概率图模型

概率图模型使用图的方式表示概率分布。为了在图中添加各种概率，首先总结一下随机变量分布的一些规则：

$$
\begin{align}
&Sum\ Rule:p(x_1)=\int p(x_1,x_2)dx_2\\
&Product\ Rule:p(x_1,x_2)=p(x_1|x_2)p(x_2)\\
&Chain\ Rule:p(x_1,x_2,\cdots,x_p)=\prod\limits_{i=1}^pp(x_i|x_{i+1,x_{i+2}}\cdots x_p)\\
&Bayesian\ Rule:p(x_1|x_2)=\frac{p(x_2|x_1)p(x_1)}{p(x_2)}
\end{align}
$$

可以看到，在链式法则中，如果数据维度特别高，那么的采样和计算非常困难，我们需要在一定程度上作出简化，在朴素贝叶斯中，作出了条件独立性假设。在 Markov 假设中，给定数据的维度是以时间顺序出现的，给定当前时间的维度，那么下一个维度与之前的维度独立。在 HMM 中，采用了齐次 Markov 假设。在 Markov 假设之上，更一般的，加入条件独立性假设，对维度划分集合 A,B,C，使得 $X_A\perp X_B|X_C$。

概率图模型采用图的特点表示上述的条件独立性假设，节点表示随机变量，边表示条件概率。概率图模型可以分为三大理论部分：

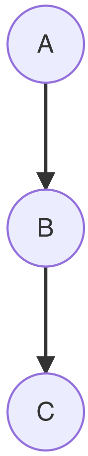
1. 表示：
   1. 有向图（离散）：贝叶斯网络
   2. 高斯图（连续）：高斯贝叶斯和高斯马尔可夫网路
   3. 无向图（离散）：马尔可夫网络
2. 推断
   1. 精确推断
   2. 近似推断
      1. 确定性近似（如变分推断）
      2. 随机近似（如 MCMC）
3. 学习
   1. 参数学习
      1. 完备数据
      2. 隐变量：E-M 算法
   2. 结构学习

有向图-贝叶斯网络

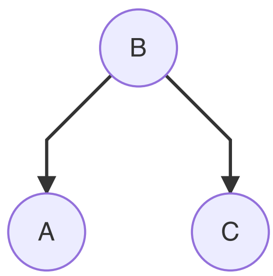
已知联合分布中，各个随机变量之间的依赖关系，那么可以通过拓扑排序（根据依赖关系）可以获得一个有向图。而如果已知一个图，也可以直接得到联合概率分布的因子分解：

$$p(x_1,x_2,\cdots,x_p)=\prod_{i=1}^p p(x_i|x_{parent(i)}) \tag{5}$$

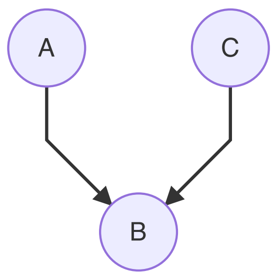
那么实际的图中条件独立性是如何体现的呢？在局部任何三个节点，可以有三种结构：



$$\begin{aligned} p(A,B,C)=p(A)p(B|A)p(C|B)=p(A)p(B|A)p(C|B,A) \\ \Longrightarrow p(C|B)=p(C|B,A) \\ \Leftrightarrow p(C|B)p(A|B)=p(C|A,B)p(A|B)=p(C,A|B) \\ \Longrightarrow C\perp A|B \end{aligned} \tag{6}$$



$$\begin{aligned} p(A,B,C)=p(A|B)p(B)p(C|B)=p(B)p(A|B)p(C|A,B) \\ \Longrightarrow p(C|B)=p(C|B,A) \\ \Leftrightarrow p(C|B)p(A|B)=p(C|A,B)p(A|B)=p(C,A|B) \\ \Longrightarrow C\perp A|B \end{aligned} \tag{7}$$



$$
\begin{aligned}
p(A,B,C)=p(A)p(C)p(B|C,A)=p(A)p(C|A)p(B|C,A)\\
\Longrightarrow p(C)=p(C|A)\\
\Leftrightarrow C\perp A
\end{aligned}
\tag{8}
$$

