PRML（Pattern Recognition And Machine Learning）读书会

第八章　Graphical Models

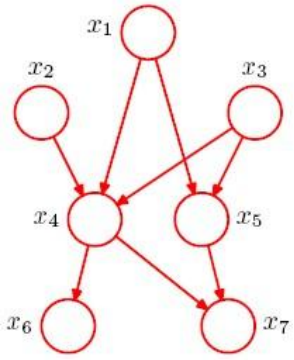
主讲人 网神

（新浪微博：@豆角茄子麻酱凉面）

QQ 群 177217565

读书会微信公众平台请扫描下面的二维码

网神(66707180) 18:52:10
今天的内容主要是：
1.贝叶斯网络和马尔科夫随机场的概念，联合概率分解，条件独立表示；2.图的概率推断 inference。
图模型是用图的方式表示概率推理 ，将概率模型可视化，方便展示变量之间的关系，概率图分为有向图和无向图。有向图主要是贝叶斯网络，无向图主要是马尔科夫随机场。对两类图，prml 都讲了如何将联合概率分解为条件概率，以及如何表示和判断条件依赖。
先说贝叶斯网络，贝叶斯网络是有向图，用节点表示随机变量，用箭头表示变量之间的依赖关系。一个例子:



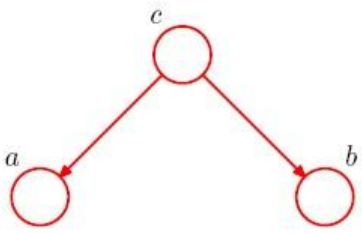
这是一个有向无环图，这个图表示的概率模型如下:

p(x1,x2,...x7)= $p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5).$ 形式化一下，贝叶斯网络表示的联合分布是：

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

其中 pa_k 是 xk 的所有父节点。

以上是贝叶斯网络将联合概率分解为条件概率的方法，比较直观易懂，就不多说了。下面说一下条件独立的表示和判断方法。条件独立是，给定 a,b,c 三个节点，如果 p(a,b|c)=p(a|c)p(b|c)，则说给定 c，a 和 b 条件独立。当然 a, b, c 也可以是三组节点，这里只以单个节点为例。用图表示，有三种情况 。
第一种情况如图：



c 位于两个箭头的尾部，称作 tail-to-tail，这种情况，c 未知的时候，a，b 是不独立的。c 已知的时候，a,b 条件独立。来看为什么，首先，这个图联合概率如下：
在 c 未知的时候,p(a,b)如下求解:

$$p(a,b) = \sum_c p(a|c)p(b|c)p(c).$$

可以看出，无法得出:p(a,b)=p(a)p(b)，所以 a,b 不独立。
如果 c 已知，则：

$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned}$$

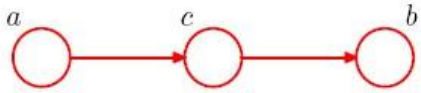
所以 a,b 条件独立于 c。条件独立用以下符号表示：

$$a \perp\!\!\!\perp b \mid c.$$

a,b 不独立的符号表示:：

$$a \not\perp\!\!\!\perp b \mid \emptyset$$

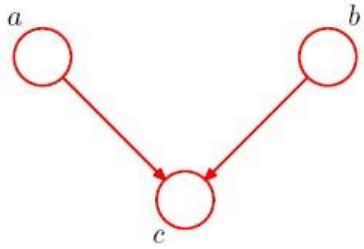
这是图表示的条件独立的第一种形式，叫做 tail-to-tail。第二种是 tail-to-head，如图：



这种情况也是 c 未知时，a 和 b 不独立。c 已知时，a 和 b 条件独立于 c，推导如下：

$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= \frac{p(a)p(c|a)p(b|c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned}$$

第三种情况是 head-to-head，如图：



这种情况反过来了，c 未知时，a 和 b 是独立的；但当 c 已知时，a 和 b 不满足条件独立，
因为：$p(a,b,c) = p(a)p(b)p(c|a,b).$
计算该概率的边界概率，得

