第4_2节 贝叶斯网络理论及方法

—— 参数学习(2)

4.3 贝叶斯参数估计方法

- 给定网络拓扑结构 S 和训练样本集 D，利用先验知识 $p(\theta|S)$，确定贝叶斯网络模型各结点处的条件概率密度 $p(\theta|D,S)$。
  - $p(\theta|S)$ 为拓扑结构 S 下参数 θ 的先验概率。
  - $p(\theta|D,S)$ 为拓扑结构 S、训练样本集 D 下参数 θ 的后验概率。

4.3.1 参数θ的先验概率选取：

- 先验分布的选取原则：
  - 共轭分布
  - 杰弗莱原则
  - 最大熵原则

1. 共轭分布族

- 共轭分布，即要求后验分布与先验分布属于同一分布类型。它的一般描述为：

设样本 $x_1, x_2, \dots, x_n$ 对参数 θ 的条件分布为 $p(x_1, x_2, \dots, x_n|\theta)$，如果先验分布密度函数为 $\pi(\theta)$，它决定的后验密度 $\pi(\theta|x)$ 与它同属于一种类型，则称这种分布为共轭分布。

常用的共轭分布：

- 二项分布、多项分布
- 泊松分布
- 正态分布、多变量正态分布
- **Gamma**分布等

- 二项分布$B(n, \theta)$中的成功概率θ的共轭先验分布是贝塔分布$Be(\alpha, \beta)$;
- 泊松分布 $p(\theta)$中的均值θ的共轭先验分布是伽玛分布$Gamma(\alpha, \beta)$;
- 在方差已知时，正态均值θ的共轭先验分布是正态分布 $N(\mu, \sigma^2)$;
- 在均值已知时，正态方差σ^2的共轭先验分布是倒伽玛分布 $IGamma(\alpha, \beta)$。

表4.1 常用的一些共轭先验分布

总体分布	参数	共轭先验分布	后验分布的期望
正态分布 $N(\theta, \sigma^2)$	均值	正态分布 $N(\mu, \tau^2)$	$\dfrac{\tau^2 x + \mu\sigma^2}{\tau^2 + \sigma^2}$
正态分布 $N(\theta, \sigma^2)$	方差	倒 Γ 分布 $IGa(a, b)$	
二项分布 $B(n, \theta)$	成功概率	β 分布 $Be(\alpha, \beta)$	$\dfrac{\alpha + x}{\alpha + \beta + x + n}$
Poisson分布 $\pi(\theta)$	均值	Γ 分布 $Ga(\alpha, \beta)$	$\dfrac{\alpha + x}{\beta + 1}$
指数分布	均值的倒数	Γ 分布 $Ga(\alpha, \beta)$	

二项分布的成功概率θ 的共轭先验分布是贝塔分布

证明：设总体 $x \sim B(n, \theta)$，则 $p(x) \propto \theta^{x} \cdot (1-\theta)^{n-x}$。

再设θ 的先验分布为贝塔分布 $Be(\alpha,\beta)=\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$，其中参数$\alpha,\beta$已知。由此可以写出$\theta$的后验分布：

$$\pi(\theta \mid x) \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}, \quad 0<\theta<1$$

这是贝塔分布的核，其密度函数为：

$$\pi(\theta \mid x)=\frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)}\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}, \quad 0<\theta<1$$

2. 最大熵原则：

- 最大熵原则，指的是在参数θ变化范围内熵取得最大的分布。
- “均匀分布”是基于最大熵原则的。设随机变量 x 取有限个值 $a_1, a_2, \ldots\ldots, a_n$，相应的概率记为 $p_1, p_2, \ldots\ldots, p_n$，则 x 的熵 $H(x)$ 最大的充分必要条件是：

$$p_1 = p_2 = \ldots\ldots = p_n = \frac{1}{n}$$

贝叶斯假设：无信息先验分布应选取在θ取值范围内的均匀分布，即：

$$\pi(\theta)=\begin{cases} c, \theta \in \Omega \\ 0, \theta \notin \Omega \end{cases}$$

