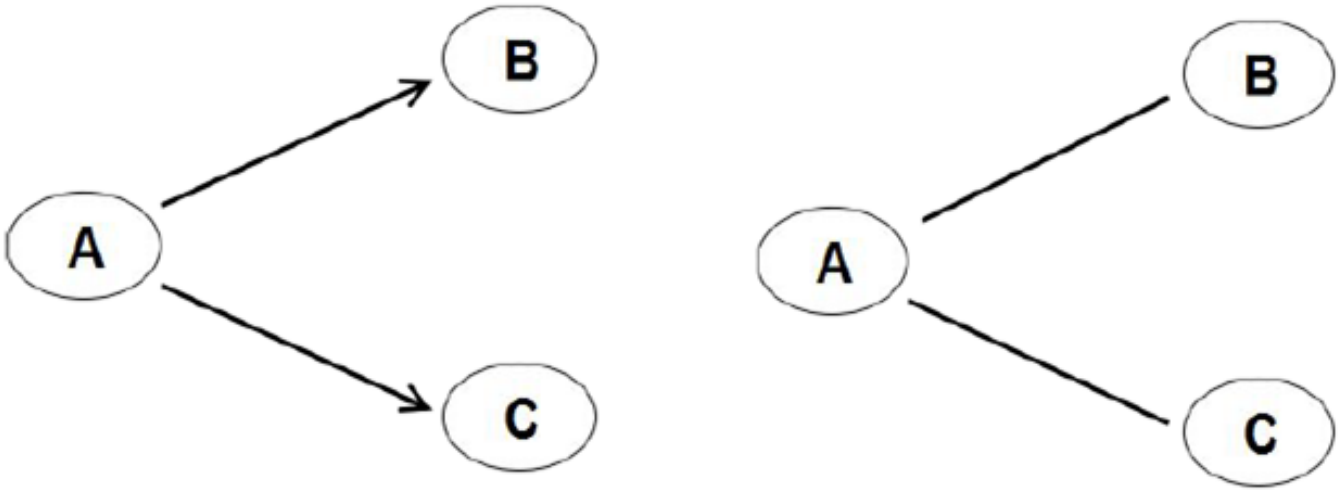
概率图模型

Probabilistic Graphical Models

- 概率图模型(PGM)
  - 图论+概率论=概率图模型: 是指一种用图结构来描述多元随机变量之间条件独立关系的概率模型
  - 节点：随机变量或一组随机变量
  - 连接弧：随机变量之间的依赖关系



- 图模型有三个基本问题：
  - 表示问题：对于一个概率模型，如何通过图结构来描述变量之间的依赖关系。
  - 推断问题：在已知部分变量时计算其它变量的概率分布。
  - 学习问题：图模型的学习包括图结构的学习(Automate Machine learning)和参数的学习。
- 有向图（因果关系）
  - 朴素贝叶斯（Naïve Bayes Classifier）
  - 贝叶斯网络（Basyesian Network）
  - 隐马尔科夫模型（Hidden Markov Model）
- 无向图（依赖关系）
  - 马尔可夫随机场（Markov Random Field）
  - 条件随机场（Conditional Random Field）