对这种结构，A,C 不与 B 条件独立。

从整体的图来看，可以引入 D 划分的概念。对于类似上面图 1和图 2的关系，引入集合A，B，那么满足 $A\perp B|C$ 的 C 集合中的点与 A,B 中的点的关系都满足图 1，2，满足图3 关系的点都不在 C 中。D 划分应用在贝叶斯定理中：

$$
p(x_i|x_{-i})=\frac{p(x)}{\int p(x)dx_i}=\frac{\prod\limits_{j=1}^p p(x_j|x_{parents(j)})}{\int\prod\limits_{j=1}^p p(x_j|x_{parents(j)})dx_i}
\tag{9}
$$

可以发现，上下部分可以分为两部分，一部分是和 x_i 相关的，另一部分是和 x_i 无关的，而这个无关的部分可以相互约掉。于是计算只涉及和 x_i 相关的部分。

与 x_i 相关的部分可以写成：

$$
p(x_i|x_{parents(i)})p(x_{child(i)}|x_i)
\tag{10}
$$

这些相关的部分又叫做 Markov 毯。

实际应用的模型中，对这些条件独立性作出了假设，从单一到混合，从有限到无限（时间，空间）可以分为：

1. 朴素贝叶斯，单一的条件独立性假设 $p(x|y)=\prod\limits_{i=1}^p p(x_i|y)$，在 D 划分后，所有条件依赖的集合就是单个元素。
2. 高斯混合模型：混合的条件独立。引入多类别的隐变量 $z_1,z_2,\cdots,z_k$， $p(x|z)=\mathcal{N}(\mu,\Sigma)$，条件依赖集合为多个元素。
3. 与时间相关的条件依赖
    1. Markov 链
    2. 高斯过程（无限维高斯分布）
4. 连续：高斯贝叶斯网络
5. 组合上面的分类
    - GMM 与时序结合：动态模型
        - HMM（离散）
        - 线性动态系统 LDS（Kalman 滤波）
        - 粒子滤波（非高斯，非线性）

无向图-马尔可夫网络（马尔可夫随机场）

无向图没有了类似有向图的局部不同结构，在马尔可夫网络中，也存在 D 划分的概念。直接将条件独立的集合 $x_A\perp x_B|x_C$ 划分为三个集合。这个也叫全局 Markov。对局部的节点，$x\perp(X-Neighbour(x))|Neighbour(x)$。这也叫局部 Markov。对于成对的节点：$x_i\perp x_j|x_{-i-j}$，其中 i,j 不能相邻。这也叫成对 Markov。事实上上面三个点局部全局成对是相互等价的。

有了这个条件独立性的划分，还需要因子分解来实际计算。引入团的概念：

> 团，最大团：图中节点的集合，集合中的节点之间相互都是连接的叫做团，如果不能再添加节点，那么叫最大团。

利用这个定义进行的 x 所有维度的联合概率分布的因子分解为，假设有 K 个团，Z 就是对所有可能取值求和：

$$p(x)=\frac{1}{Z}\prod\limits_{i=1}^{K}\phi(x_{ci}) \tag{11}$$

$$Z=\sum\limits_{x\in\mathcal{X}}\prod\limits_{i=1}^{K}\phi(x_{ci}) \tag{12}$$

其中 $\phi(x_{ci})$ 叫做势函数，它必须是一个正值，可以记为：

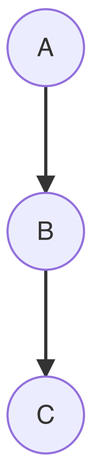
$$\phi(x_{ci})=\exp(-E(x_{ci})) \tag{13}$$

这个分布叫做 Gibbs 分布（玻尔兹曼分布）。于是也可以记为：$p(x)=\frac{1}{Z}\exp(-\sum\limits_{i=1}^KE(x_{ci}))$。这个分解和条件独立性等价（Hammesley-Clifford 定理），这个分布的形式也和指数族分布形式上相同，于是满足最大熵原理。