贝叶斯假设在很多情况下都是合理的。但是，当θ的取值范围为无限区间时，就无法在θ上定义一个正常的均匀分布。而且，贝叶斯假设不满足θ函数的不变性。

例4.3

- 考虑正态标准差σ，它的定义域是$(0,+\infty)$。若定义一个变换

$$\eta=\sigma^2\in(0,+\infty)$$

即η是正态方差，在$(0,+\infty)$上η与σ是一对一的关系。如果σ的无信息先验分布是常数，则η的无信息先验分布也应该是常数，并成比例。

但是，按照概率运算法则得到的结果并非如此。

若设$\pi(\sigma)$为σ的密度函数，那么依据概率运算法则，η的密度函数应为：

$$\pi(\eta)=\left|\frac{d\sigma}{d\eta}\right|\cdot\pi(\sigma)=\frac{1}{2\sqrt{\eta}}\pi(\sigma)$$

即η的无信息先验函数与σ的无信息先验分布的比例系数是$\eta^{-1/2}$，并非常数，因此与贝叶斯假设矛盾。

3. 杰弗莱原则：

- **Jeffreys提出的选取先验分布的原则是一种不变原理，较好地解决了Bayes假设中的一个矛盾：即若对参数 θ 选用均匀分布为先验分布，则其函数 $g(\theta)$ 的先验分布往往不是均匀的。**

杰弗莱原则：

- 设θ的先验分布为$\pi(\theta)$，$g(\theta)$是θ的函数。若按照与θ先验分布相同原则决定 $g(\theta)$ 的先验分布 $\pi[g(\theta)]$，则应有关系式：

$$\pi(\theta)=\pi[g(\theta)]\cdot\left|\frac{dg(\theta)}{d\theta}\right|$$

- 若选取的 $\pi(\theta)$ 符合上式，则θ及θ的函数 $g(\theta)$ 的先验分布是一致的。困难之处在于如何找到满足上式的 $\pi(\theta)$。

Jeffreys利用Fisher信息阵的不变性，找到了符合要求的$\pi(\theta)$

- **Fisher信息阵不变性定理**：设$I(\theta)$、$I[g(\theta)]$分别表示参数θ、$g(\theta)$的Fisher信息阵，则有：

$$|I(\theta)|^{1/2} = \left|\frac{d[g(\theta)]}{d\theta}\right| \cdot |I[g(\theta)]|^{1/2}$$

- 由此获得 Jeffreys (杰弗莱) 原则的先验密度$\pi(\theta)$ 的求取方法：$\pi(\theta) \propto |I(\theta)|^{1/2}$

Fisher信息阵定义：

1) 似然函数：设样本$x=(x_1, x_2, \ldots, x_n)$，它的条件概率为$p(x \mid \theta_i)$，$i=1, 2, \ldots, n$。则其似然函数为：