模型表示

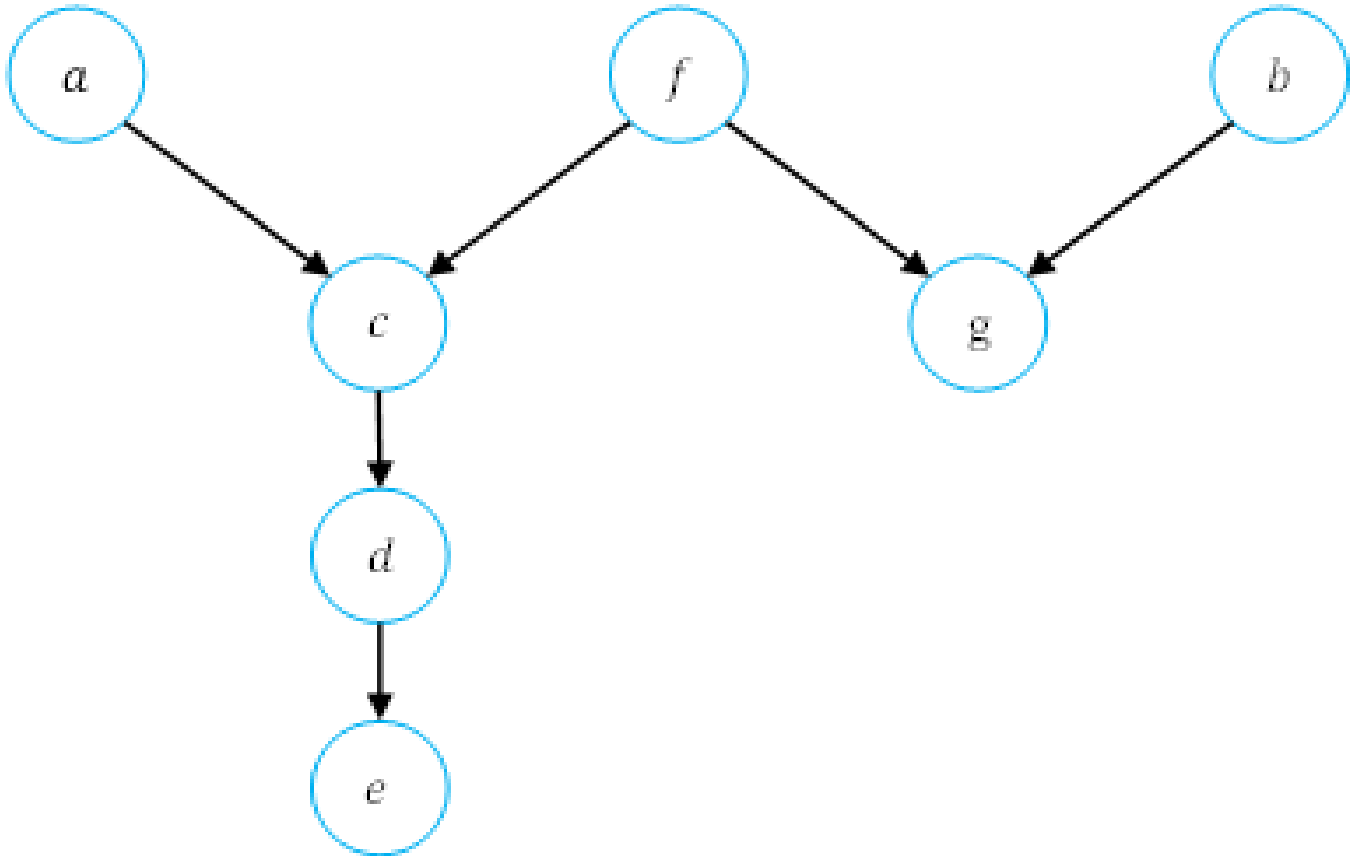
贝叶斯网络:

- 有向无环图$G = (V, E)$, 其中V表示有向图中节点的集合，与领域的随机变量一一对应；E表示图中有向边的集合，反映了变量之间的因果依赖关系。
- 父节点到子节点的条件概率分布。

贝叶斯网络定义的紧凑的联合分布表示为

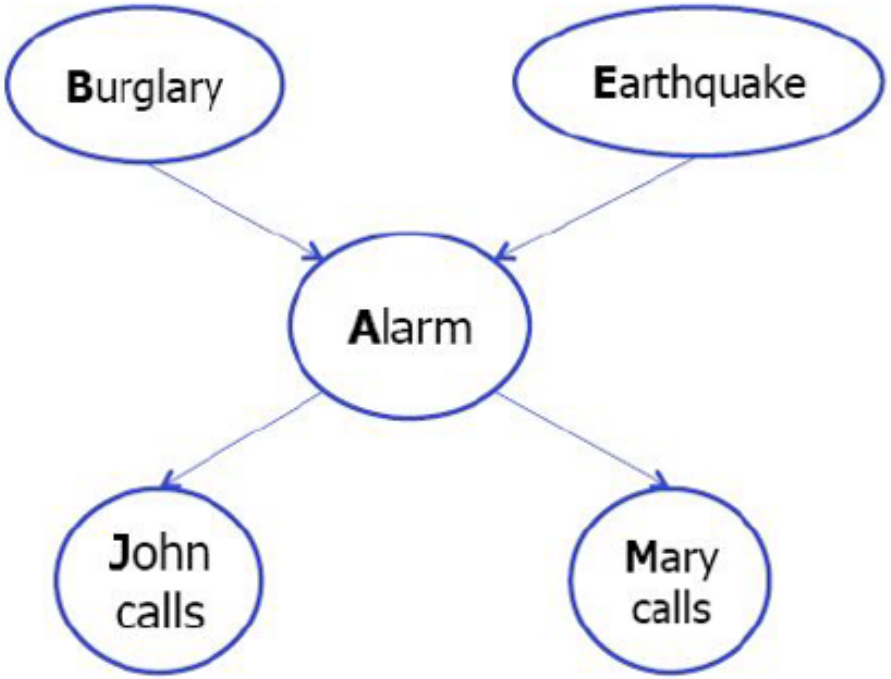
$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \mathbf{pa}(X_i))$$

