阵P和分布$\pi(x)$满足:

$$\pi(i)P_{ij} = \pi(j)P_{ji}$$

则$\pi(x)$是马氏链的平稳分布，上式被称为细致平稳条件.

假设我们已经有一个转移矩阵为Q马氏链$q(i,j)$表示从状态i转移到状态j 的概率，，显然，通常情况下

$$p(i)q(ij) \neq p(j)q(j,i)$$

我们可否对马氏链做一个改造，使得细致平稳条件成立呢?

MCMC

引入一个$\alpha(i,j)$我们希望

$$p(i)q(ij)\alpha(i,j) = p(j)q(j,i)\alpha(j,i)$$

取什么样的$\alpha(i,j)$以上等式能成立呢?

MCMC

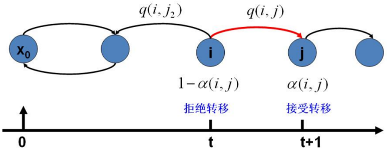
按照对称性，我们可以取

$$\alpha(i,j) = p(j)q(j,i), \qquad \alpha(j,i) = p(i)q(ij)$$

所以有

$$p(i)\underbrace{q(i,j)\alpha(i,j)}_{Q'(i,j)} = p(j)\underbrace{q(j,i)\alpha(j,i)}_{Q'(j,i)} \quad (**) \tag{1}$$

MCMC

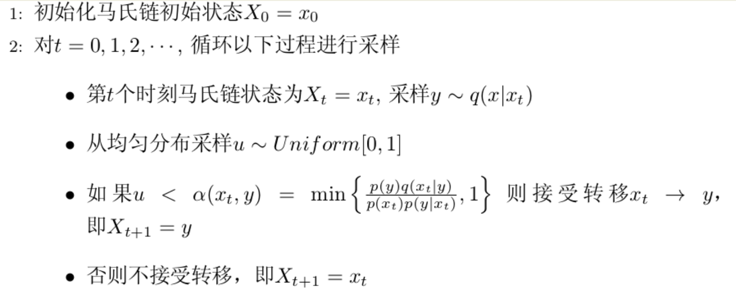


Algorithm 1 MCMC Sampling

```
初始化马氏链初始状态$X_0 = x_0$
对$t = 0, 1, 2, \cdots$，循环以下过程进行采样
    • 第$t$个时刻马氏链状态为$X_t = x_t$，采样$y \sim q(x|x_t)$
    • 从均匀分布采样$u \sim Uniform[0,1]$
    • 如果$u < \alpha(x_t, y) = p(y)q(x_t|y)$ 则接受转移$x_t \rightarrow y$，即$X_{t+1} = y$
    • 否则不接受转移，即$X_{t+1} = x_t$
```

Metropolis-Hastings

Algorithm 2 Metropolis-Hastings Sampling



```
初始化马氏链初始状态$X_0 = x_0$
对$t = 0, 1, 2, \cdots$, 循环以下过程进行采样
    • 第$t$个时刻马氏链状态为$X_t = x_t$, 采样$y \sim q(x|x_t)$
    • 从均匀分布采样$u \sim Uniform[0, 1]$
    • 如果$u < \alpha(x_t, y) = \min\left\{\frac{p(y)q(x_t|y)}{p(x_t)p(y|x_t)}, 1\right\}$ 则接受转移$x_t \rightarrow y$, 即$X_{t+1} = y$
    • 否则不接受转移, 即$X_{t+1} = x_t$
```

Gibbs Sampling

考察x坐标相同的两个点$A(x_1, y_1), B(x_1, y_2)$ 我们发现

$$p(x_1, y_1)p(y_2|x_1) = p(x_1)p(y_1|x_1)p(y_2|x_1)$$
$$p(x_1, y_2)p(y_1|x_1) = p(x_1)p(y_2|x_1)p(y_1|x_1)$$

