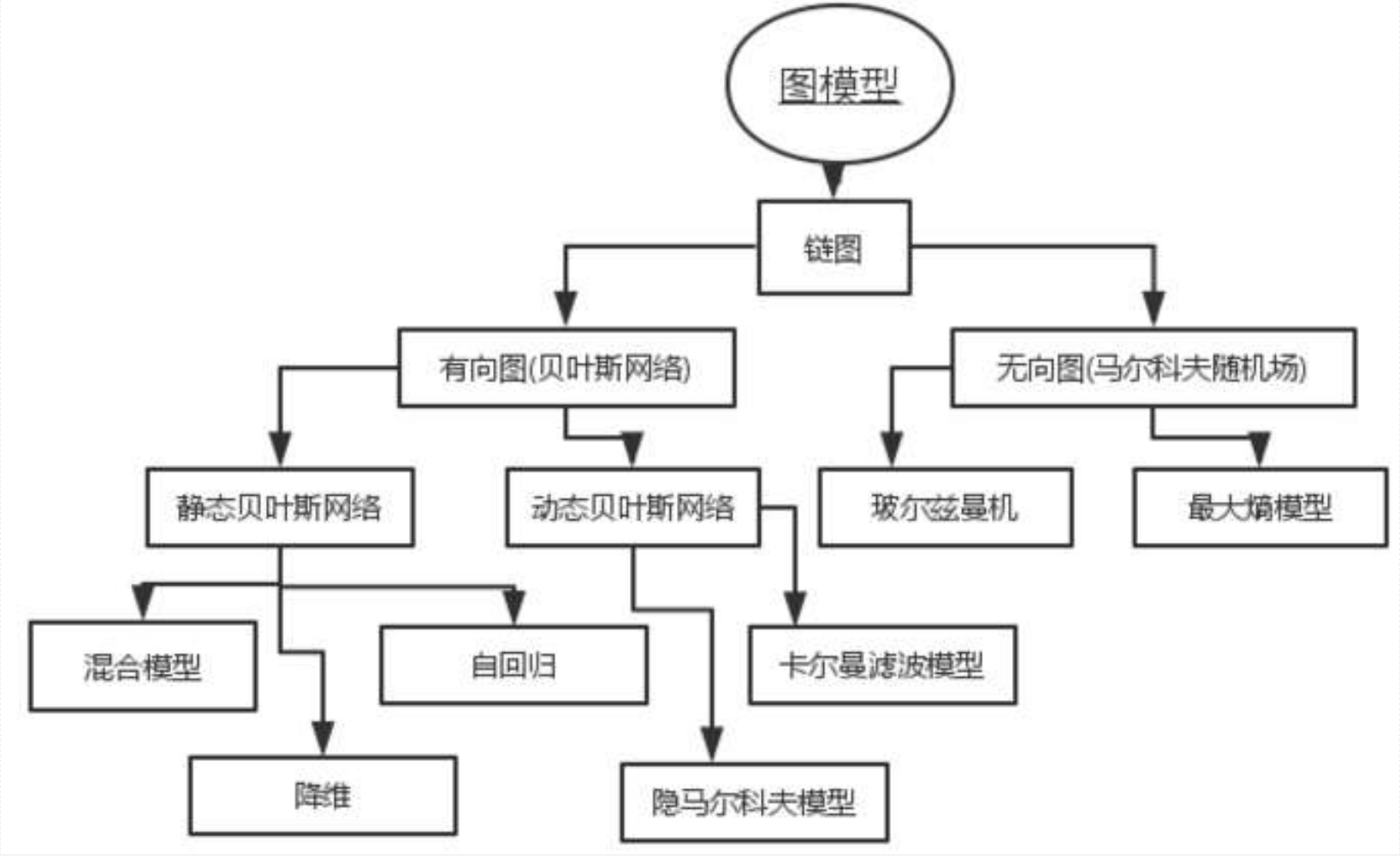
概率图模型

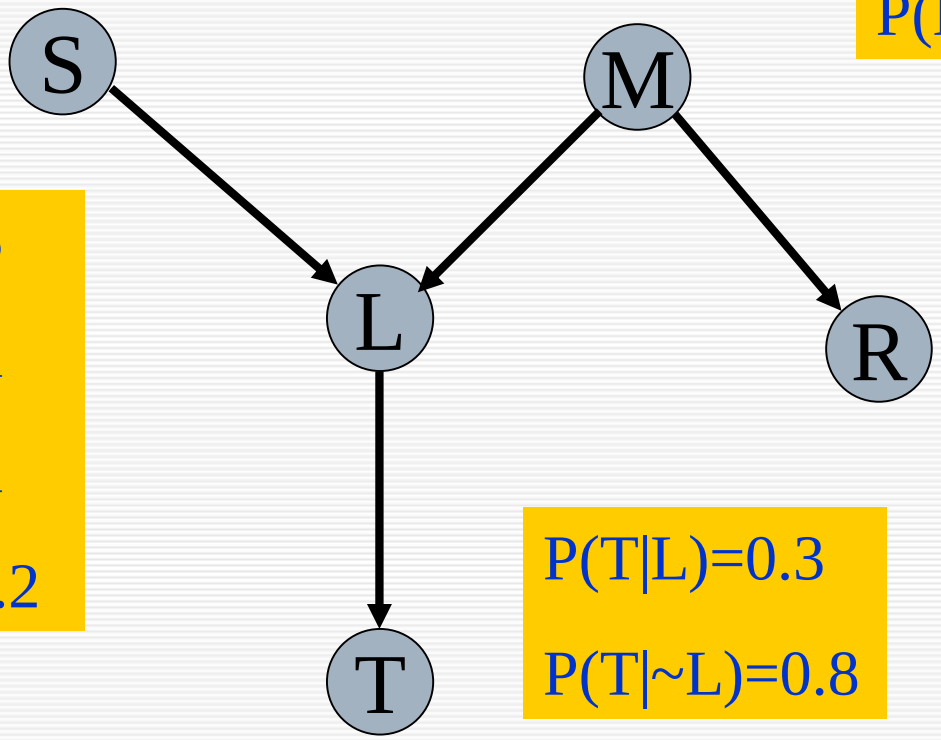
信息学院-黄浩

2022年3月14日星期一

概述


图模型
链图
有向图(贝叶斯网络)
无向图(马尔科夫随机场)
静态贝叶斯网络
动态贝叶斯网络
玻尔兹曼机
最大熵模型
混合模型
自回归
卡尔曼滤波模型
降维
隐马尔科夫模型