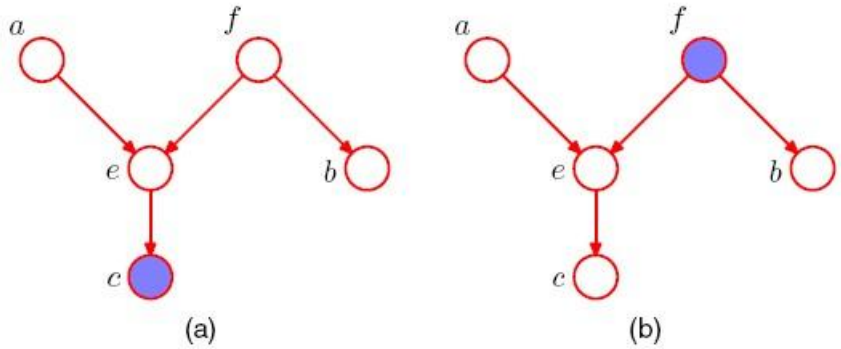
$$p(a,b) = p(a)p(b)$$

所以 a 和 b 相互独立.
但 c 已知时：

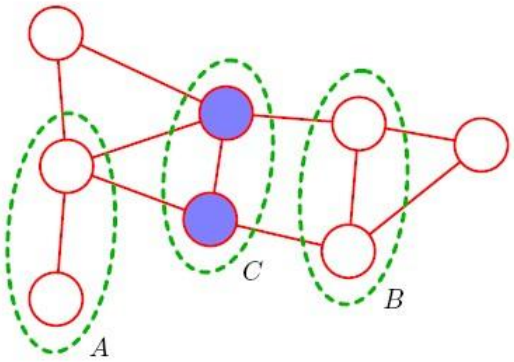
$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= \frac{p(a)p(b)p(c|a,b)}{p(c)} \end{aligned}$$

无法得到 p(a,b|c)=p(a|c)p(b|c)

将这三种情况总结，就是贝叶斯网络的一个重要概念，D-separation，这个概念的内容就是:
A,B,C 三组节点，如果 A 中的任意节点与 B 的任意节点的所有路径上，存在以下节点，就说 A 和 B 被 C 阻断:
1, A 到 B 的路径上存在 tail-to-tail 或 head-to-tail 形式的节点，并且该节点属于 C
2. 路径上存在 head-to-head 的节点，并且该节点不属于 C
举个例子：

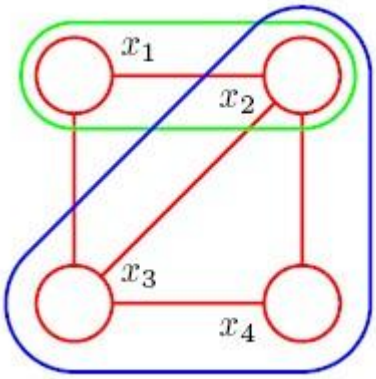


左边图上，节点 f 和节点 e 都不是 d-separation.因为 f 是 tail-to-tail，但 f 不是已知的，因此 f 不属于 C. e 是 head-to-head，但 e 的子节点 c 是已知的，所以 e 也不属于 C。
speedmancs<speedmancs@qq.com> 19:23:05
2 漏了一点，该节点包括所有后继
网神(66707180) 19:23:21
右边图，f 和 e 都是 d-separation.理由与上面相反.对，是漏了这一点。看到这个例子才想起来，这部分大家有什么问题没？
speedmancs<speedmancs@qq.com> 19:24:54
这个还是抽象了一些，我之前看的 prml 这一章，没看懂，后来看了 PGM 前三章，主要看了那个学生成绩的那个例子，就明白了。姑妄记之，其实蛮不好记的。讲得挺好的，继续。
网神(66707180) 19:27:14
因为有了这些条件独立的规则，可以将图理解成一个 filter。
既给定一系列随机变量，其联合分布 p(x1,x2...,xn)理论上可以分解成各种条件分布的乘积，但过一遍图，不满足图表示依赖关系和条件独立的分布就被过滤掉。所以图模型，用不同随机变量的连接表示各种关系，可以表示复杂的分布模型。
接下来是马尔科夫随机场，是无向图，也叫马尔科夫网络，马尔科夫网络也有条件独立属性。
用 MRF(malkov random field)表示马尔科夫网络，MRF 因为是无向的，所以不存在 tail-to-tail 这些概念。MRF 的条件独立如图：