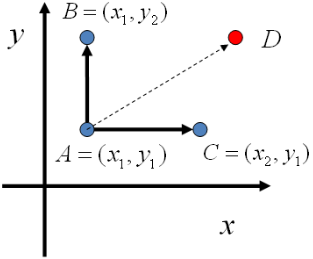
所以得到

$$p(x_1, y_1)p(y_2|x_1) = p(x_1, y_2)p(y_1|x_1) \quad (***) \tag{2}$$

即

$$p(A)p(y_2|x_1) = p(B)p(y_1|x_1)$$

Gibbs Sampling

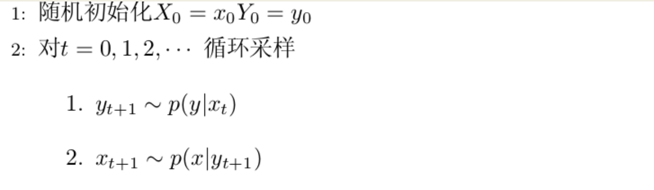


$$\begin{array}{ll}
Q(A\to B)=p(y_B|x_1) & \text{如果}\quad x_A=x_B=x_1\\
Q(A\to C)=p(x_C|y_1) & \text{如果}\quad y_A=y_C=y_1\\
Q(A\to D)=0 & \text{其它}
\end{array}$$

$$p(X)Q(X\to Y)=p(Y)Q(Y\to X)$$

Gibbs Sampling

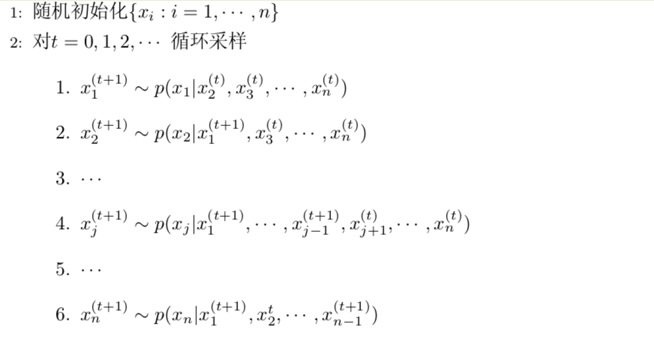
2-Gibbs Sampling



```
随机初始化$X_0 = x_0 Y_0 = y_0$
对$t = 0, 1, 2, \cdots$ 循环采样
    1. $y_{t+1} \sim p(y|x_t)$
    2. $x_{t+1} \sim p(x|y_{t+1})$
```

Gibbs Sampling

n-Gibbs Sampling



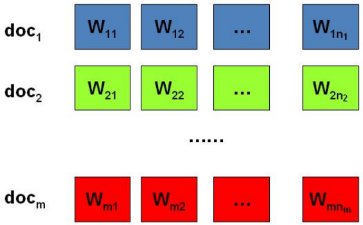
```
随机初始化$\{x_i : i = 1, \cdots, n\}$
对$t = 0, 1, 2, \cdots$ 循环采样
    1. $x_1^{(t+1)} \sim p(x_1 | x_2^{(t)}, x_3^{(t)}, \cdots, x_n^{(t)})$
    2. $x_2^{(t+1)} \sim p(x_2 | x_1^{(t+1)}, x_3^{(t)}, \cdots, x_n^{(t)})$
    3. $\cdots$
    4. $x_j^{(t+1)} \sim p(x_j | x_1^{(t+1)}, \cdots, x_{j-1}^{(t+1)}, x_{j+1}^{(t)}, \cdots, x_n^{(t)})$
    5. $\cdots$
    6. $x_n^{(t+1)} \sim p(x_n | x_1^{(t+1)}, x_2^{t}, \cdots, x_{n-1}^{(t+1)})$
```

Outline

1. Supplementary knowledge
   - statistical distribution
     - Binomal and Multinomal
     - Gama
     - Beta
     - Dirichlet
   - Bayes Estimation
     - 贝叶斯公式
     - 几个概念
   - Statistical Sampling
     - Classical Sampling
     - MCMC Sampling
     - Gibbs Sampling
2. Text Model
   - Unigram Model
   - Topic Model and PLSA
   - LDA(Latent Dirichlet Allocation)

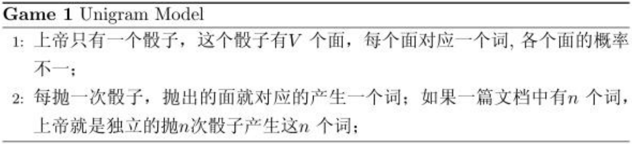
每篇文档从人的观察来说就是有序的词的序列

$$d = (w_1, w_2, \cdots, w_n)$$


$\mathbf{doc}_1$ W_{11} W_{12} ... W_{1n_1}
$\mathbf{doc}_2$ W_{21} W_{22} ... W_{2n_2}
......
$\mathbf{doc}_m$ W_{m1} W_{m2} ... W_{mn_m}