贝叶斯网络概述



P(S)=0.3

P(M)=0.6

P(L|M^S)=0.05

P(L|M^~S)=0.1

P(L|~M^S)=0.1

P(L|~M^~S)=0.2

P(R|M)=0.3

P(R|~M)=0.6

P(T|L)=0.3

P(T|~L)=0.8

T: The lecture started by 9:00

L: The lecturer arrives late

R: The lecture concerns robots

M: The lecturer is Manuel

S: It is sunny

贝叶斯网络概述

- 贝叶斯网络
  - 用来表示变量间连接概率的图形模式，它提供了一种自然的表示因果信息的方法，用来发现数据间的潜在关系。在这个网络中，用节点表示变量，有向边表示变量间的依赖关系。同时每个节点都对应着一个条件概率分布表(CPT)，指明了该节点与父节点之间概率依赖的数量关系。

贝叶斯网络概述

- 贝叶斯网络
  - 贝叶斯网络(Bayesian network)，又称信念网络(Belief Network)，或有向无环图模型(directed acyclic graphical model)，是一种概率图模型，于1985年由Judea Pearl首先提出。它是一种模拟人类推理过程中因果关系的不确定性处理模型，其网络拓朴结构是一个有向无环图(DAG)

贝叶斯网络概述

- 贝叶斯网络
- 贝叶斯网络的有向无环图中的节点表示随机变量，它们可以是可观察到的变量，或隐变量、未知参数等。认为有因果关系（或非条件独立）的变量或命题则用箭头来连接。若两个节点间以一个单箭头连接在一起，表示其中一个节点是“因(parents)”，另一个是“果(children)”，两节点就会产生一个条件概率值。

贝叶斯网络概述

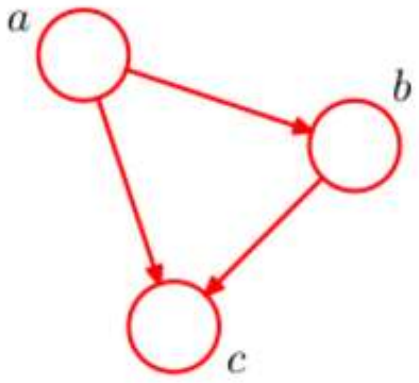
给出$p(a,b,c)$的联合分布形式

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

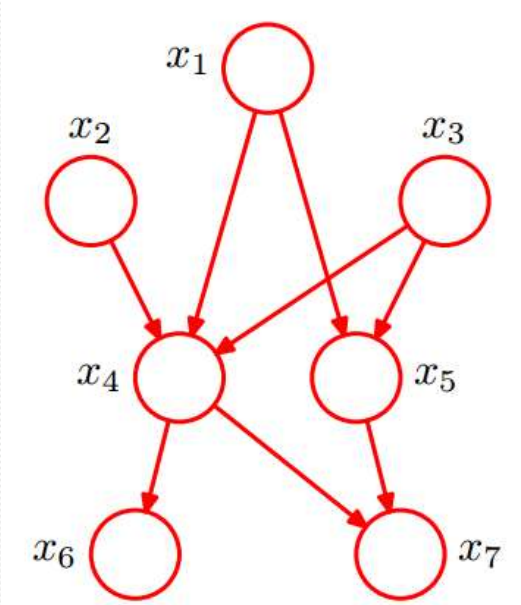
继续

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

得到下图基本表示



贝叶斯网络概述



分布为

$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$$

所以可以用图(无环，即DAG(directed acyclic graph))表示任意的概率分布。

贝叶斯网络概述

$$p(\mathbf{t}, \boldsymbol{w}) = p(\boldsymbol{w}) \prod_{n=1}^{N} p(t_n \mid \boldsymbol{w}) \tag{8.6}$$

图模型表示的联合概率分布如图8.3所示。

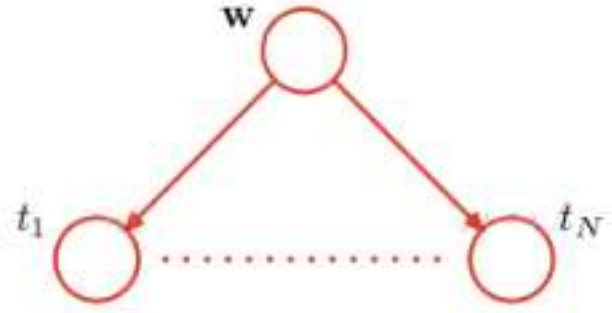


图 8.3: 有向图模型表示联合概率分布 (8.6)

贝叶斯网络概述

复杂模型的板(plate)表示

当我们开始处理更加复杂的模型时,我们会看到,像图8.3那样显式地写出 t 1 , . . . , t N 的结点是很不方便的。于是,我们引入一种图结构,使得多个结点可以更简洁地表示出来。这种图结构中,我们画出一个单一表示的结点 t n ,然后用一个被称为板(plate)的方框圈起来,标记为 N ,表示有 N 个同类型的点。用这种方式重新表示图8.3,我们得到了图8.4所示的图。

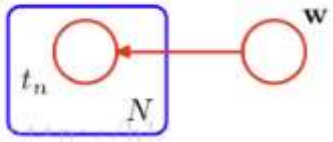


图 8.4: 一种更加简洁的方式表示图8.3中的图，其中我们引入了一个板（plate）（标记为N的方框）来表示N个结点，这些结点中，只有一个例子t_n被显式地画出。

贝叶斯网络概述

显式表示参数和随机变量

我们有时会发现,显式地写出模型的参数和随机变量是很有帮助的。此时,公式(8.6)就变成了

$$p(\mathbf{t}, \boldsymbol{w} \mid \mathbf{x}, \alpha, \sigma^2) = p(\boldsymbol{w} \mid \alpha) \prod_{n=1}^{N} p(t_n \mid \boldsymbol{w}, x_n, \sigma^2)$$

对应地,我们可以在图表示中显式地写出 x 和 α 。为了这样做,我们会遵循下面的惯例:随机变量由空心圆表示,确定性参数由小的实心圆表示。如果我们让图8.4包含确定性参数,我们就得到了图8.5。