$$L(\theta \mid x)=\prod_{i=1}^{n} p(x \mid \theta_i)$$

2) 相应的对数似然函数为：

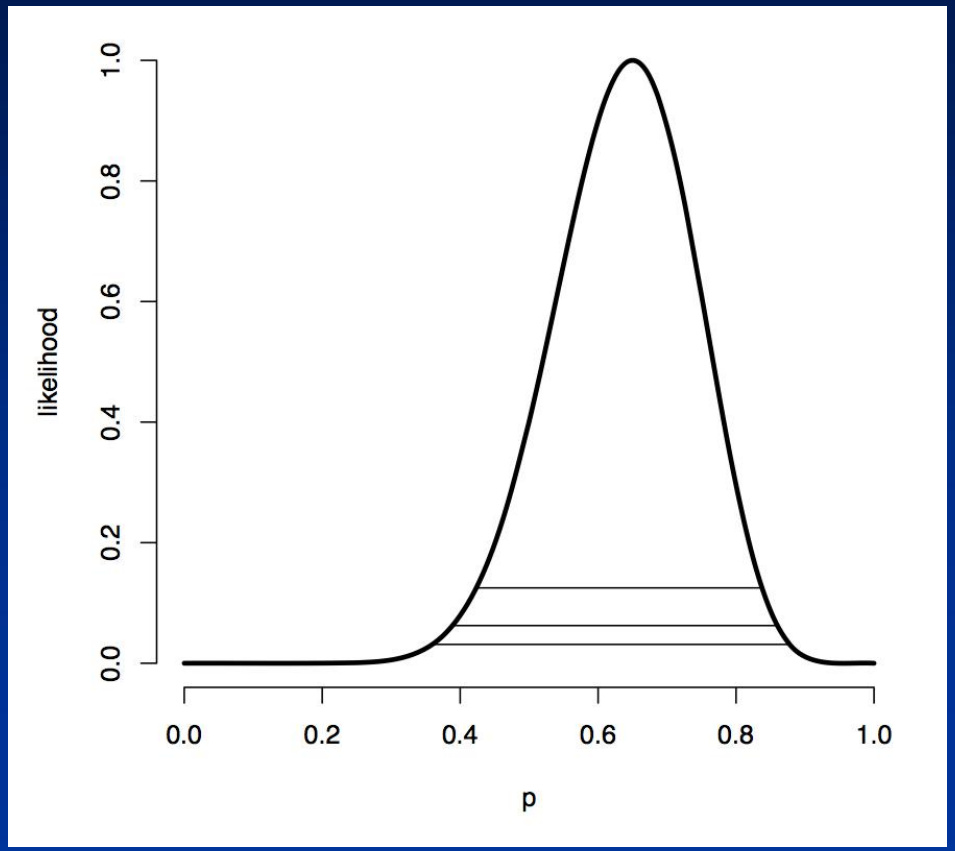
$$l(\theta)=\log L(\theta \mid x)$$

3) 观测Fisher信息量按下式计算：

$$\hat{I}(\theta) = -\frac{\partial^2}{\partial\theta} l(\theta)$$

4) 期望Fisher信息量则定义为：

$$I(\theta) = E\left[\hat{I}(\theta)\right]$$
$$= -E\left\{\frac{\partial^2}{\partial\theta^2} l(\theta)\right\}$$
$$= -E\left\{\frac{\partial^2}{\partial\theta^2} \log L(\theta|x)\right\}$$



对数似然函数越平滑，代表样本对于参数估计的能力越差；越高耸，代表样本对于参数估计的能力越好。此时样本的Fisher信息量 $I(\theta)$ 也越大。

图4.1 对数似然函数 $l(\theta)$ 图像

5) 对数似然函数 $l(\theta)$ 的一阶导数也叫计分函数 S：

$$S(x|\theta)=\frac{\partial[l(\theta)]}{\partial\theta}$$

一般情况下，可以证明：$E[S(x|\theta)]=0$

6) Fisher信息量是S 的二阶矩 $I(\theta)=E[S(x|\theta)^2]$，则有：

$$I(\theta)=E\left[S(x|\theta)^2\right]-E\left[S(x|\theta)\right]^2=Var\left[S(x|\theta)\right]$$

上式说明Fisher信息阵就是样本 x 似然函数MLE的方差。

一维和二维的Fisher信息阵：

- 假设待估计的未知参数$\theta = (\theta_1, \theta_2, \ldots, \theta_p)$，

一般情况下，很容易证明：

当$p = 1$时，它的Fisher信息阵为：$I(\theta) = -E\left(\frac{\partial l}{\partial \theta}\right)^2$

当 $p = 2$时，它的Fisher信息阵为：

$$I(\theta) = \begin{pmatrix} -E\left(\frac{\partial l}{\partial \theta_1}\right)^2 & -E\left(\frac{\partial l}{\partial \theta_1} \cdot \frac{\partial l}{\partial \theta_2}\right) \\ -E\left(\frac{\partial l}{\partial \theta_1} \cdot \frac{\partial l}{\partial \theta_2}\right) & -E\left(\frac{\partial l}{\partial \theta_2}\right)^2 \end{pmatrix}$$