其中，$\mathbf{pa}(X_i)$表示节点X_i的父节点。



根据贝叶斯网络的紧凑的联合分布，上图贝叶斯网络的联合分布为：

$$P(a, b, \ldots, g) = P(a)P(b)P(f)P(c|a, f)P(g|f, b)P(d|c)P(e|d)$$



根据概率分布的乘法公式，其联合分布为：

$$P(B, E, A, J, M) = P(B)P(E|B)P(A|B, E)P(A|B, E, A)P(M|B, E, A, J)$$

根据贝叶斯网络的紧凑的联合分布，其公式为：

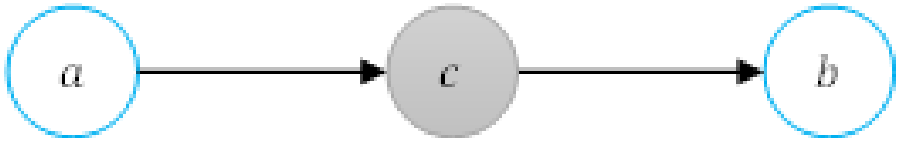
$$P(B, E, A, J, M) = P(B)P(E)P(A|B, E)P(J|A)P(M|A)$$

- 独立性(Independent): 设两个随机变量a和b，如果它们的联合概率分布满足$p(a,b) = p(a)p(b)$, 则称随机变量a和b是相互独立的，记为$a \perp b$.
- 条件独立性(Conditional Independent):
  如果在给定一个额外的随机变量c后，有$p(a,b|c) = p(a|c)p(b|c)$, 此时称a和b是条件独立的, 记为$a \perp b|c$.
- 定理：贝叶斯网络中每一个节点在给定其父节点的条件下与其他非后代节点条件独立, 即：

$$\forall k, X_k \perp NonDesc(X_k)|\mathbf{pa}(X_k)$$

其中，$NonDesc(X_k)$表示除$\mathbf{pa}(X_k)$之外X_k的非后代节点。

- 变量依赖的三种基本结构
  - 顺序结构：头到尾
  - 发散结构：尾到尾
  - 汇总结构：头到头
- 顺序结构：节点c连接了一个箭头的头部和另一个箭头的尾部。顺序结构具有条件独立性：在给定c的条件下，a和b条件独立。

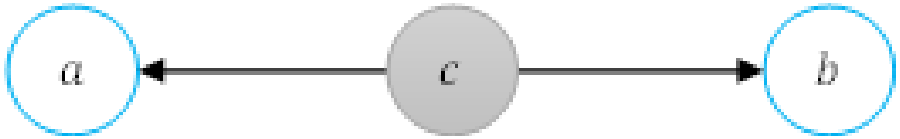


概率图模型的联合分布为：

$$P(a,b,c) = P(a)P(b|a)P(b|c)$$

$$P(a,b|c) = \frac{P(a,b,c)}{P(c)} = \frac{P(a)P(b|a)P(b|c)}{P(c)} = P(a|c)P(b|c)$$

- 发散结构：节点c连接两个箭头的尾部。发散结构具有条件独立性：在给定c的条件下，a和b条件独立。

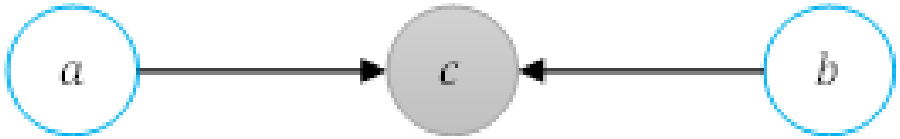


概率图模型的联合分布为:

$$P(a,b,c) = P(c)P(a|c)P(b|c)$$

$$P(a,b|c) = \frac{P(a,b,c)}{P(c)} = \frac{P(c)P(a|c)P(b|c)}{P(c)} = P(a|c)P(b|c)$$