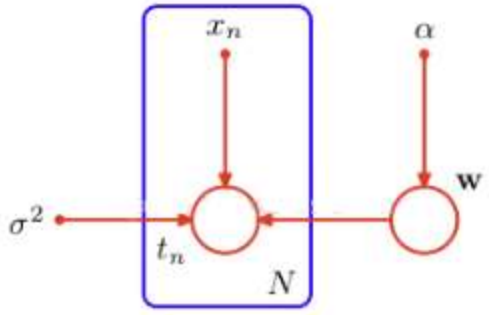


图 8.5: 本图给出了与图8.4相同的模型，但是显式地画出了确定性参数，用小的实心圆表示。

贝叶斯网络概述

观测变量和潜在变量

当我们将图模型应用于机器学习或者模式识别的问题中时,我们通常将某些随机变量设置为具体的值,例如将变量{t n } 根据多项式曲线拟合中的训练集进行设置。在图模型中,我们通过给对应的结点加上阴影的方式来表示这种观测变量(observed variables)。于是,图8.5所示的图中,如果 {t n } 是观测变量,那么就变成了图8.6。注意, w 不是观测变量,因此 w 是潜在变量(latent variable)的一个例子。潜在变量也被称为隐含变量(hidden variable)。

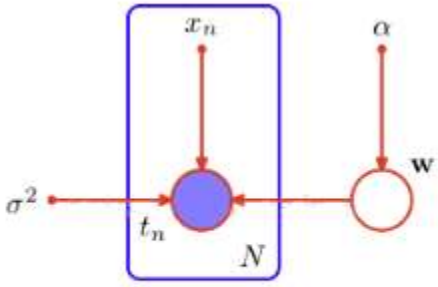


图 8.6: 与图8.5相同，但是结点$\{t_n\}$被标记为阴影，表示对应的随机变量被设置成它们在训练集里的观测值。

贝叶斯网络概述

系数 w 的的后验概率

观测到了 {t n } 的值,如果必要的话,我们可以计算系数 w 的的后验概率。现阶段,我们注意到,这是贝叶斯定理的一个直接应用。

$$p(\boldsymbol{w} \mid \mathbf{t}) \propto p(\boldsymbol{w}) \prod_{n=1}^{N} p(t_n \mid \boldsymbol{w}) \tag{8.7}$$

观测数据为条件的t 的概率分布

其中,我们再一次省略了确定性参数,使得记号简洁。

通常,我们对于 w 这样的参数本身不感兴趣,因为我们的最终目标是对输入变量进行预测。假设给定一个输入值 x ,我们想找到以观测数据为条件的对应的 t 的概率分布。描述这个问题的图模型如图8.7所示。以确定性参数（x）为条件,这个模型的所有随机变量（t, w）的联合分布为

$$p(\widehat{t}, \mathbf{t}, \boldsymbol{w} \mid \widehat{x}, \mathbf{x}, \alpha, \sigma^2) = \left[\prod_{n=1}^{N} p(t_n \mid x_n, \boldsymbol{w}, \sigma^2)\right] p(\boldsymbol{w} \mid \alpha) p(\widehat{t} \mid \widehat{x}, \boldsymbol{w}, \sigma^2) \tag{8.8}$$

然后,根据概率的加和规则,对模型参数 w 积分,即可得到 t 的预测分布

$$p(\widehat{t} \mid \widehat{x}, \mathbf{x}, \mathbf{t}, \alpha, \sigma^2) \propto \int p(\widehat{t}, \mathbf{t}, \boldsymbol{w} \mid \widehat{x}, \mathbf{x}, \alpha, \sigma^2)\, \mathrm{d}\boldsymbol{w}$$

其中我们隐式地将 t 中的随机变量设置为数据集中观测到的具体值。

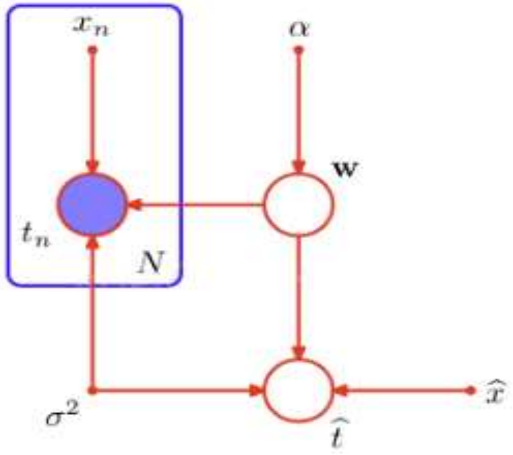


图 8.7: 多项式回归模型，对应于图8.6。同时画出了一个新的输入值$\widehat{x}$以及对应的模型精度$\widehat{t}$。

贝叶斯网络概述

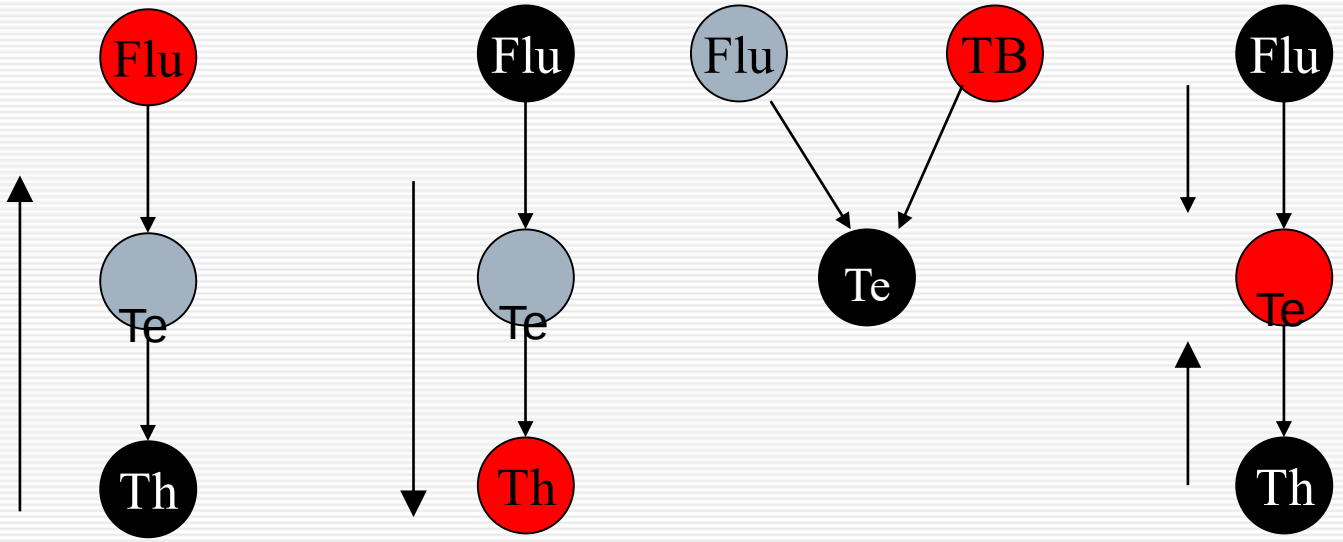
- 贝叶斯网络的特点：
  - 双向推理能力（诊断、因果和混合推理）
  - 快速的调试和重构能力
  - 具有较强的概率统计基础
  - 用于人工智能和专家系统的不确定推理（优于早期的基于规则的模式）。