例4.4

- 样本x 服从正态分布 $N(\mu,\sigma^2)$，其对数似然为：

$$l(\mu,\sigma)=-\frac{1}{2}ln(2\pi)-ln\,\sigma-\frac{(x-\mu)^2}{2\sigma^2}$$

它的二阶偏导数分别为：

$$\frac{\partial^2 l}{\partial\mu^2}=-\frac{1}{\sigma^2}$$

$$\frac{\partial^2 l}{\partial\mu\partial\sigma}=-\frac{2(x-\mu)}{\sigma^2}$$

$$\frac{\partial^2 l}{\partial\sigma^2}=\frac{1}{\sigma^2}-\frac{3(x-\mu)^2}{\sigma^4}$$

由于$E(x-\mu)=0$，$E(x-\mu)^2=\sigma^2$，故样本x的Fisher信息阵为：

$$I(\mu,\sigma)=\begin{bmatrix} -E\dfrac{\partial^2 l}{\partial\mu^2} & -E\dfrac{\partial^2 l}{\partial\mu\partial\sigma} \\ -E\dfrac{\partial^2 l}{\partial\mu\partial\sigma} & -E\dfrac{\partial^2 l}{\partial\sigma^2} \end{bmatrix}=\begin{bmatrix} \dfrac{1}{\sigma^2} & 0 \\ 0 & \dfrac{2}{\sigma^2} \end{bmatrix}$$

例4.5

- 设 $x_1, x_2, \ldots, x_n$ 是一组来自正态分布 $N(\mu, 1)$ 的独立同分布样本，求参数 μ 的先验分布。

解： 例4.2 已算得正态总体下的Fisher信息阵，因此可得到n 个样本分布下的Fisher信息阵为：

$$I_n(\mu, 1) = \begin{bmatrix} n & 0 \\ 0 & 2n \end{bmatrix}, \qquad |I_n(\mu, 1)| = 2n^2$$

所以，取 $\pi(\mu) \propto n / n = 1$

例4.6

- 设 $x_1, x_2, \ldots, x_n$ 是一组来自正态分布 $N(0, \sigma^2)$ 的独立同分布样本，求 σ 与 $\delta = \sigma^2$ 的先验分布。

解： 利用例4.2 得到的结果，可求得：

$$I_n(0, \sigma^2) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{bmatrix}, \qquad \left| I_n(0, \sigma^2) \right| = \frac{2n^2}{\sigma^4}$$

所以，取 $\pi(\sigma) \propto \frac{n}{\sigma^2} \cdot \frac{1}{n} = \frac{1}{\sigma^2}$，同理 $\pi(\delta) \propto \frac{1}{\delta^2}$

4.3.2 贝叶斯参数估计

- 在Bayesian 网络结构 *S* 已确定的条件下，其参数学习可分为：
  - 数据完整下的参数学习
  - 数据不完整(有缺损)下的参数学习

1. 单参数完整数据下的贝叶斯估计

未知参数θ 的估计就是计算其后验概率分布 $p(\theta \mid D)$，以及下一个样本的概率分布 $p(D_{m+1} \mid D)$。

利用θ的先验知识 $\pi(\theta)$，结合4.2.3小节中推导出来的似然函数 $L(\theta \mid D)$，用贝叶斯公式求出 θ 的后验分布 $\pi(\theta \mid D)$：

$$\pi(\theta \mid D) \propto \pi(\theta) \cdot L(\theta \mid D)$$

将图钉的二项似然函数表达代入上式，得到

$$\pi(\theta \mid D) \propto \theta^{m_h} \cdot (1-\theta)^{m_t} \cdot \pi(\theta)$$