如果 A 的任意节点和 B 的任意节点的任意路径上，都存在至少一个节点属于 C

speedmancs<speedmancs@qq.com> 19:33:16
无向图的条件独立 比有向图简单多了。
网神(66707180) 19:33:18
那么 A 和 B 条件独立于 C，可以理解为，如果 C 的节点都是已知的，就阻断了 A 和 B 的所有路径。
网神(66707180) 19:33:49
嗯 ,MRF 的 概率分解就概念比较多了 ,不像有向图那么直观 ,MRF 联合概率分解成条件概率。用到了 clique 的概念，我翻译成“团”，就是图的一个子图，子图上两两节点都有连接. 例如这个图，最大团有两个，分别是(x1,x2,x3)和(x2,x3,x4)：



MRF 的联合概率分解另一个概念是 potential function，联合概率分解成一系列 potential 函数的乘积:

$$p(\mathbf{x}) = \frac{1}{Z}\prod_{C}\psi_C(\mathbf{x}_C).$$

$\mathbf{x}_C$是一个最大团的所有节点，一个 potential 函数 $\psi_C(\mathbf{x}_C)$ ，是最大团的一个函数.

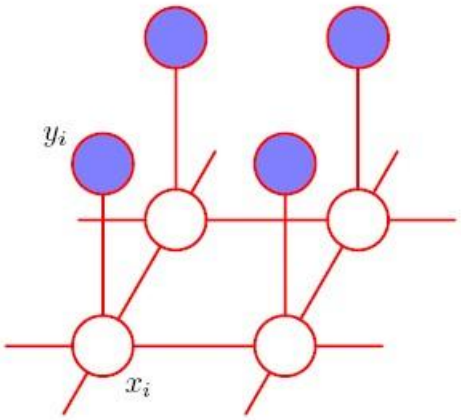
这个函数具体的定义是依赖具体应用的，一会举个例子.
上面式子里那个 Z 是 normalization 常量：

$$Z = \sum_{\mathbf{x}}\prod_{C}\psi_C(\mathbf{x}_C)$$

speedmancs<speedmancs@qq.com> 19:42:47
这个 Z 很麻烦
网神(66707180) 19:43:15
在这个式子里 $p(\mathbf{x}) = \frac{1}{Z}\prod_{C}\psi_C(\mathbf{x}_C).$ ，p(x)是一系列 potential 函数的乘积。换一种理解方式，定义将 potential 函数表示成指数函数:

$$\psi_C(\mathbf{x}_C) = \exp\{-E(\mathbf{x}_C)\}$$

这样 p(x)就可以表示成一系列 E(Xc)的和的指数函数，E(Xc)叫做能量函数，这么转换之后，可以将图理解成一个能量的集合，他的值等于各个最大团的能量的和.先举个例子看看 potential 函数和能量函数在具体应用中是什么样的，大家再讨论。
要把噪声图片尽量还原成 原图，用图的方式表示噪声图和还原后的图，每个像素点是一个节点:



上面那层 Yi，是噪声图，紫色表示这些是已知的，是观察值。下面那层 Xi 是未知的，要求出 Xi，使 Xi 作为像素值得到的图，尽量接近无噪声图片。每个 xi 的值，与 yi 相关，也与相邻的 xj 相关。这里边，最大团是(xi, yi)和(xi, xj)，两类最大团。

对于(xi, yi)，选择能量函数 E(xi,yi)= $-\eta x_i y_i$ ；对于(xi,xj)，选择能量函数 E(xi, xj)= $-\beta x_i x_j$

两个能量函数的意思都是，如果 xi 和 yi (或 xi 和 xj)的值相同，则能量小；如果不同，则能量大.
整个图的能量如下：

$$E(\mathbf{x},\mathbf{y}) = h\sum_i x_i - \beta\sum_{\{i,j\}} x_i x_j - \eta\sum_i x_i y_i$$