Flu: 感冒

Te: 体温

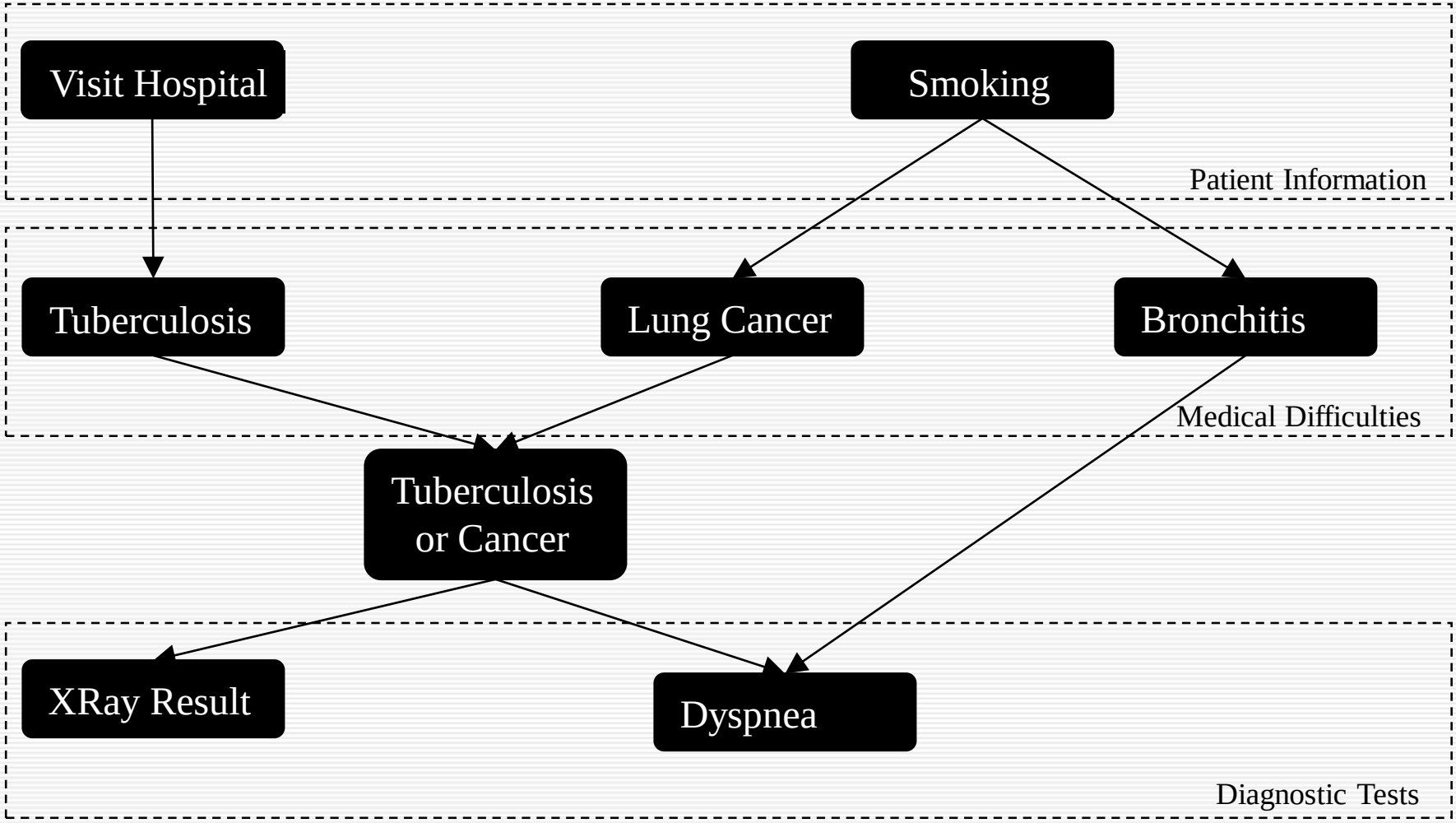
TB: 肺结核

Th: 温度计度数



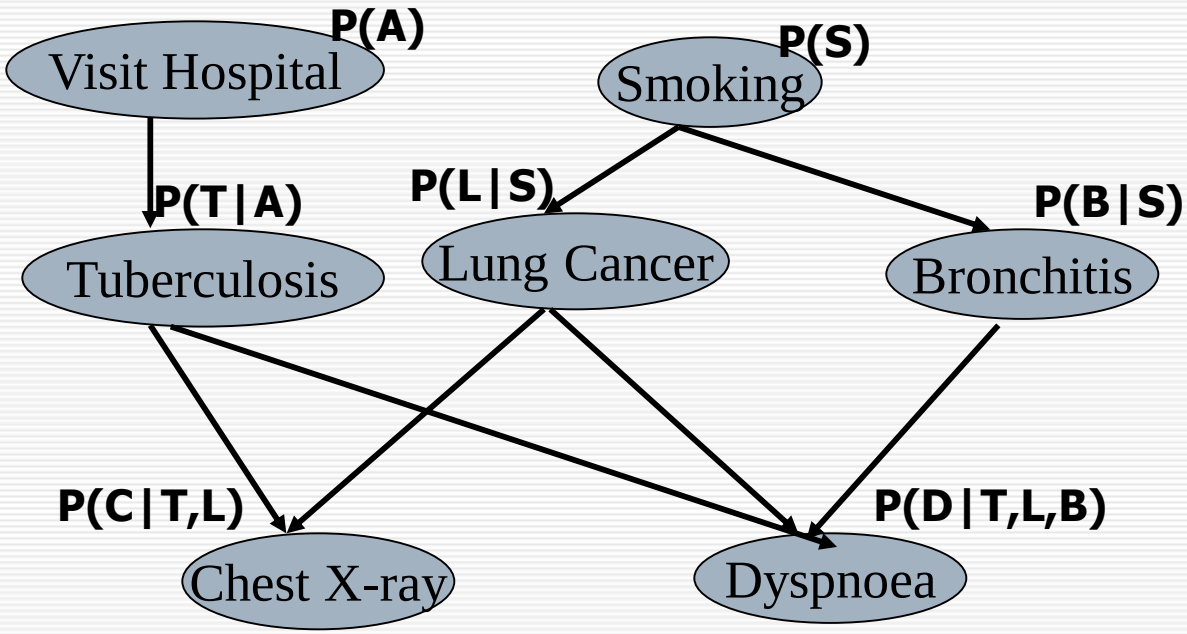
Diagnostic inference　*Causal inference*　*Intercausal inference*　*Mixed inference*