先验分布$\pi(\theta)$，一般假设它是贝塔分布 $Be\,[\alpha_h, \alpha_t]$，即

$$\pi(\theta) = \frac{\Gamma(\alpha_t + \alpha_h)}{\Gamma(\alpha_t) \cdot \Gamma(\alpha_h)} \cdot \theta^{\alpha_h - 1} \cdot (1-\theta)^{\alpha_t - 1}$$

实际上就是做如下假设：关于θ的先验知识相当于已投掷图钉$\alpha_h + \alpha_t$次，其中 α_h 次头朝上， α_t 次尾朝上。

Beta分布概念：

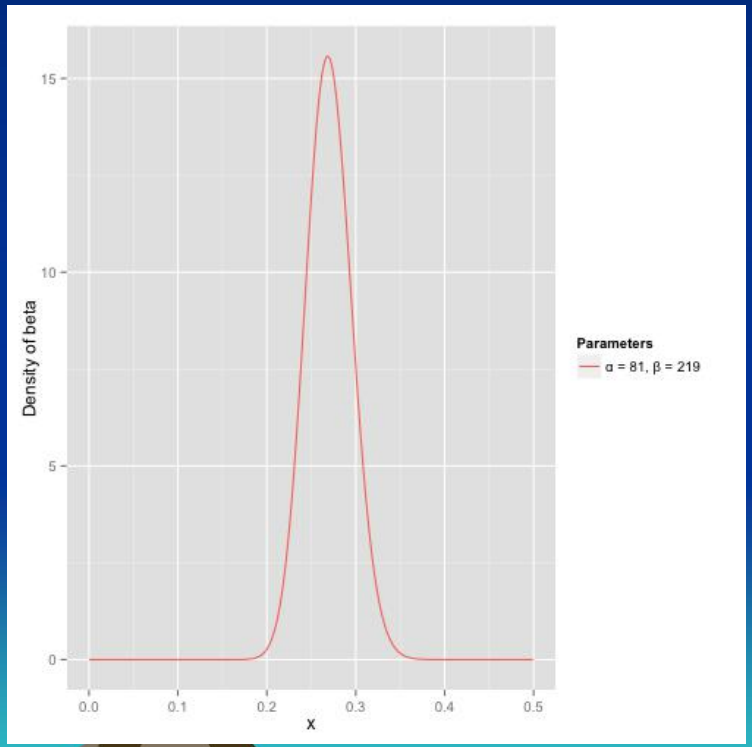
BETA分布可以看作对一个概率的概率分布估计，当你不知道一个东西的具体概率是多少时，它可以给出了所有概率出现的可能性大小。

例如：有一个棒球运动员，我们希望能够预测他在这一赛季中的棒球击球率是多少。可以直接计算棒球击球率，用击中的数除以击球数。

但是如果这个棒球运动员只打了一次，而且还命中了，那么他就击球率就是100%了，这显然是不合理的。因为根据棒球的历史信息，正常的击球率应该是0.215到0.36之间。

可以用一个二项分布的先验分布来表示。就是用BETA分布，表示在没有看到这个运动员打球之前，就有了一个大概的范围。

根据击球率应该是0.215到0.36之间，设置 $\alpha = 81$，$\beta = 219$。其均值：$\alpha/(\alpha+\beta)$ $=81/(81+219)=0.27$



Beta分布与二项分布的共轭先验性质：

二项分布的似然函数：

$$P(data|\theta) \propto \theta^{z}(1-\theta)^{N-z}$$

$$z = \sum_{i=1}^{N} X_i$$

beta分布

$$Beta(a,b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \propto \theta^{a-1}(1-\theta)^{b-1}$$

在beta分布中，B函数是一个标准化函数，它只是为了使得这个分布的概率密度积分等于1才加上的。

将上式代入贝叶斯公式，得到

$$\pi(\theta \mid D) \propto \theta^{m_h+\alpha_h-1} \cdot (1-\theta)^{m_t+\alpha_t-1}$$