$h\sum_i x_i$ 是一个偏置项

speedmancs<speedmancs@qq.com> 19:56:55
偏置项是一种先验吧
网神(66707180) 19:57:28
有了这个能量函数，接下来就是求出 Xi，使得能量 E(x,y)最小。求最小，书上简单说了一下，我的理解也是用梯度下降类似的方法。
speedmancs<speedmancs@qq.com> 19:57:34
这里表示-1 的点更多吧。
网神(66707180) 19:58:20
对偏执项的作用，书上这么解释：
Such a term has the effect of biasing the model towards pixel values that have one particular sig n in preference to the other.
speedmancs<speedmancs@qq.com> 19:59:29
恩，对，让能量最小
网神(66707180) 19:59:39
为了使 E(x,y)尽量小，是尽量让 xi 选择-1，而不是 1。E(x,y)越小，得到的图就越接近无噪声的图，因为：

$$p(\mathbf{x},\mathbf{y}) = \frac{1}{Z}\exp\{-E(\mathbf{x},\mathbf{y})\}$$

所以 E(x,y)越小，p(x,y)就越大。

η<liyitan2144@163.com> 20:01:13

是不是可以这样看，$\beta \sum_{\{i,j\}} x_i x_j$ 表示平滑；$\eta \sum_i x_i y_i$ 表示似然。

网神(66707180) 20:02:18

liyitan2144 说得好，

speedmancs<speedmancs@qq.com> 20:02:33

恩，所以这是一个 产生式模型。

网神(66707180) 20:02:59

有向图和无向图的概率分解 和 条件独立 都说完了.有向图和无向图是可以互相转换的，有向图转换成无向图，如果每个节点都只有少于等于 1 个父节点，比较简单。如果有超过 1 个父节点，就需要在转换之后的无向图上增加一些边，来避免都是有向图上的一些关系，这部分就不细说了。

下面要说图的 inference 了。前面大家有啥要讨论的？先讨论一下吧。

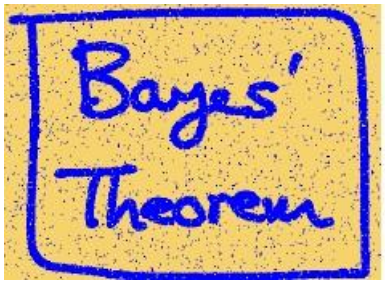
==========讨论==========

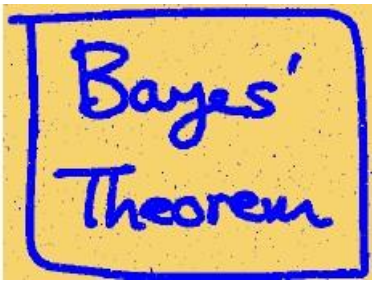
speedmancs<speedmancs@qq.com> 20:06:45

刚才那个例子中，那几个 beta, h 等参数如何得到？

网神(66707180) 20:07:30

那个是用迭代求解的方法 ,求得这几个参数 ,书上提到了两种方法 ,一种 ICM,iterated conditional modes 一种 max-product 方法 ,其中 max-product 方法效果比较好,在后面的 inference 一节里详细讲了这个方法，而 ICM 方法只是提了一下，原理没有细说.两种方法的效果看下图,左边是 ICM 的结果，右边是 max-product 的方法：





speedmancs<speedmancs@qq.com> 20:15:07

稍微插一句，刚才那个 denoise 的例子，最好的那个结果是 graph cut，而且那几个 beta 参数是事先固定了。

η<liyitan2144@163.com> 20:16:12

graph cut 是指？

kxkr<lxfkxkr@126.com> 20:16:30

the graph-cut algorithm on the right. ICM produces an image where 96% of the pixels agree with the original image, whereas the corresponding number for graph-cut is 99%.