Unigram Model

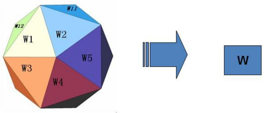
假设我们的词典中一共有V个词$v_1, v_2, \cdots v_V$



Game 1 Unigram Model

1: 上帝只有一个骰子，这个骰子有V 个面，每个面对应一个词, 各个面的概率不一；

2: 每抛一次骰子，抛出的面就对应的产生一个词；如果一篇文档中有n 个词，上帝就是独立的抛n次骰子产生这n 个词；



$\vec{p}$ $w \sim \mathrm{Mult}(w|\vec{p})$

Unigram Model

对于一篇文档$d = \overrightarrow{w} = (w_1, w_2, \cdots, w_n)$，该文档被生成的概率就是

$$p(\overrightarrow{w}) = p(w_1, w_2, \cdots, w_n) = p(w_1)p(w_2)\cdots p(w_n)$$

所以如果语料中有多篇文档$\mathcal{W} = (\overrightarrow{w_1}, \overrightarrow{w_2}, \ldots, \overrightarrow{w_m})$，则该语料的概率是

$$p(\mathcal{W}) = p(\overrightarrow{w_1})p(\overrightarrow{w_2})\cdots p(\overrightarrow{w_m})$$

Unigram Model

假设语料中总的词频是N，在所有的N个词中,如果我们关注每个词v_i的发生次数n_i，那么$\overrightarrow{n} = (n_1, n_2, \cdots, n_V)$正好是一个多项分布

$$p(\overrightarrow{n}) = Mult(\overrightarrow{n}|\overrightarrow{p}, N) = \binom{N}{\overrightarrow{n}} \prod_{k=1}^{V} p_k^{n_k}$$

此时，语料的概率是

$$p(\mathcal{W}) = p(\overrightarrow{w_1})p(\overrightarrow{w_2}) \cdots p(\overrightarrow{w_m}) = \prod_{k=1}^{V} p_k^{n_k}$$

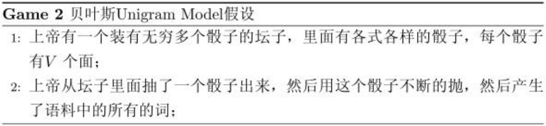
Unigram Model

我们很重要的一个任务就是估计上帝拥有的这个骰子的各个面的概率是多大，按照统计学家中频率派的观点，使用最大似然估计最大化$P(\mathcal{W})$，于是参数p_i的估计值就是

$$\hat{p}_i = \frac{n_i}{N}.$$

贝叶斯Unigram Model

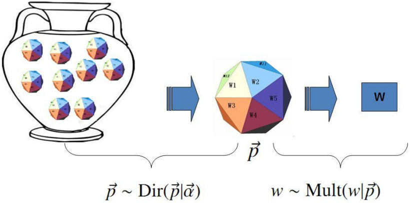
在贝叶斯学派看来，一切参数都是随机变量，以上模型中的骰子不是唯一固定的，它也是一个随机变量。所以按照贝叶斯学派的观点，上帝是按照以下的过程在玩游戏的



Game 2 贝叶斯Unigram Model假设

1: 上帝有一个装有无穷多个骰子的坛子，里面有各式各样的骰子，每个骰子有V 个面；

2: 上帝从坛子里面抽了一个骰子出来，然后用这个骰子不断的抛，然后产生了语料中的所有的词；



$\vec{p} \sim \mathrm{Dir}(\vec{p}|\vec{\alpha})$ $w \sim \mathrm{Mult}(w|\vec{p})$

贝叶斯Unigram Model

$$p(\overrightarrow{n}) = Mult(\overrightarrow{n}|\overrightarrow{p}, N)$$

$$Dir(\overrightarrow{p}|\overrightarrow{\alpha}) = \frac{1}{\Delta(\overrightarrow{\alpha})}\prod_{k=1}^{V} p_k^{\alpha_k-1} \quad \overrightarrow{\alpha} = (\alpha_1, \cdots, \alpha_V)$$