从上式可看出，θ的后验分布 $\pi(\theta \mid D)$也是贝塔分布，其参数为 $m_k+\alpha_h$ 和 $m_t+\alpha_t$，即

$$\pi(\theta \mid D) = \frac{\Gamma(m_h+\alpha_h+m_t+\alpha_t)}{\Gamma(m_h+\alpha_h)\cdot\Gamma(m_t+\alpha_t)} \cdot \theta^{m_h+\alpha_h-1} \cdot (1-\theta)^{m_t+\alpha_t-1}$$

并有下一个样本的预测概率：

$$p(D_{m+1}=h \mid D) = \frac{m_h+\alpha_h}{m_h+\alpha_h+m_t+\alpha_t} = \frac{m_h+\alpha_h}{m+\alpha}$$

其中，$m=m_k+m_t$，$\alpha=\alpha_h+\alpha_t$。

当样本量 m 很小的时候，这个估计主要依赖于先验知识；

当样本量 m 增大时，这个估计越来越多地依赖于数据，越来越接近最大似然估计 $\frac{m_h}{m}$，而先验知识的影响逐渐减小。

2. 多参数下的贝叶斯参数学习

在节点 x_i 共有 r_i 个取值1, 2, …, r_i，其父节点 $\pi(x_i)$ 的取值共有 q_i 个组合的条件下，参数θ在数据 D 下的似然函数为：

$$L(\theta \mid D) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{m_{ijk}}$$

其中，θ_{ijk} 是 $x_i = k$ 和 $\pi(x_i) = j$ 时的参数θ；

m_{ijk} 是数据 D 中满足 $x_i = k$ 和 $\pi(x_i) = j$ 的样本数量。

根据贝叶斯公式，有

$$\pi(\theta \mid D) \propto \pi(\theta) \cdot \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{m_{ijk}}$$

假设θ的先验分布 $\pi(\ \theta\ )$服从Dirichlet分布，即

$$\pi(\theta) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_s)}{\Gamma(\alpha_1) \cdot \ldots \cdot \Gamma(\alpha_s)} \cdot \theta_1^{\alpha_1 - 1} \ldots \cdot \theta_s^{\alpha_s - 1}$$

其中$\Gamma(\ )$是Γ函数，$\alpha_1, \ldots, \alpha_s$是 θ分布的参数，称为超参数。

Dirichlet(狄利克雷)分布：

DIRICHLET分布是BETA分布的多元推广。BETA分布是二项式分布的共轭分布，DIRICHLET分布是多项式分布的共轭分布。

$$P(\theta_1, \ldots, \theta_m) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{m} \theta_k^{\alpha_k - 1}$$

在贝叶斯独立性假设下，参数 θ 的先验分布有：

$$\pi(\theta)=\prod_{i=1}^{n}\pi(\theta_{i\bullet\bullet})=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\pi(\theta_{ij\bullet})\propto\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{\alpha_{ijk}-1}$$

把它代入贝叶斯公式，得到

$$\pi(\theta\mid D)\propto\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{m_{ijk}+\alpha_{ijk}-1}$$

也就是说，θ 的后验分布 $\pi(\theta\mid D)$也是一个Dirichlet分布：$D[m_{ij1}+\alpha_{ij1}, m_{ij2}+\alpha_{ij2}, \ldots, m_{ijri}+\alpha_{ijri}]$

由于 $\pi(\theta_{ij\cdot}|D)$服从 Dirichlet 分布，因此节点 x_i 在第 j 种父节点组合下，取第 k 个值的概率为：

$$\theta_{x_i=k|j}=\frac{m_{ijk}+\alpha_{ijk}}{\sum_{k=1}^{r_i}(m_{ijk}+\alpha_{ijk})}$$