- 汇总结构：节点c连接了两个箭头的头部。汇总结构不具有条件独立性：在给定c的条件下，a和b条件不独立。

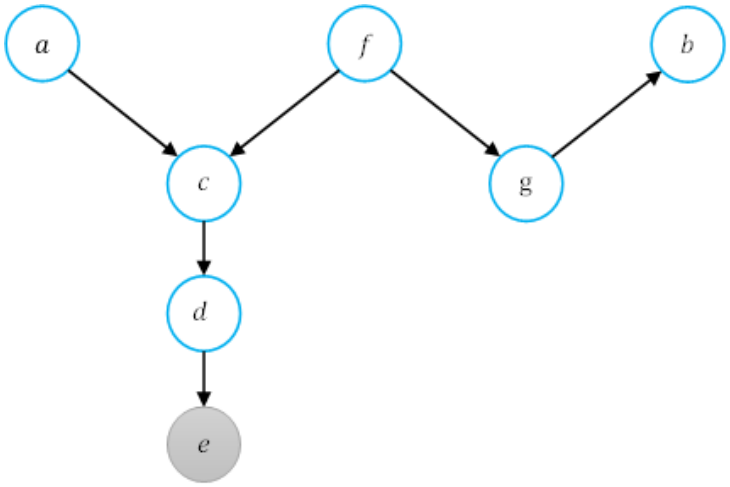


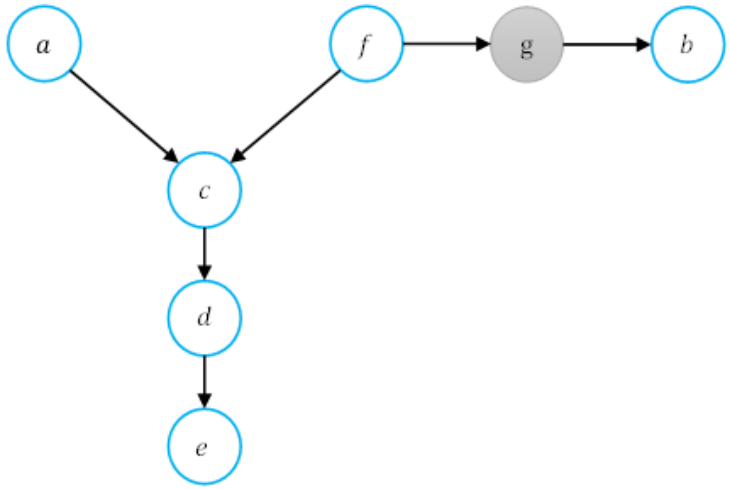
概率图模型的联合分布为:

$$P(a,b,c) = P(a)P(b)P(c|a,b)$$

$$P(a,b) = \int_c P(a,b,c)\mathbf{d}c = \int_c P(a)P(b)P(c|a,b)\mathbf{d}c = P(a)P(b)$$

- d-划分
  设$G = (V, E)$是一个贝叶斯网络（有向图模型），集合A,B,C是V中相互不相交的子集，其中C中的节点是被观测到的。考虑从A中节点到B中节点所有可能的路径， 如果路径上存在一个节点满足如下两个条件之一，那么该条路径是被阻隔的：
  - 路径上的箭头以头到尾或者尾到尾的方式交汇于这个结点，且这个节点在集合C中.
  - 箭头以头到头的方式交汇于这个结点，且这个结点和它的所有后继都不在集合C中
- 如果从A中节点到B中节点的所有路径都被阻隔，就可以说从A到B的路径是被阻隔的，或者叫d-分隔的。
- 对于一个贝叶斯网络，在给定观测的节点集合C的条件下，如果A到B的所有路径都被阻隔，就可以说A和B在给定C的情况下条件独立，图中所有变量的联合概率分布将会满足$A \perp B | C$。


a
f
b
c
g
d
e

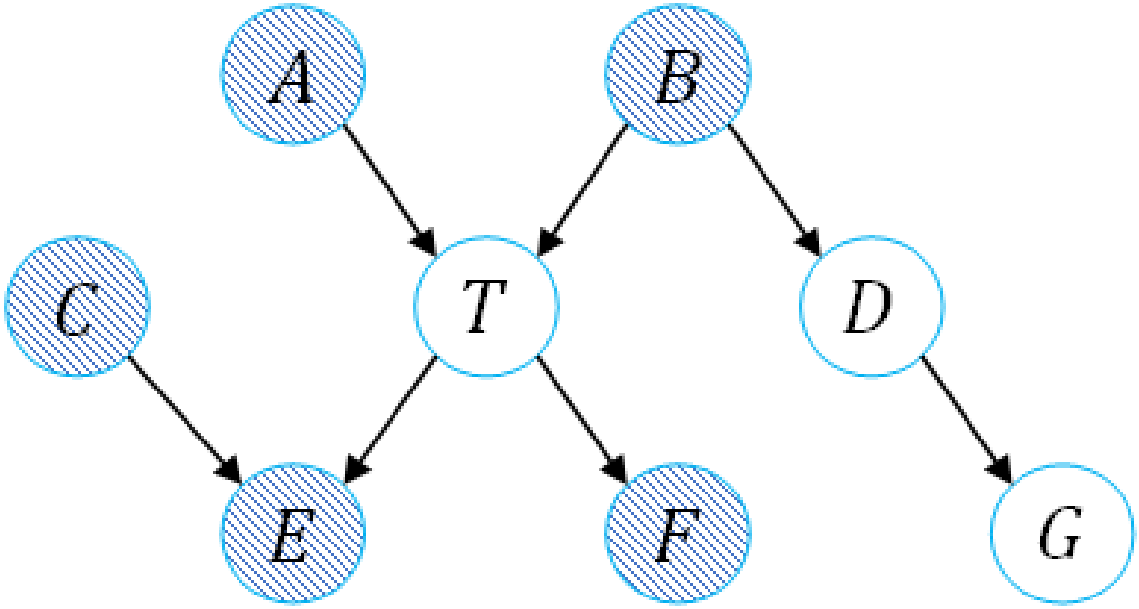

a
f
g
b
c
d
e

马尔科夫毯(Markov blanket)