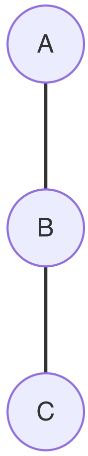
两种图的转换-道德图

我们常常想将有向图转为无向图，从而应用更一般的表达式。

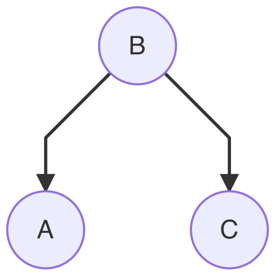
1. 链式：



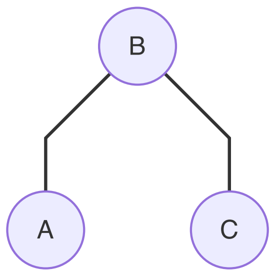
   直接去掉箭头，$p(a,b,c)=p(a)p(b|a)p(c|b)=\phi(a,b)\phi(b,c)$：



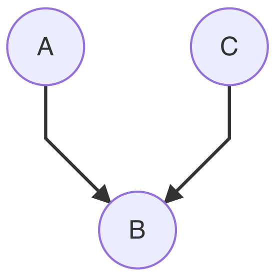
2. V 形：



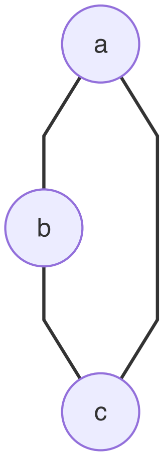
由于 $p(a,b,c)=p(b)p(a|b)p(c|b)=\phi(a,b)\phi(b,c)$，直接去掉箭头：



3. 倒 V 形：



由于 $p(a,b,c)=p(a)p(c)p(b|a,c)=\phi(a,b,c)$，于是在 a,c 之间添加线：



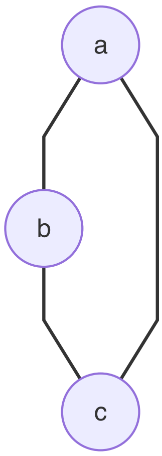
观察着三种情况可以概括为：

1. 将每个节点的父节点两两相连
2. 将有向边替换为无向边

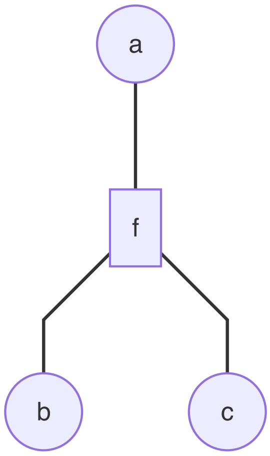
更精细的分解-因子图

对于一个有向图，可以通过引入环的方式，可以将其转换为无向图（Tree-like graph），这个图就叫做道德图。但是我们上面的 BP 算法只对无环图有效，通过因子图可以变为无环图。

考虑一个无向图：



可以将其转为：



其中 $f=f(a,b,c)$。因子图不是唯一的，这是由于因式分解本身就对应一个特殊的因子图，将因式分解：$p(x)=\prod\limits_s f_s(x_s)$ 可以进一步分解得到因子图。

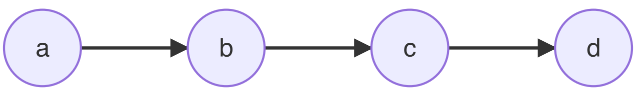
推断

推断的主要目的是求各种概率分布，包括边缘概率，条件概率，以及使用 MAP 来求得参数。通常推断可以分为：

1. 精确推断
    1. Variable Elimination(VE)
    2. Belief Propagation(BP, Sum-Product Algo)，从 VE 发展而来
    3. Junction Tree，上面两种在树结构上应用，Junction Tree 在图结构上应用
2. 近似推断
    1. Loop Belief Propagation（针对有环图）
    2. Mente Carlo Interference：例如 Importance Sampling，MCMC
    3. Variational Inference

推断-变量消除（VE）

变量消除的方法是在求解概率分布的时候，将相关的条件概率先行求和或积分，从而一步步地消除变量，例如在马尔可夫链中：



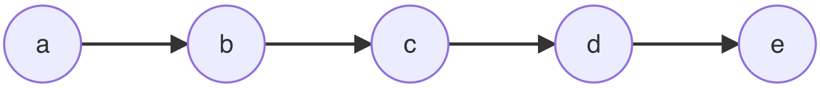
$$p(d)=\sum_{a,b,c}p(a,b,c,d)=\sum_c p(d|c)\sum_b p(c|b)\sum_a p(b|a)p(a) \tag{14}$$