上式得到的就是贝叶斯网络中，条件概率表里节点间在不同取值组合下的概率值。

而下一个样本D_{m+1}的概率为：

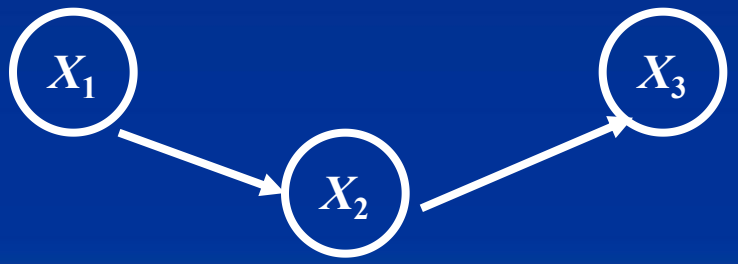
$$p(D_{m+1} \mid D) = \prod_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \chi(i,j,k : D_{m+1}) \frac{m_{ijk} + \alpha_{ijk}}{\sum_{k=1}^{r_i} (m_{ijk} + \alpha_{ijk})}$$

其中：

$$\chi(i,j,k : D_{m+1}) = \begin{cases} 1, & D_{m+1}\text{满足}x_i = k\text{且}\pi(x_i) = j \\ 0, & D_{m+1}\text{不满足上述条件} \end{cases}$$

例4.7

如图(*a*)所示贝叶斯网*S*，其中所有变量均取二值，1或2。设图 (*b*)是关于*S* 的一组i.i.d.数据，考虑计算*S* 参数的贝叶斯估计。



(*a*) 贝叶斯网 *S*

	X_1	X_2	X_3
D_1	1	1	1
D_2	2	2	2
D_3	1	1	2
D_4	2	2	2

(*b*) 完整数据

首先，假设先验分布 $\pi(\theta)$ 是乘积Dirichlet分布，且其超参数分别如下：

α^0_{1jk}

j \ k	1	2
1	2	2

α^0_{2jk}

j \ k	1	2
1	1	1
2	1	1

α^0_{3jk}

j \ k	1	2
1	1	1
2	1	1

θ的先验知识相当于如下的虚拟数据：

	X_1	X_2	X_3
$\mathbf{D'}_1$	1	1	1
$\mathbf{D'}_2$	1	2	1
$\mathbf{D'}_3$	2	1	2
$\mathbf{D'}_4$	2	2	2

加上原有的4个数据，可得：

	X_1	X_2	X_3
D_1	1	1	1
D_2	2	2	2
D_3	1	1	2
D_4	2	2	2

	X_1	X_2	X_3
D_5	1	1	1
D_6	1	2	1
D_7	2	1	2
D_8	2	2	2

由前面可知，后验分布 $p(\theta \mid D)$ 也是乘积Dirichlet分布，其超参数为 α：

α_{1jk}

j \ k	1	2
1	4	4

α_{2jk}

j \ k	1	2
1	3	1
2	1	3

α_{3jk}

j \ k	1	2
1	2	2
2	1	3

下一个样本D_{m+1}的条件分布 $p(D_{m+1} \mid D)$概率参数 θ_i 可以表示为如下表：

$p(x_1)$

x_1	1	2
	4/8	4/8

$p(x_2 \mid x_1)$

x_1 \ x_2	1	2
1	3/4	1/4
2	1/4	3/4

$p(x_3 \mid x_2)$

x_2 \ x_3	1	2
1	2/4	2/4
2	1/4	3/4

假设下一个样本 D_{m+1}是 $\{x_1=2, x_2=1, x_3=1\}$，则根据 $p(D_{m+1} \mid D)$的计算公式，可得该样本出现的概率为：

$$p(D_{m+1} \mid D) = \prod_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \chi(i,j,k : D_{m+1}) \frac{m_{ijk} + \alpha_{ijk}}{\sum_{k=1}^{r_i} (m_{ijk} + \alpha_{ijk})}$$

$$= \left[\chi(1,1,2) \cdot \frac{4}{8}\right] \times \left[\chi(2,2,1) \cdot \frac{1}{4}\right] \times \left[\chi(3,1,1) \cdot \frac{2}{4}\right] = 0.0625$$