一个医疗诊断的例子


Visit Hospital
Smoking
Patient Information
Tuberculosis
Lung Cancer
Bronchitis
Medical Difficulties
Tuberculosis or Cancer
XRay Result
Dyspnea
Diagnostic Tests

例子说明

贝叶斯网络是表示变量间概率依赖关系的有向无环图



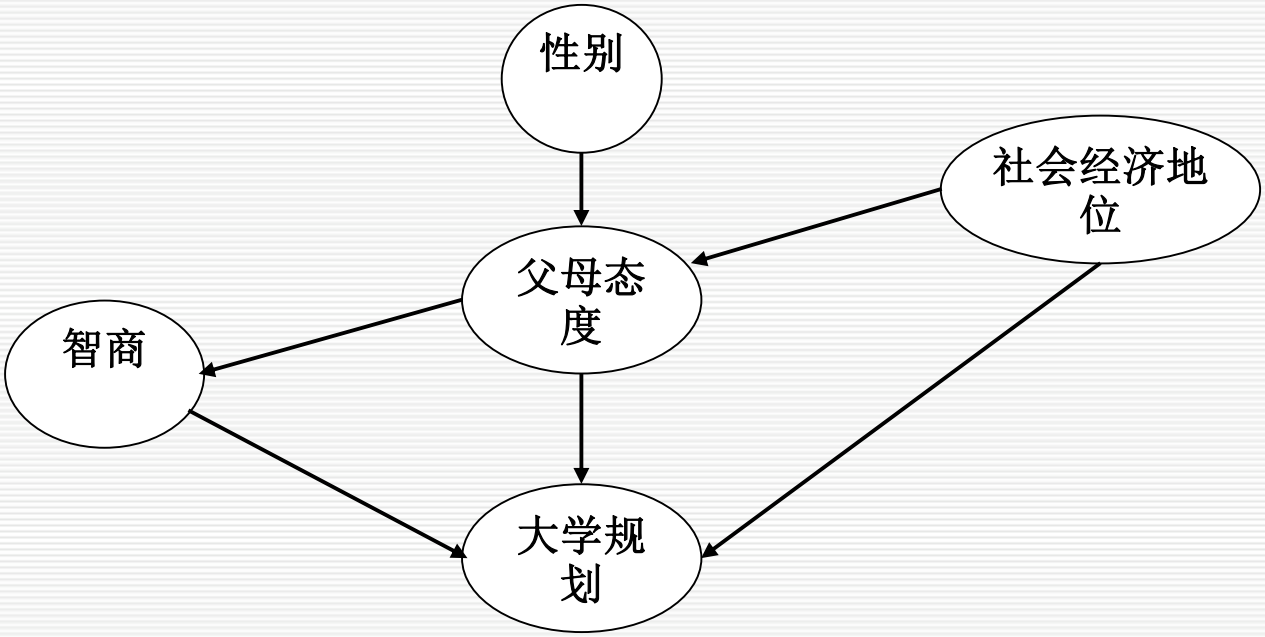
CPT:

T	L	B	D=0	D=1
0	0	0	0.1	0.9
0	0	1	0.7	0.3
0	1	0	0.8	0.2
0	1	1	0.9	0.1
...				