变量消除的缺点很明显：

1. 计算步骤无法存储
2. 消除的最优次序是一个 NP-hard 问题

推断-信念传播（BP）

为了克服 VE 的第一个缺陷-计算步骤无法存储。我们进一步地对上面的马尔可夫链进行观察：

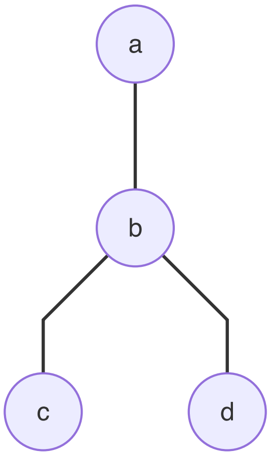


要求 $p(e)$，当然使用 VE，从 a 一直消除到 d，记 $\sum\limits_a p(a)p(b|a)=m_{a\to b(b)}$，表示这是消除 a 后的关于 b 的概率，类似地，记 $\sum\limits_b p(c|b)m_{a\to b}(b)=m_{b\to c}(c)$。于是 $p(e)=\sum\limits_d p(e|d)m_{b\to c}(c)$。进一步观察，对 $p(c)$：

$$p(c)=[\sum_b p(c|b)\sum_a p(b|a)p(a)]\cdot[\sum_d p(d|c)\sum_e p(e)p(e|d)] \tag{15}$$

我们发现了和上面计算 $p(e)$ 类似的结构，这个式子可以分成两个部分，一部分是从 a 传播过来的概率，第二部分是从 e 传播过来的概率。

一般地，对于图（只对树形状的图）：



这四个团（对于无向图是团，对于有向图就是概率为除了根的节点为1），有四个节点，三个边：

$$p(a,b,c,d)=\frac{1}{Z}\phi_a(a)\phi_b(b)\phi_c(c)\phi_d(d)\cdot\phi_{ab}(a,b)\phi_{bc}(c,b)\phi_{bd}(d,b) \tag{16}$$

套用上面关于有向图的观察，如果求解边缘概率 $p(a)$，定义 $m_{c\to b}(b)=\sum\limits_c \phi_c(c)\phi_{bc}(bc)$，$m_{d\to b}(b)=\sum\limits_d \phi_d(d)\phi_{bd}(bd)$，$m_{b\to a}(a)=\sum\limits_b \phi_{ba}(ba)\phi_b(b)m_{c\to b}(b)_{d\to b}m(b)$，这样概率就一步步地传播到了 a：

$$p(a)=\phi_a(a)m_{b\to a}(a) \tag{17}$$

写成一般的形式，对于相邻节点 i,j：

$$m_{j\to i}(i)=\sum_j \phi_j(j)\phi_{ij}(ij)\prod_{k\in Neighbour(j)-i} m_{k\to j}(j) \tag{18}$$

这个表达式，就可以保存计算过程了，只要对每条边的传播分别计算，对于一个无向树形图可以递归并行实现：

1. 任取一个节点 a 作为根节点

2. 对这个根节点的邻居中的每一个节点，收集信息（计算入信息）
3. 对根节点的邻居，分发信息（计算出信息）

推断-Max-Product 算法

在推断任务中，MAP 也是常常需要的，MAP 的目的是寻找最佳参数：

$$(\hat{a},\hat{b},\hat{c},\hat{d})=\mathop{argmax}_{a,b,c,d}p(a,b,c,d|E) \tag{19}$$

类似 BP，我们采用信息传递的方式来求得最优参数，不同的是，我们在所有信息传递中，传递的是最大化参数的概率，而不是将所有可能求和：

$$m_{j\to i}=\max_{j}\phi_j\phi_{ij}\prod_{k\in Neighbour(j)-i}m_{k\to j} \tag{20}$$

于是对于上面的图：

$$\max_a p(a,b,c,d)=\max_a\phi_a\phi_{ab}m_{c\to b}m_{d\to b} \tag{21}$$

这个算法是 Sum-Product 算法的改进，也是在 HMM 中应用给的 Viterbi 算法的推广。