kxkr<lxfkxkr@126.com> 20:17:18

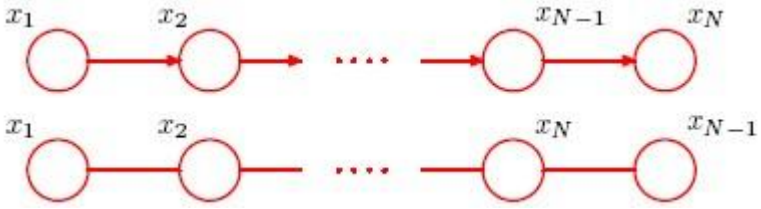
tribution. However, for certain classes of model, including the one given by (8.42), there exist efficient algorithms based on *graph cuts* that are guaranteed to find the global maximum (Greig *et al.*, 1989; Boykov *et al.*, 2001; Kolmogorov and Zabih, 2004). The lower right panel of Figure 8.30 shows the result of applying a graph-cut algorithm to the de-noising problem.

网神(66707180) 20:18:11
这个例子只是为了说明能量函数和潜函数.所以不一定是最佳方法，这两段截屏是 prml 上的，咋没看到捏
kxkr<lxfkxkr@126.com> 20:19:05
那几个参数如何得到 文中好像并没有说，只是说了 ICM、graph-cut 能够得到最后的去噪图像，第一段 在 figure8.30，第二个截图在 section8.4，figure8.32 下面。
网神(66707180) 20:20:15
看到了，先不管这个吧，主要知道能量函数是啥样的就行了
kxkr<lxfkxkr@126.com> 20:21:25
嗯
==========讨论结束==========
网神(66707180) 20:22:24
接着说 inference 了，inference 就是已知一些变量的值，求另一些变量的概率

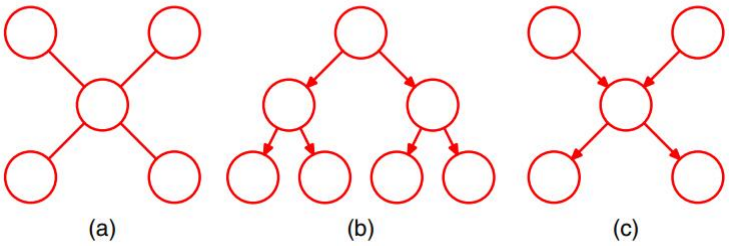
比如上图，已知 y，求 x 的概率
这个简单的图可以用典型的贝叶斯法则来求

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}.$$

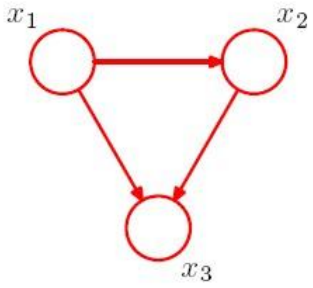
对于复杂点的情况，比如链式图：



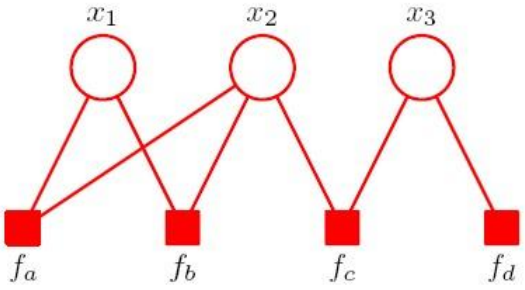
为了求 p(xn)，就是求 xn 的边界概率。这里都假设 x 的值是离散的。如果是连续的，就是积分，为了求这个边界概率，做这个累加动作，如果 x 的取值是 k 个，则要做 k 的(N-1)次方次计算，利用图结构，可以简化计算，这个简化方法就不讲了。
下面讲通用的图 inference 的方法，就是 factor graph 方法，链式图或树形图，都比较好求边界概率。
所以 factor graph 就是把复杂的图转换成树形图，针对树形图来求.先说一下什么是树形图，树形图有两种情况：
1. 每个节点只有一个父节点。
2. 如果有的节点有多个父节点，必须图上每个节点之间只有一条路径。
这是树形图的三种情况：



对于无向图，只要无环，都可以看做树形图；对于有向图，必须每两个节点之间只有一条路径；中间那种是典型的树.右边那种多个有多个节点的叫 polytree。
这种书结构的图，都比较好求边界概率. 具体怎么求，就不说了。
这里主要说怎么把复杂的图转换成树形图.，这种转换引入 factor 节点，从而将普通的图转换成 factor 图.
先看个例子：