此处，$\Delta(\overrightarrow{\alpha})$ 就是归一化因子$Dir(\overrightarrow{\alpha})$，即

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

可以推出后验分布是

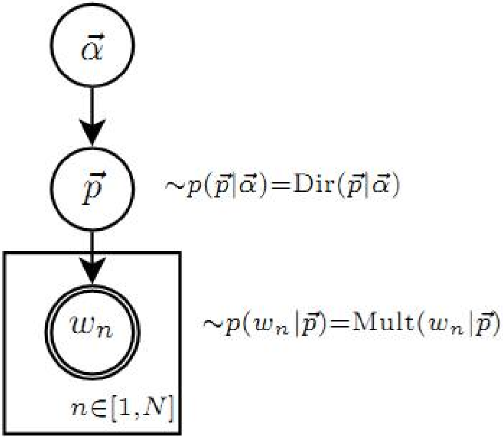
$$p(\overrightarrow{p}|\mathcal{W}, \overrightarrow{\alpha}) = Dir(\overrightarrow{p}|\overrightarrow{n} + \overrightarrow{\alpha}) = \frac{1}{\Delta(\overrightarrow{n} + \overrightarrow{\alpha})}\prod_{k=1}^{V} p_k^{n_k+\alpha_k-1} d\overrightarrow{p}$$

贝叶斯Unigram Model

对每一个p_i，我们用下式做参数估计$\hat{p}_i = \frac{n_i+\alpha_i}{\sum_{i=1}^{V}(n_i+\alpha_i)}$进一步，我们可以计算出文本语料的产生概率为

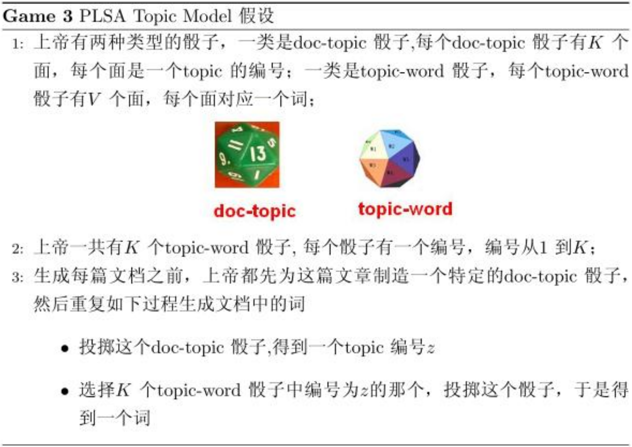
$$\begin{aligned} p(\mathcal{W}|\overrightarrow{\alpha}) &= \int p(\mathcal{W}|\overrightarrow{p})p(\overrightarrow{p}|\overrightarrow{\alpha})d\overrightarrow{p} \\ &= \int \prod_{k=1}^{V} p_k^{n_k} Dir(\overrightarrow{p}|\overrightarrow{\alpha})d\overrightarrow{p} \\ &= \int \prod_{k=1}^{V} p_k^{n_k} \frac{1}{\Delta(\overrightarrow{\alpha})} \prod_{k=1}^{V} p_k^{\alpha_k-1} d\overrightarrow{p} \\ &= \frac{1}{\Delta(\overrightarrow{\alpha})} \int \prod_{k=1}^{V} p_k^{n_k+\alpha_k-1} d\overrightarrow{p} \\ &= \frac{\Delta(\overrightarrow{n}+\overrightarrow{\alpha})}{\Delta(\overrightarrow{\alpha})} \end{aligned} \tag{3}$$

贝叶斯Unigram Model


$\vec{\alpha}$
$\vec{p}$
$\sim p(\vec{p}|\vec{\alpha})=\mathrm{Dir}(\vec{p}|\vec{\alpha})$
w_n
$\sim p(w_n|\vec{p})=\mathrm{Mult}(w_n|\vec{p})$
$n\in[1,N]$

Topic Model and PLSA

人类思考和写文章的行为都可以认为是上帝的行为，那么在PLSA 模型中，上帝是按照如下的游戏规则来生成文本的。



Game 3 PLSA Topic Model 假设

1: 上帝有两种类型的骰子，一类是doc-topic 骰子,每个doc-topic 骰子有K 个面，每个面是一个topic 的编号；一类是topic-word 骰子，每个topic-word 骰子有V 个面，每个面对应一个词；