$$P(A, S, T, L, B, C, D) = P(A)\,P(S)\,P(T|A)\,P(L|S)\,P(B|S)\,P(C|T,L)\,P(D|T,L,B)$$

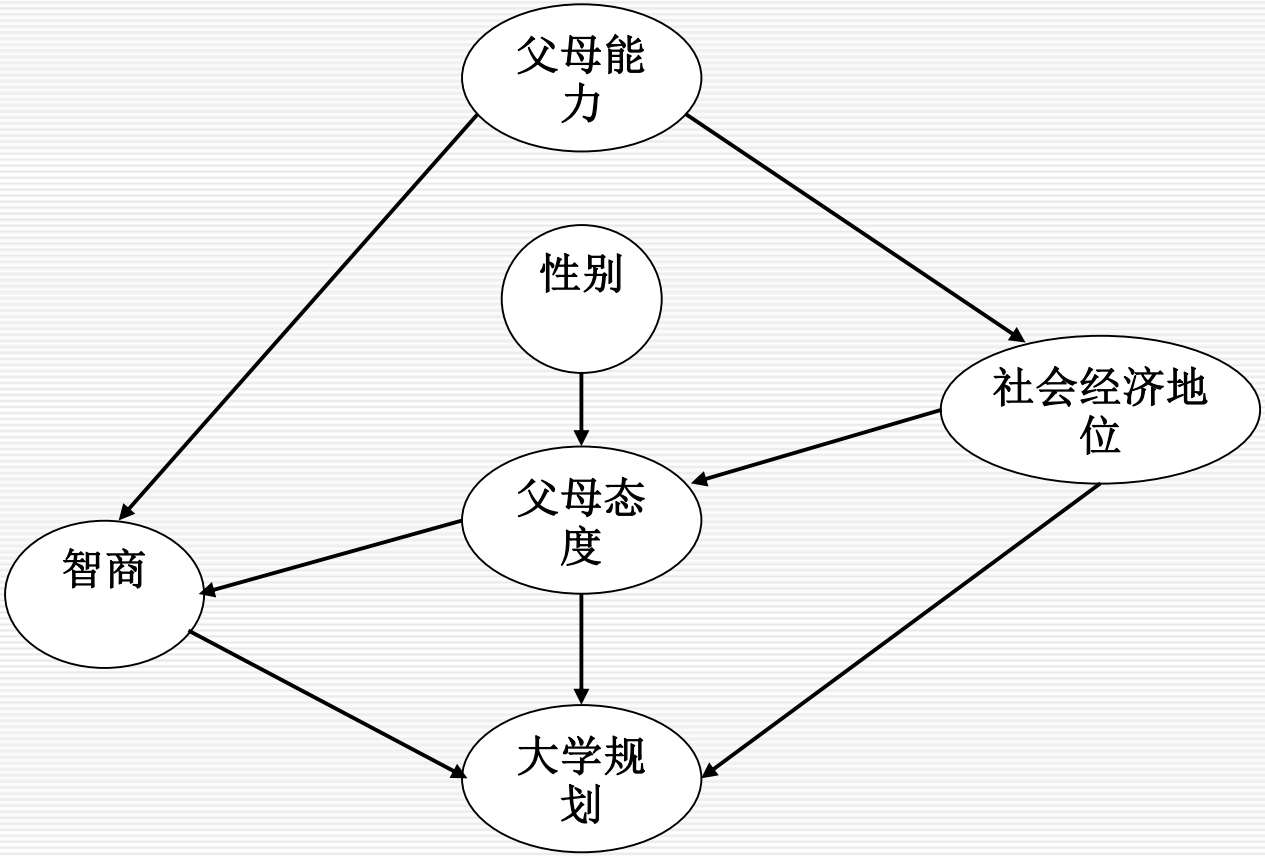
条件独立性假设 ➡ 有效的表示

例子说明



1968年对美国威斯康星州高中生的一项调查

例子说明



1968年对美国威斯康星州高中生的一项调查

贝叶斯网络学习

- 贝叶斯网络的构造：
  - 确定为建立网络模型有关的变量及其解释
  - 建立一个表示条件独立断言的有向无环图

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i \mid \mathbf{pa}_i)$$

  - 指派局部概率分布 $p(x_i \mid \mathbf{pa}_i)$

贝叶斯网络学习

- 学习情况的分类：
- 一、完整数据的学习
  - 参数学习
  - 结构学习
- 二、不完整数据的学习
  - 参数学习
  - 结构学习

贝叶斯网络学习

- **完整数据集的参数学习：**
- 在网络结构已知的情况下，参数学习的目标是根据参数的先验分布和数据样本计算参数的后验分布

$$P(\theta_m \mid D,m)=\frac{P(\theta_m \mid m)P(D \mid \theta_m,m)}{P(D \mid m)}=\frac{P(\theta_m \mid m)P(D \mid \theta_m,m)}{\int P(\theta_m \mid m)P(D \mid \theta_m,m)d\theta_m}$$

D为数据集　　m为网络结构模型　　θ_m 是参数向量

贝叶斯网络学习

- **完整数据集的结构学习**
- 基于打分函数和网络结构搜索的方法
- 基于打分函数和搜索的贝叶斯网络的学习问题可以描述如下：给定一个训练数据集，找到一个与数据集匹配程度最好的网络 。通常的做法是引入一个打分函数，用它给每一个网络结构和数据集的匹配程度打分，选择适宜的搜索算法搜索分值最高的网络模型。而常用的打分函数有两种：基于贝叶斯统计学的BDe打分函数和基于编码理论的MDL打分函数
- 优点
  - 具有较好的数学解释
  - 能在精度和网络复杂度之间折中处理
  - 同时考虑结构和局部条件概率
- 缺点
  - 计算强度大

贝叶斯网络学习

- **完整数据集的结构学习**
  - 基于相关性分析的学习方法
- 基于相关性分析的网络结构学习算法可以分为三个阶段：第一阶段通过计算每一对节点的互信息来测量节点间的相关程度，并以此来构造一个初始网络；第二阶段通过计算条件互信息来决定两节点是否条件独立，如不是则添加相应的边；第三阶段检查当前网络中的每一条边，如果暂时移开某一条边后，它连接的两个节点条件独立，则永久删除；否则保留
  - 优点
    - 在直觉上符合贝叶斯网络的构成
    - 独立测试的形式与结构学习无关
  - 缺点
    - 对单个的独立测试的错误敏感
    - 计算强度大

贝叶斯网络学习

- **不完整数据集的参数学习：**
- 精确处理不完整数据中的网络参数学习问题是难于实现的。因此，目前主要采用一些近似方法
  - Monte-Carlo方法
  - Gaussian 近似和Laplace近似
  - 贝叶斯信息标准(BIC)
  - EM（expectation-maximization）算法

贝叶斯网络学习

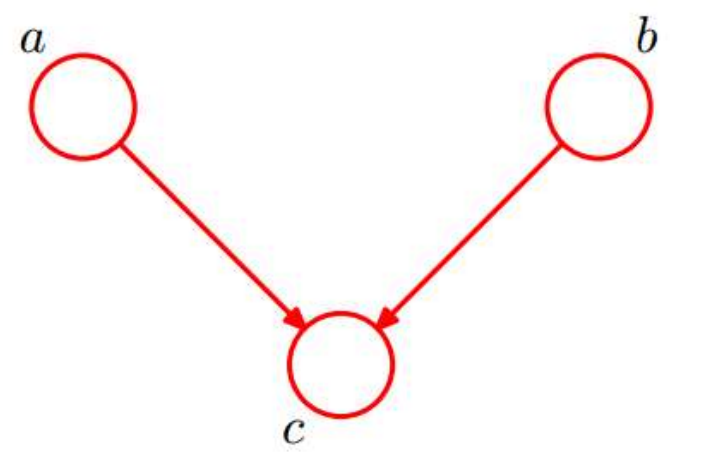
- **不完整数据集的结构学习**
- 从不完整数据中学习网络结构比从完整数据中学习要困难得多，因为事件发生的次数无法统计出来，所以评分函数无法分解成只与局部结构相关的因式，这样就使得：
  - 必须执行推理过程计算待评判的网络结构的分值
  - 为了给网络结构配置最佳参数，必须利用EM算法或基于梯度的方法执行非线性的优化过程
  - 搜索算法对网络结构任一点的局部改动，都将影响网络其它局部结构的评估，因而为评判当前网络结构必须评估其所有的“邻居”
- 常用方法是吉布斯采样、MCMC