在所考虑的随机变量的全集U中，对于给定的变量$x \in U$和变量集$MB \subset U(x \notin MB)$，若有

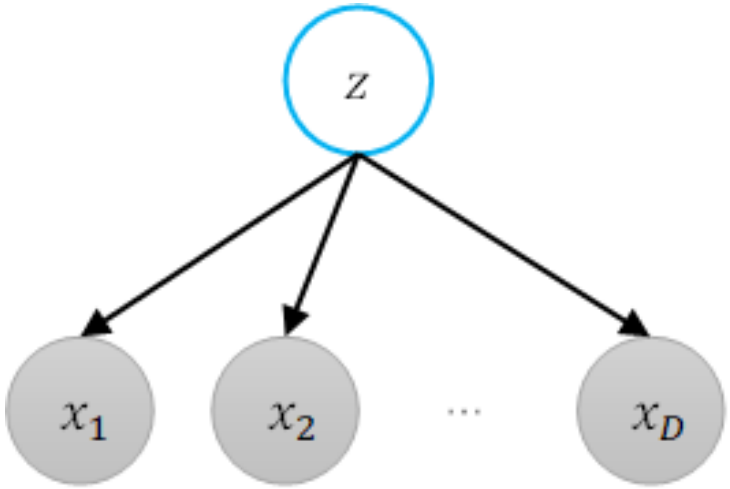
$$x \perp \{U - MB - \{x\}\}|MB$$

则称满足上述条件的最小变量集MB为x的马尔可夫毯。



朴素贝叶斯网络（naïve Bayes）

假设$\mathbf{x}$是一个D维的样本向量，其中每一个元素x_d表示$\mathbf{x}$的一个特征或属性，$z = 1, 2, \cdots, C$表示数据$\mathbf{x}$所属的类别，例如$z = c$表示数据属于第c类。在朴素贝叶斯分类器中，D个特征在已知类别的条件下相互独立。



如上图，其中观测到的特征变量$x_1, x_2, \cdots, x_D$依赖于类别变量z，并且在已知类别z的条件下，特征之间$x_1, x_2, \cdots, x_D$是条件独立的。

由于朴素贝叶斯网络中各个特征之间是条件独立的，因此每一个样本的类条件概率分布有如下表示

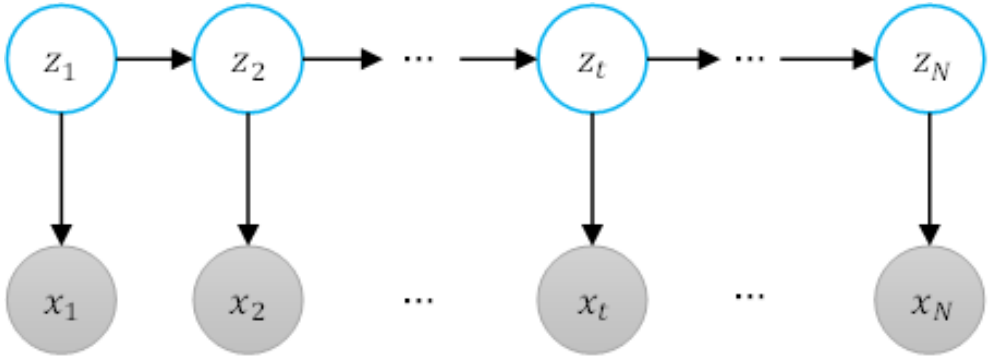
$$P(\mathbf{x}|z=c)=\prod_{d=1}^{D}P(x_d|z=c)$$

已知类条件概率分布，贝叶斯分类器通过计算后验概率$p(z=c|\mathbf{x})$进行决策。根据贝叶斯公式可以得到后验概率为

$$P(z=c|\mathbf{x})=\frac{P(z=c)P(\mathbf{x}|z=c)}{P(\mathbf{x})}=\frac{P(z=c)}{P(\mathbf{x})}\prod_{d=1}^{D}P(x_d|z=c)$$

由于对于所有类别$p(\mathbf{x})$都是相同的，朴素贝叶斯分类器等价于寻找一个c类别，使得$P(\mathbf{x})\prod_{d=1}^{D}P(x_d|z=c)$最大。