doc-topic topic-word

2: 上帝一共有K 个topic-word 骰子, 每个骰子有一个编号，编号从1 到K；

3: 生成每篇文档之前，上帝都先为这篇文章制造一个特定的doc-topic 骰子，然后重复如下过程生成文档中的词

- 投掷这个doc-topic 骰子,得到一个topic 编号z
- 选择K 个topic-word 骰子中编号为z的那个，投掷这个骰子，于是得到一个词

Topic Model and PLSA

游戏中的K个topic-word 骰子，我们可以记为$\vec{\varphi}_1, \cdots, \vec{\varphi}_K$，对于包含$M$篇文档的语料$C = (d_1, d_2, \cdots, d_M)$中的每篇文档$d_m$，都会有一个特定的doc-topic骰子$\vec{\theta}_m$，所有对应的骰子记为$\vec{\theta}_1, \cdots, \vec{\theta}_M$。为了方便，我们假设每个词$w$都是一个编号，对应到topic-word 骰子的面。于是在PLSA 这个模型中，第m篇文档d_m 中的每个词的生成概率为

$$p(w|d_m) = \sum_{z=1}^{K} p(w|z)p(z|d_m) = \sum_{z=1}^{K} \varphi_{zw}\theta_{mz}$$

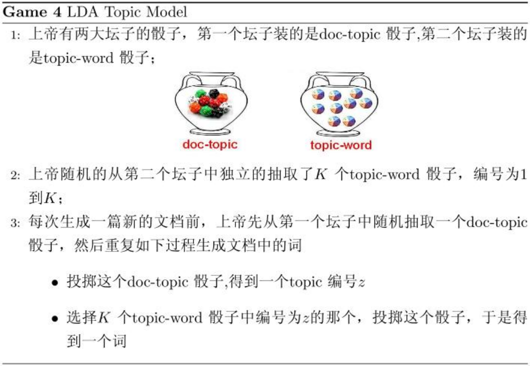
所以整篇文档的生成概率为

$$p(\vec{w}|d_m) = \prod_{i=1}^{n}\sum_{z=1}^{K} p(w_i|z)p(z|d_m) = \prod_{i=1}^{n}\sum_{z=1}^{K} \varphi_{zw_i}\theta_{dz}$$

求解PLSA 这个Topic Model 的过程汇总，模型参数并容易求解，可以使用著名的EM 算法进行求得局部最优解

LDA Model

对于上述的PLSA 模型，贝叶斯学派显然是有意见的，doc-topic 骰子$\vec{\theta}_m$和topic-word 骰子$\vec{\varphi}_k$都是模型中的参数，参数都是随机变量，怎么能没有先验分布呢？在LDA 模型中, 上帝是按照如下的规则玩文档生成的游戏的


doc-topic
topic-word

Game 4 LDA Topic Model

1: 上帝有两大坛子的骰子，第一个坛子装的是doc-topic 骰子,第二个坛子装的是topic-word 骰子；

2: 上帝随机的从第二个坛子中独立的抽取了K 个topic-word 骰子，编号为1到K；

3: 每次生成一篇新的文档前，上帝先从第一个坛子中随机抽取一个doc-topic 骰子，然后重复如下过程生成文档中的词

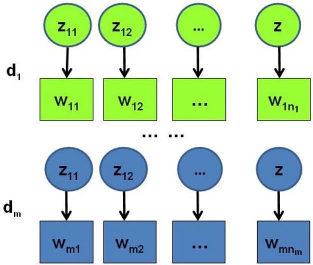
- 投掷这个doc-topic 骰子,得到一个topic 编号z
- 选择K 个topic-word 骰子中编号为z的那个，投掷这个骰子，于是得到一个词

LDA Model

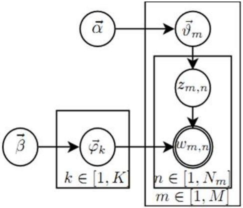
假设语料库中有M篇文档，所有的的word和对应的topic如下表示

$$\vec{\mathbf{w}} = (\vec{w}_1, \cdots, \vec{w}_M)$$
$$\vec{\mathbf{z}} = (\vec{z}_1, \cdots, \vec{z}_M)$$

其中，$\vec{w}_m$表示第m篇文档中的词，$\vec{z}_m$示这些词对应的topic 编号。



LDA Model



这个概率图可以分解为两个主要的物理过程:

- $\overrightarrow{\alpha} \to \overrightarrow{\theta}_m \to z_{m,n}$，这个过程表示在生成第$m$篇文档的时候，先从第一个坛子中抽了一个doc-topic 骰子$\overrightarrow{\theta}_m$，然后投掷这个骰子生成了文档中第$n$个词的topic编号$z_{m,n}$
- $\overrightarrow{\beta} \to \overrightarrow{\varphi}_k \to w_{m,n}|k = z_{m,n}$，这个过程表示用如下动作生成语料中第$m$篇文档的第$n$个词：在上帝手头的$K$个topic-word 骰子$\overrightarrow{\varphi}_k$中，挑选编号为$k = z_{m,n}$的那个骰子进行投掷，然后生成word $w_{m,n}$；