贝叶斯网络推理

- 网络结构一　　D-分离

$$\sum_c P(a,b,c) = \sum_c P(a)*P(b)*P(c|a,b)$$

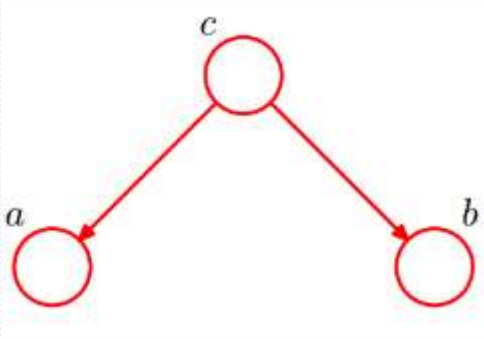
$$\Rightarrow P(a,b) = P(a)*P(b)$$

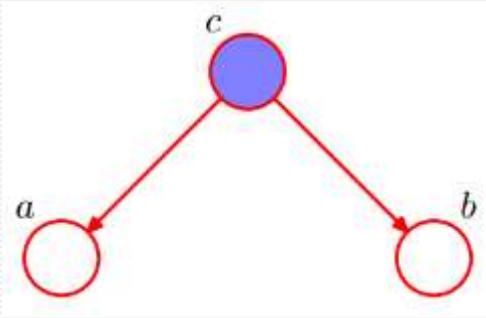


即在**c未知的条件下，a、b被阻断(blocked)，是独立的**，称之为head-to-head条件独立。

贝叶斯网络推理

- 网络结构二



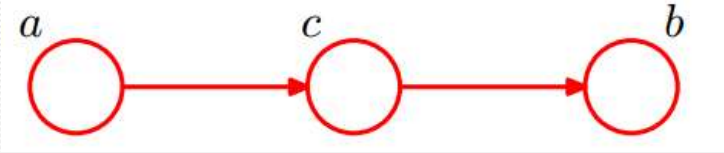


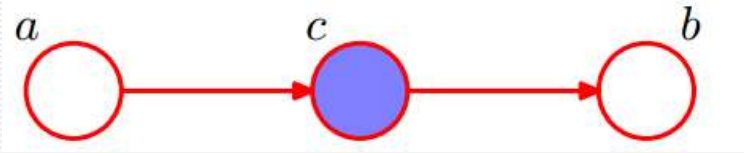
考虑c未知，跟c已知这两种情况：
在c未知的时候，有：P(a,b,c)=P(c)*P(a|c)*P(b|c)，此时，没法得出P(a,b) = P(a)P(b)，即c未知时，a、b不独立。
在c已知的时候，有：P(a,b|c)=P(a,b,c)/P(c)，然后将P(a,b,c)=P(c)*P(a|c)*P(b|c)带入式子中，得到：P(a,b|c)=P(a,b,c)/P(c) = P(c)*P(a|c)*P(b|c) / P(c) = P(a|c)*P(b|c)，即c已知时，a、b独立。

在c给定的条件下，a，b被阻断(blocked)，是独立的，称之为tail-to-tail条件独立

贝叶斯网络推理

- 网络结构三





$$
\begin{aligned}
& P(a, b \mid c) \\
= \; & P(a, b, c) / P(c) \\
= \; & P(a) * P(c \mid a) * P(b \mid c) \; / \; P(c) \\
= \; & P(a, c) * \; P(b \mid c) \; / \; P(c) \\
= \; & P(a \mid c) \; * \; P(b \mid c)
\end{aligned}
$$

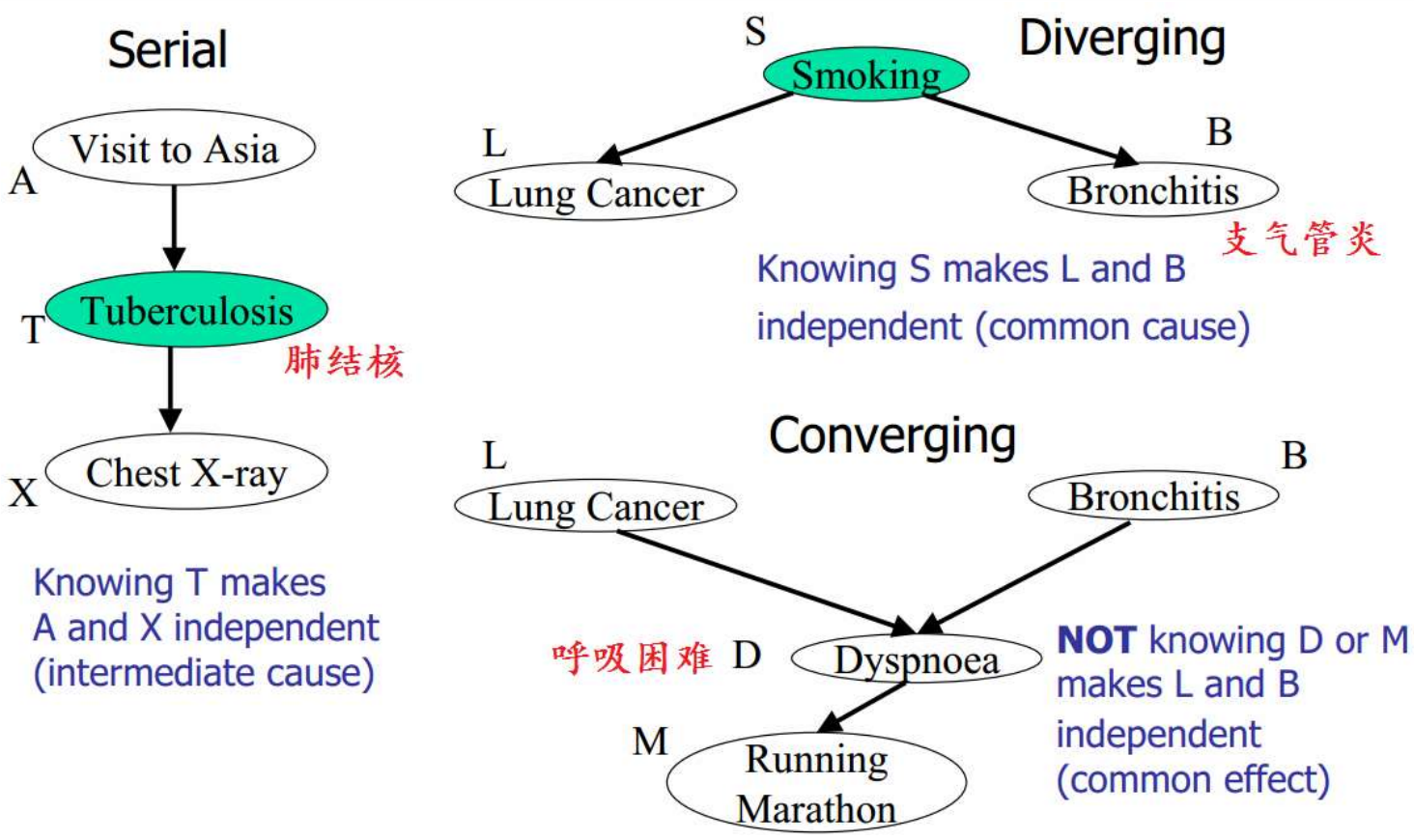
在c给定的条件下，a，b被阻断(blocked)，是独立的，称之为head-to-tail条件独立。

贝叶斯网络推理

将上述结点推广到结点集，则是：对于任意的结点集A，B，C，考察所有通过A中任意结点到B中任意结点的路径，若要求A，B条件独立，则需要所有的路径都被阻断(blocked)，即满足下列两个前提之一：