隐马尔可夫模型（hidden Markov model）

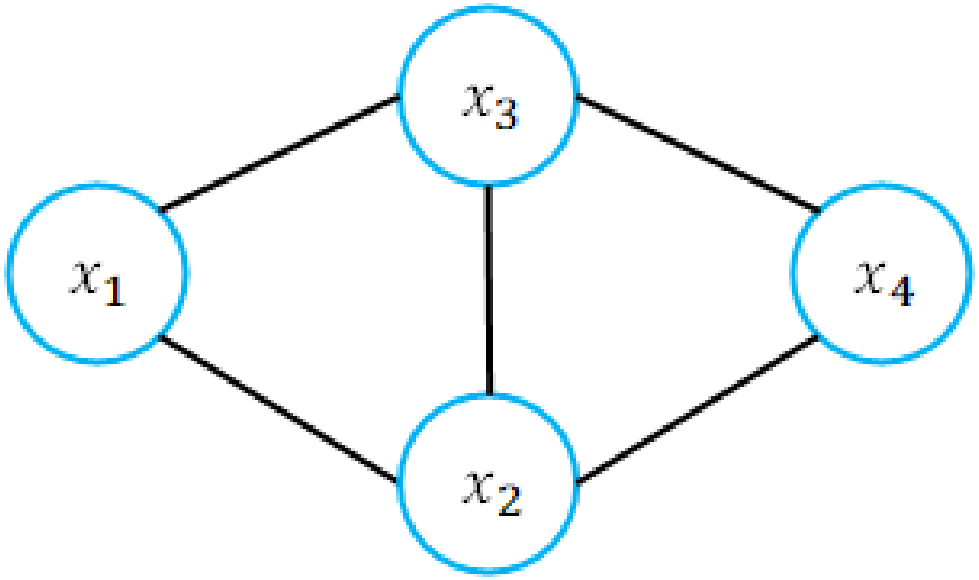


在任意时刻t，给定z_t，观测变量x_t的取值与其他隐变量无关。同时，对于任意时刻的隐变量z_t，如果给定其前一时刻的隐变量z_{t-1}，则与更早时刻的隐变量没有关系。对应的联合概率分布为：

$$P(x_1, z_1, x_2, z_2, \ldots, x_N, z_N) = P(z_1)P(x_1|z_1)\prod_{t=2}^{N} P(z_t|z_{t-1})P(x_t|z_t)$$

无向图模型

- 无向图模型(马尔科夫随机场或马尔科夫网络)，是一类用无向图来描述一组具有局部马尔科夫性质的随机向量X的联合概率分布的模型
  - 节点表示随机变量
  - 无向边表示变量间的依赖关系



无向图的马尔可夫性：任一变量x_k在给定它的邻居的情况下条件独立于所有其他变量, 即

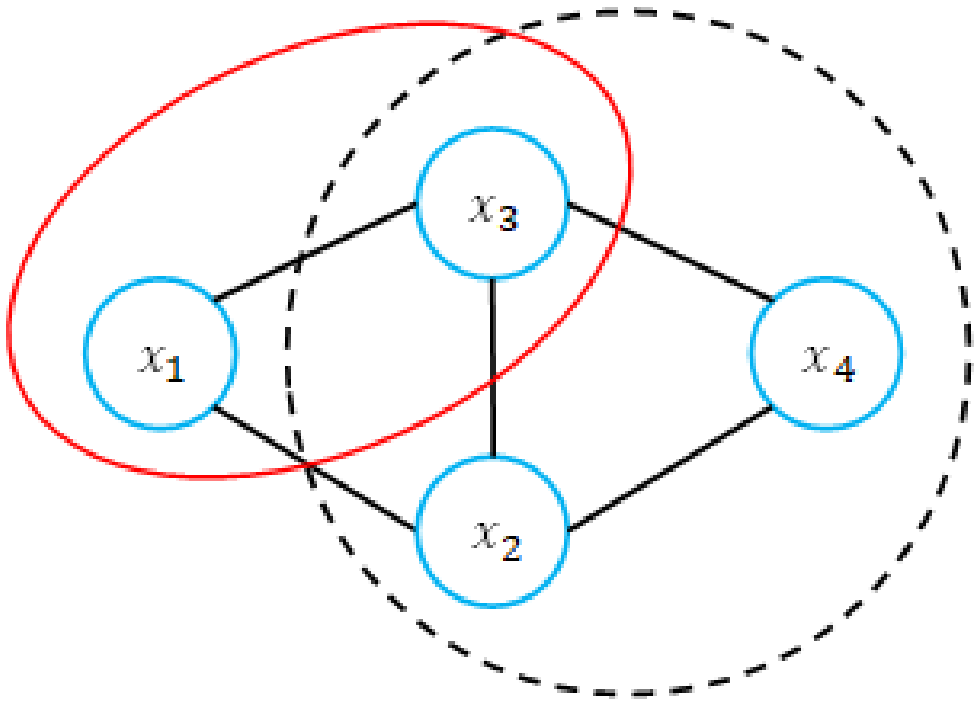
$$x_k \perp z | Ne(x_k)$$

其中z表示除了x_k和它的邻居之外的其他变量。

无向图模型的联合分布一般以全连通子图为单位进行分解。

- 无向图的一个全连通子图，称为团（clique），即团内所有节点之间都有边相连。
- 最大团：在所有团中，如果一个团不能被其它的团包含

下图除单点团外共有7个团，包括$\{x_1, x_2\}$、$\{x_1, x_3\}$、$\{x_2, x_3\}$、$\{x_3, x_4\}$、$\{x_2, x_4\}$、$\{x_1, x_2, x_3\}$和$\{x_2, x_3, x_4\}$.