LDA Model

由于语料中M篇文档的topics 生成过程相互独立，所以我们得到M个相互独立的Dirichlet-Multinomial 共轭结构，从而我们可以得到整个语料中topics 生成概率

$$\begin{aligned} p(\vec{\mathbf{z}}|\vec{\alpha}) &= \prod_{m=1}^{M} p(\vec{z}_m|\vec{\alpha}) \\ &= \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})} \qquad (*) \end{aligned}$$

LDA Model

而语料中K个topics 生成words 的过程相互独立，所以我们得到K个相互独立的Dirichlet-Multinomial 共轭结构，从而我们可以得到整个语料中词生成概率

$$\begin{aligned}
p(\overrightarrow{\mathbf{w}}|\overrightarrow{\mathbf{z}},\overrightarrow{\beta}) &= p(\overrightarrow{\mathbf{w}}'|\overrightarrow{\mathbf{z}}',\overrightarrow{\beta}) \\
&= \prod_{k=1}^{K} p(\overrightarrow{w}_{(k)}|\overrightarrow{z}_{(k)},\overrightarrow{\beta}) \\
&= \prod_{k=1}^{K} \frac{\Delta(\overrightarrow{n}_k+\overrightarrow{\beta})}{\Delta(\overrightarrow{\beta})} \qquad (**)
\end{aligned}$$

LDA Model

结合(*) 和(**) 于是我们得到

$$
\begin{aligned}
p(\overrightarrow{\mathbf{w}}, \overrightarrow{\mathbf{z}} | \overrightarrow{\alpha}, \overrightarrow{\beta}) &= p(\overrightarrow{\mathbf{w}} | \overrightarrow{\mathbf{z}}, \overrightarrow{\beta}) p(\overrightarrow{\mathbf{z}} | \overrightarrow{\alpha}) \\
&= \prod_{k=1}^{K} \frac{\Delta(\overrightarrow{n}_k + \overrightarrow{\beta})}{\Delta(\overrightarrow{\beta})} \prod_{m=1}^{M} \frac{\Delta(\overrightarrow{n}_m + \overrightarrow{\alpha})}{\Delta(\overrightarrow{\alpha})} (***)
\end{aligned}
$$

Gibbs Sampling

语料库$\vec{\mathbf{z}}$中的第i个词我们记为z_i，其中$i=(m,n)$是一个二维下标，对应于第m篇文档的第n个词，用$\neg i$表示去除下标为i的词。按照Gibbs Sampling 算法的要，我们要求得任一个坐标轴i对应的条件分布$p(z_i=k|\vec{\mathbf{z}}_{\neg i},\vec{\mathbf{w}})$则由贝叶斯法则，我们容易得到

$$p(z_i=k|\vec{\mathbf{z}}_{\neg i},\vec{\mathbf{w}}) \propto p(z_i=k,w_i=t|\vec{\mathbf{z}}_{\neg i},\vec{\mathbf{w}}_{\neg i})$$

$\vec{\theta}_m,\vec{\varphi}_k$ 的后验分布都是Dirichlet:

$$p(\vec{\theta}_m|\vec{\mathbf{z}}_{\neg i},\vec{\mathbf{w}}_{\neg i}) = Dir(\vec{\theta}_m|\vec{n}_{m,\neg i}+\vec{\alpha})$$
$$p(\vec{\varphi}_k|\vec{\mathbf{z}}_{\neg i},\vec{\mathbf{w}}_{\neg i}) = Dir(\vec{\varphi}_k|\vec{n}_{k\text{fi}\neg i}+\vec{\beta})$$