对于这个有向图，p(x1,x2,x3)=p(x1)p(x2|x1)p(x3|x1,x2)，等号右边的三个概率作为三个 factor
每个 factor 作为一个节点，加入新的 factor 图中：

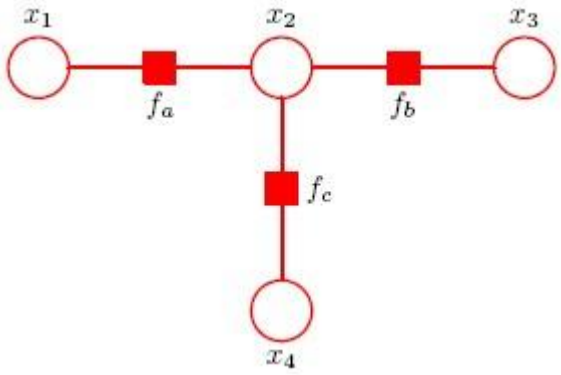


上面的 x1,x2,x3 是原图的随机变量 ，下面的 fa,fb,fc,fd 是 factor 节点，只有随即变量节点和 factor 节点之间有连接，每类节点自身互不连接，每个 factor 连接的变量节点，是相互依赖的节点。
kxkr<lxfkxkr@126.com> 20:42:10
就是一个 clique 中的节点吧
网神(66707180) 20:42:19
对，严格的说，也不是。
kxkr<lxfkxkr@126.com> 20:44:15
对于有向图 ？
网神(66707180) 20:45:10
x1,x2,x3 是一个最大团. 可以只用一个 factor 节点，如中间那个图
kxkr<lxfkxkr@126.com> 20:45:17
嗯

网神(66707180) 20:45:22
也可以用多个 factor 节点，如右边图，但什么情况下用一个 factor 什么情况用多个 factor，我没想明白
kxkr<lxfkxkr@126.com> 20:47:10
嗯，继续
网神(66707180) 20:47:21
转换成 factor 图后，就是树形图了，符合前面树形图的两种情况，这时候，要求一个节点或一组节点的边界概率，用一种叫做 sum-product 的方法，已求一个节点的边界概率为例.
η<liyitan2144@163.com> 20:49:50
网神(66707180) 20:46:31
但什么情况下用一个factor什么情况用多个factor 虽然也对具体的应用情景不清楚，但是一个 factor 能表达的信息比使用多个 factor 能表达的信息多.
网神(66707180) 20:51:06
单个的 factor 表达的信息多，而且节点越少，计算越简单，所以是不是尽量用少的 factor？max-product 求单个节点的边界概率，其思想是以该节点为 root。
η<liyitan2144@163.com> 20:52:38
但是，从设计模型的角度看，节点少了，一个节点设计的复杂程度就大了，不一定容易设计，拆解成多个 factor，每个 factor 都很简单，设计方便。
kxkr<lxfkxkr@126.com> 20:54:37
文中貌似倾向于一个单个的 factor

If we are given a distribution that is expressed in terms of an undirected graph, then we can readily convert it to a factor graph. To do this, we create variable nodes corresponding to the nodes in the original undirected graph, and then create additional factor nodes corresponding to the maximal cliques $\mathbf{x}_s$. The factors $f_s(\mathbf{x}_s)$ are then set equal to the clique potentials. Note that there may be several different factor graphs that correspond to the same undirected graph. These concepts are illustrated in Figure 8.41.