无向图中的联合概率分布可以分解为一系列定义在最大团上的非负函数的乘积形式

$$P(X) = \frac{1}{Z} \prod_{c \in C} \phi_c(X_c)$$

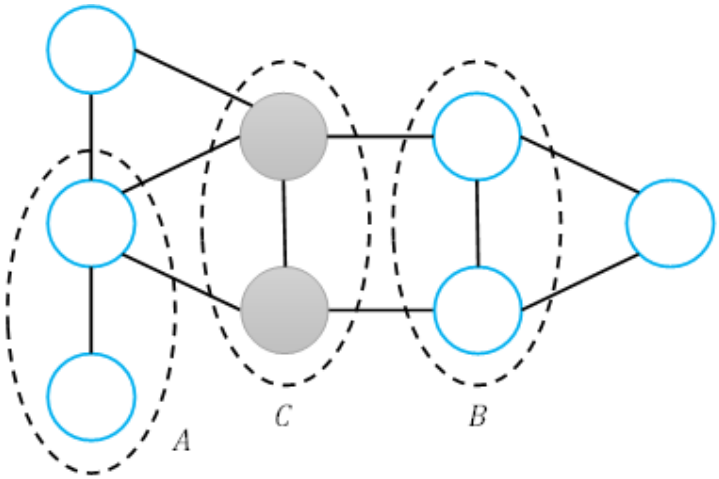
其中C为G中的最大团集合，$\phi_c(X_c) \geq 0$是定义在团c上的势能函数(potential function), Z是配分函数(partition function)，用于将乘积归一化为概率分布形式。

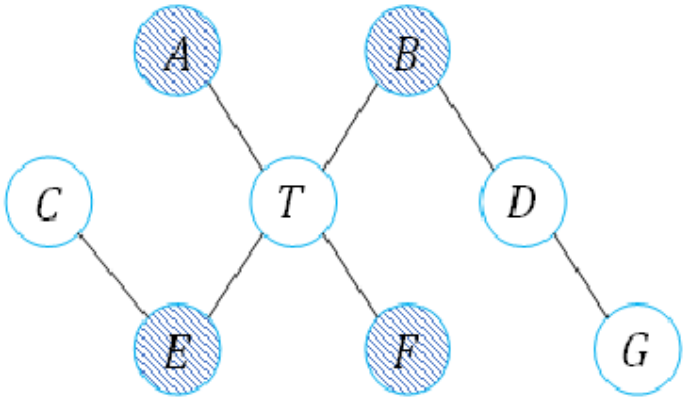
$$Z = \sum_{X \in \mathcal{X}} \prod_{c \in C} \phi_c(X_c)$$

其中$\mathcal{X}$为随机变量的取值空间。

- Hammersley-Clifford定理：如果一个分布$P(X) > 0$满足无向图中的局部马尔可夫性，即，当且仅当$P(X)$可以表示为一系列定义在最大团上的非负函数的乘积形式，即

$$P(X) = \frac{1}{Z} \prod_{c \in C} \phi_c(X_c)$$





对数线性模型

势能函数一般定义为:

$$\phi_c(\mathbf{x}_c|\theta_c) = \exp(\theta_c^T f_c(x_c))$$

其中函数$f_c(\mathbf{x}_c)$为定义在$\mathbf{x}_c$上的特征向量，θ_c为权重向量。这样联合概率分布的对数形式为

$$\log P(\mathbf{x}|\theta) = \sum_{c \in C} \theta_c^T f_c(x_c) - \log Z(\theta)$$

其中θ代表所有势能函数中的参数$\{\theta_c\}$。这种形式的无向图模型也称为对数线性模型或最大熵模型。

如果用对数线性模型来建模条件概率分布$P(y|\mathbf{x})$，那么带有参数的条件概率分布表示$P(y|\mathbf{x},\theta)$为

$$P(y|\mathbf{x}) = \frac{1}{Z(\mathbf{x},\theta)} \exp(\theta^T f(\mathbf{x},y))$$

其中$Z(\mathbf{x},\theta) = \sum_y \exp(\theta^T f(\mathbf{x},y))$

这种对数线性模型也称为条件最大熵模型或softmax回归模型。

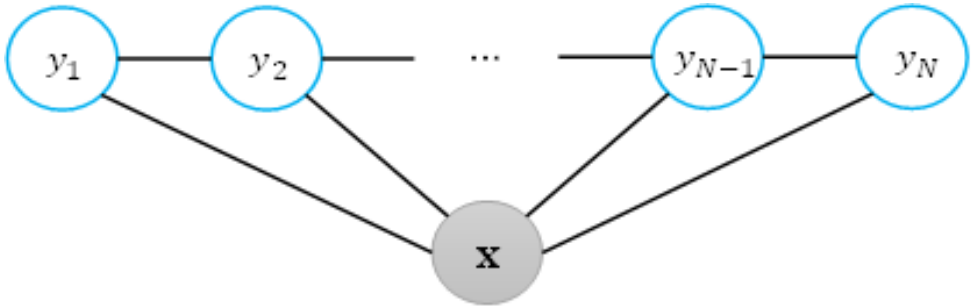
条件随机场

假设条件随机场的最大团集合为C，则其条件概率分布为

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x},\theta)} \exp(\sum_{c\in C} \theta_c^T f_c(\mathbf{x}, y_c))$$