Gibbs Sampling

$$
\begin{aligned}
& p(z_i = k | \overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}) \propto p(z_i = k, w_i = t | \overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}) \\
& = \int p(z_i = k, w_i = t, \overrightarrow{\theta}_m, \overrightarrow{\varphi}_k | \overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}) d \overrightarrow{\theta}_m d \overrightarrow{\varphi}_k \\
& = \int p(z_i = k, \overrightarrow{\theta}_m | \overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}) p(w_i = t, \overrightarrow{\varphi}_k | \overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}) d \overrightarrow{\theta}_m d \overrightarrow{\varphi}_k \\
& = \int p(z_i = k | \overrightarrow{\theta}_m) p(\overrightarrow{\theta}_m | \overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}) p(w_i = t | \overrightarrow{\varphi}_k) p(\overrightarrow{\varphi}_k | \overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}) d \overrightarrow{\theta}_m d \overrightarrow{\varphi}_k \\
& = \int p(z_i = k | \overrightarrow{\theta}_m) Dir(\overrightarrow{\theta}_m | \overrightarrow{n}_{m,\neg i} + \overrightarrow{\alpha}) d \overrightarrow{\theta}_m \\
& \quad \int p(w_i = t | \overrightarrow{\varphi}_k) Dir(\overrightarrow{\varphi}_k | \overrightarrow{n}_{k,\neg i} + \overrightarrow{\beta}) d \overrightarrow{\varphi}_k \\
& = \int \theta_{mk} Dir(\overrightarrow{\theta}_m | \overrightarrow{n}_{m,\neg i} + \overrightarrow{\alpha}) d \overrightarrow{\theta}_m \cdot \int \varphi_{kt} Dir(\overrightarrow{\varphi}_k | \overrightarrow{n}_{k,\neg i} + \overrightarrow{\beta}) d \overrightarrow{\varphi}_k \\
& = E(\theta_{mk}) \cdot E(\varphi_{kt}) = \hat{\theta}_{mk} \cdot \hat{\varphi}_{kt}
\end{aligned}
$$

Gibbs Sampling

借助于前面介绍的Dirichlet 参数估计的公式，我们有

$$\hat{\theta}_{mk} = \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_{m,\neg i}^{(k)} + \alpha_k)}$$

$$\hat{\varphi}_{kt} = \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V}(n_{k,\neg i}^{(t)} + \beta_t)}$$

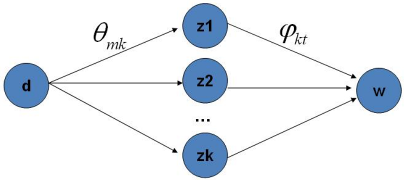
于是，我们最终得到了LDA 模型的Gibbs Sampling 公式

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

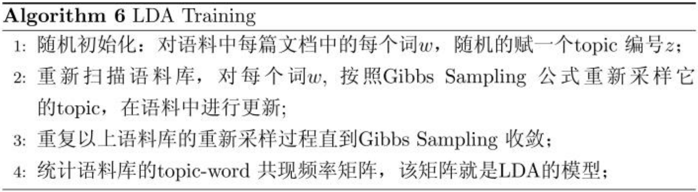
Training and Inference

有了LDA 模型，当然我们的目标有两个

- 估计模型中的参数$\vec{\varphi}_1, \cdots, \vec{\varphi}_K$和$\vec{\theta}_1, \cdots, \vec{\theta}_M$
- 对于新来的一篇文档doc_{new}，我们能够计算这篇文档的topic分布$\vec{\theta}_{new}$



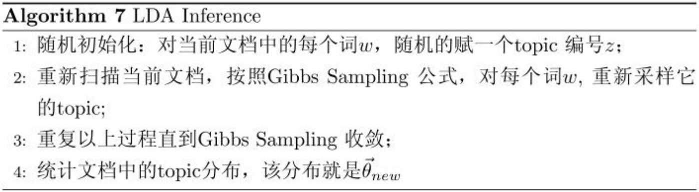
Training and Inference



Algorithm 6 LDA Training

```
随机初始化：对语料中每篇文档中的每个词$w$，随机的赋一个topic 编号$z$；
重新扫描语料库，对每个词$w$，按照Gibbs Sampling 公式重新采样它的topic，在语料中进行更新;
重复以上语料库的重新采样过程直到Gibbs Sampling 收敛；
统计语料库的topic-word 共现频率矩阵，该矩阵就是LDA的模型；
```

Training and Inference



Algorithm 7 LDA Inference

```
随机初始化：对当前文档中的每个词$w$，随机的赋一个topic 编号$z$；
重新扫描当前文档，按照Gibbs Sampling 公式，对每个词$w$, 重新采样它的topic;
重复以上过程直到Gibbs Sampling 收敛；
统计文档中的topic分布，该分布就是$\vec{\theta}_{new}$
```

Thank you for your attention!