η<liyitan2144@163.com> 20:56:57
这貌似是在表达一个通用的方法，和是把大的 clique 的 factor 拆解成小的多个 factor 没有关系。
kxkr<lxfkxkr@126.com> 20:57:53
拆成多个 factor 是不是会造成参数变多，不容易求呢，继续吧，这个有待实践。
η<liyitan2144@163.com> 21:00:00
参数应该会变多吧，不过模型其实简单了，确实有待实践，我觉得这和具体应用有关，比如在一副图像里面，如果仅仅表达“像素”的相似度，我们完全可以使用小 factor。
网神(66707180) 21:00:39
我继续，prml 这章图模型只讲了基础的概念，没讲常用的图模型实例，HMM 和 CRF 这两大主流图模型方法还没讲，感觉不够直观.
我继续说 inference. 转换成 factor 图后，求节点 xi 的边缘概率。把 xi 作为 root 节点,把求 root 边界概率理解成一个信息(message)传递的过程，从叶子节点传递概率信息到 root 节点.传递的规则是：
从叶子节点开始，如果叶子是变量节点，发送 1 给父节点,如果叶子是 factor 节点，发送 f(x)给父节点.
对于非叶子节点，如果是变量节点，将其收到的 message 相乘，发给父节点，如果是 factor 节点，将其收到的 message 和自身 f(x)相乘，然后做一个 sum，发给父节点。
举个例子：



这个图中求 x3 的边缘概率，message 传递的过程是:

$$
\begin{aligned}
\mu_{x_1 \to f_a}(x_1) &= 1 \\
\mu_{f_a \to x_2}(x_2) &= \sum_{x_1} f_a(x_1, x_2) \\
\mu_{x_4 \to f_c}(x_4) &= 1 \\
\mu_{f_c \to x_2}(x_2) &= \sum_{x_4} f_c(x_2, x_4) \\
\mu_{x_2 \to f_b}(x_2) &= \mu_{f_a \to x_2}(x_2)\mu_{f_c \to x_2}(x_2) \\
\mu_{f_b \to x_3}(x_3) &= \sum_{x_2} f_b(x_2, x_3)\mu_{x_2 \to f_b}.
\end{aligned}
$$

$\mu_{x_1 \to f_a}(x_1)$ 中的 $x_1 \to f_a$ 表示从节点 x1 传递到 fa，最后 p(x3)是等于 $\mu_{f_b \to x_3}(x_3)$ ，因为只有一个 fb 节点向其传入信息，如果要求 x2 的边界概率,因为 x2 有三个节点出入信息,分别是：

$$\mu_{f_a \to x_2}(x_2) \quad \mu_{f_c \to x_2}(x_2) \quad \mu_{f_b \to x_2}(x_2)$$

所以 p(x2)就等于这三个信息的乘积

$$\widetilde{p}(x_2) = \mu_{f_a \to x_2}(x_2)\mu_{f_b \to x_2}(x_2)\mu_{f_c \to x_2}(x_2)$$

这个结果与边界分布的定义是相符的，x2 的边界定义如下：

$$\widetilde{p}(x_2) = \sum_{x_1}\sum_{x_3}\sum_{x_4}\widetilde{p}(\mathrm{x})$$

$\widetilde{p}(x_2) = \mu_{f_a \to x_2}(x_2)\mu_{f_b \to x_2}(x_2)\mu_{f_c \to x_2}(x_2)$ 通过这个 ,可以推导出上面的边界分布. 推导如下：

$$
\begin{aligned}
\widetilde{p}(x_2) &= \mu_{f_a \to x_2}(x_2)\mu_{f_b \to x_2}(x_2)\mu_{f_c \to x_2}(x_2) \\
&= \left[\sum_{x_1} f_a(x_1, x_2)\right]\left[\sum_{x_3} f_b(x_2, x_3)\right]\left[\sum_{x_4} f_c(x_2, x_4)\right] \\
&= \sum_{x_1}\sum_{x_2}\sum_{x_4} f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4) \\
&= \sum_{x_1}\sum_{x_3}\sum_{x_4}\widetilde{p}(\mathrm{x})
\end{aligned}
$$

上面就是用 sum-product 来求边缘分布的方法。

不知道讲的是否明白，不明白就看书吧，一起研究  今天就讲到这了，大家有啥问题讨论下。

huajh7(284696304) 21:34:52
补充几点，一是 temporal model ,如 Dynamical Bayesian newtwork(DBN), plate models ，这是图模型的表达能力; 二是 belief Bropagation ，包括 exact 和 approximation ，loopy 时的收敛性; Inference 包括 MCMC,变分法。这是图模型在 tree 和 graph 的推理能力。三是 Structure Learning ，BIC score 等。这是图模型的学习能力。