其中$Z(\mathbf{x},\theta) = \sum_{\mathbf{y}} \exp(\theta_c^T f_c(\mathbf{x}, y_c))$为归一化项。

一个常用的条件随机场为下图所示的链式结构:



其条件概率分布为

$$P(\mathbf{y}|\mathbf{x};\theta) = \frac{1}{Z(\mathbf{x},\theta)} \exp(\sum_{t=1}^{T} \theta_1^T f_1(\mathbf{x}, y_t) + \sum_{t=1}^{T-1} \theta_2^T f_2(\mathbf{x}, y_t, y_{t+1}))$$

其中，$f_1(\mathbf{x}, y_t)$为状态特征，一般和位置t相关，$f_2(\mathbf{x}, y_t, y_{t+1})$为转移特征，一般可以简化为$f_2(y_t, y_{t+1})$

- 有向图和无向图的转换
  - 无向图模型可以表示有向图模型无法表示的一些依赖关系，比如循环依赖；但它不能表示有向图模型能够表示的某些关系，比如因果关系。
  - 将有向图转化为无向图，这个过程称为道德化（Moralization）。转换后的无向图称为道德图（Moral Graph）。在道德化的过程中有向图的一些独立性会丢失。

概率推断

- 概率推断: 是指利用模型结构和参数，在观测到部分变量$e = \{e_1, e_2, \cdots, e_m\}$时，计算所关心变量$q = \{q_1, q_2, \cdots, q_n\}$的概率分布或最大概率状态, 其余变量表示为$z$
- 计算部分变量的边缘概率或边缘条件概率
  - $P(q) = \sum_{e,z} P(q, e, z)$
- 计算部分变量后验概率最大的状态组合
  - $P(q|e) = \frac{P(q,e)}{P(e)} = \frac{\sum_z P(q,e,z)}{\sum_{q,z} P(q,e,z)}$

图模型的推断问题可以转换为求任意一个变量子集的边际概率分布问题

- 精确推理（基于图结构、条件独立）
  - 变量消除(Variable Elimination)
  - 置信度传播算法(Belief Propagation)
- 近似推理
  - 抽样算法（基于概率统计）
  - 变分法（基于数学规划）

模型学习

- 网络结构学习：领域专家
- 网络参数估计

不含隐变量的参数估计(最大似然估计)

- 有向图模型：所有变量x的联合概率分布可以分解为每个随机变量x_k的局部条件概率$P(x_k|\mathbf{pa}(x_k);\theta_k)$的连乘形式，其中$\theta_k$为第$k$个变量的局部条件概率的参数。

给定N个训练样本，其对数似然函数为:

$$L(D;\theta) = \frac{1}{N}\prod_{n=1}^{N}\prod_{k=1}^{K} P(x_k^n \mid \mathbf{pa}(x_k^n);\theta_k)$$

其中θ为模型中的所有参数。
因为所有变量都是可观测的，最大化对数似然$L(D;\theta)$，只需要分别地最大化每个变量的条件似然来估计其参数

$$\theta_k = \arg\max_{\theta} \sum_{n=1}^{N} \log P(x_k^n \mid \mathbf{pa}(x_k^n);\theta_k)$$

- 无向图模型：在无向图模型中，所有变量的联合概率分布可以分解为定义在最大团上的势能函数的连乘形式。以对数线性模型为例：

$$P(\mathbf{x};\theta) = \frac{1}{Z(\theta)} \exp\left(\sum_{c\in C} \exp(\theta_c^T f_c(x_c))\right.$$

其中$Z(\theta) = \sum_{\mathbf{x}} \exp\left(\sum_{c\in C} \exp(\theta_c^T f_c(x_c))\right.$

给定N个训练样本，其对数似然函数为:

$$\begin{aligned} L(D;\theta) &= \frac{1}{N}\sum_{n=1}^{N} \log P(\mathbf{x}_n;\theta) \\ &= \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{c\in C} \theta_c^T f_c(x_c^n)\right) - \log Z(\theta) \end{aligned}$$

其中θ_c 为定义在团c上的势能函数的参数。

无向图的参数估计通常采用近似的方法。

- 利用采样来近似计算这个期望；
- 坐标上升法，即固定其它参数，来优化一个势能函数的参数。

含隐变量的参数估计

- 如果图模型中包含隐变量，即有部分变量是不可观测的，就需要用EM算法进行参数估计。