A和B的“head-to-tail型”和“tail-to-tail型”路径都通过C；
A和B的“head-to-head型”路径不通过C以及C的子孙；

贝叶斯网络推理

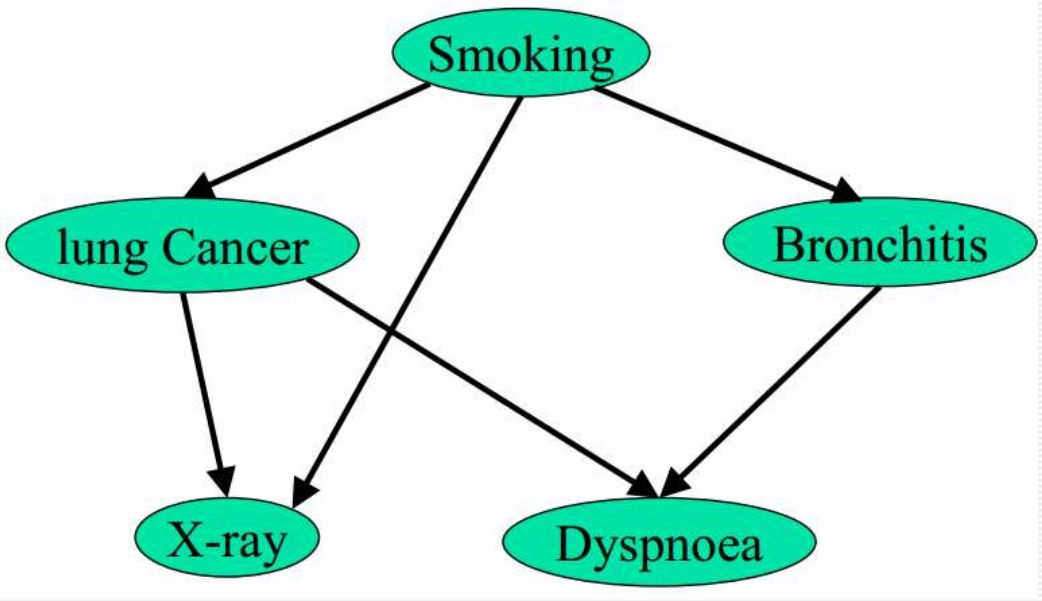


举例说明上述D-Separation的3种情况（即贝叶斯网络的3种结构形式），则是如图所示：

贝叶斯网络推理

上图中左边部分是head-to-tail，给定 T 时，A 和 X 独立；右边部分的右上角是tail-to-tail，给定S时，L和B独立；右边部分的右下角是head-to-head，未给定D时，L和B独立。

贝叶斯网络推理



对于上图，在一个人已经呼吸困难（dyspnoea）的情况下，其抽烟（smoking）的概率是多少呢？即：

$$P(\text{smoking} \mid \text{dyspnoea}=\text{yes}) = ?$$

贝叶斯网络推理

$$P(s|d=1) = \frac{P(s,d=1)}{P(d=1)} \propto P(s,d=1) =$$

$$\sum_{d=1,b,x,c} P(s)P(c|s)P(b|s)P(x|c,s)P(d|c,b) =$$

$$P(s)\sum_{d=1}\sum_{b}P(b|s)\sum_{x}\sum_{c}P(c|s)P(x|c,s)P(d|c,b)$$

Variable Elimination

$$f(s,d,b,x)$$

贝叶斯网络推理

解释下上述式子推导过程：

第二行：对联合概率关于b,x,c求和（在d=1的条件下），从而消去b,x,c，得到s和d=1的联合概率。
第三行：最开始，所有变量都在sigma(d=1,b,x,c)的后面（sigma表示对“求和”的称谓），但由于P(s)和“d=1,b,x,c”都没关系，所以，可以提到式子的最前面。而且P(b|s)和x、c没关系，所以，也可以把它提出来，放到sigma(b)的后面，从而式子的右边剩下sigma(x)和sigma(c)。

贝叶斯网络推理

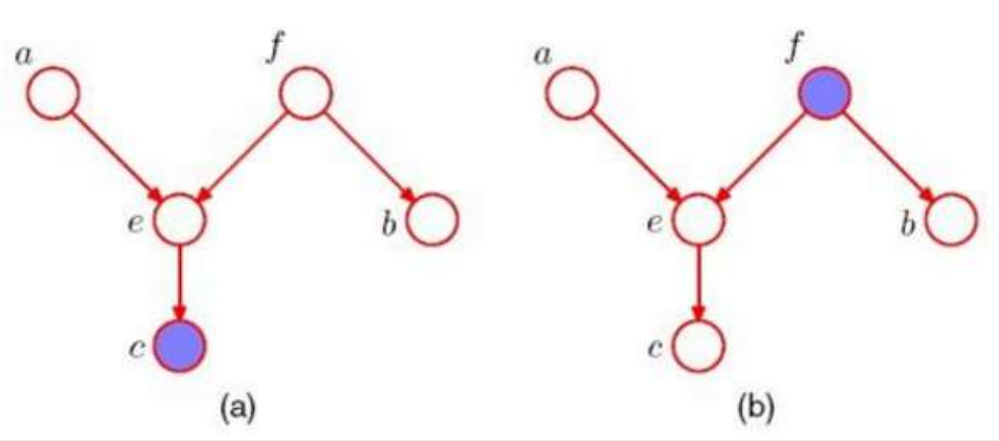
将这三种情况总结，就是贝叶斯网络的一个重要概念，D-separation，这个概念的内容就是:

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

贝叶斯网络推理



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。右边图，f和e都是d-separation.理由与上面相反.

朴素贝叶斯网络

1. 朴素贝叶斯的学习与分类
2. 朴素贝叶斯的参数估计

一、朴素贝叶斯法的学习与分类

- 基本方法
- 后验概率最大化的含义

基本方法

- 训练数据集：

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$

- 由X和Y的联合概率分布P(X,Y)独立同分布产生

- 朴素贝叶斯通过训练数据集学习联合概率分布P(X,Y)，
- 即先验概率分布：$P(Y=c_k),\quad k=1,2,\cdots,K$
- 及条件概率分布：

$$P(X=x\mid Y=c_k)=P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}\mid Y=c_k),\quad k=1,2,\cdots,K$$

- 注意：条件概率为指数级别的参数：$K\prod_{j=1}^{n}S_j$

基本方法

- 条件独立性假设：

$$P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}|Y=c_k)$$
$$=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)$$

- “朴素”贝叶斯名字由来，牺牲分类准确性。

- 贝叶斯定理：$P(Y=c_k|X=x)=\frac{P(X=x|Y=c_k)P(Y=c_k)}{\sum_k P(X=x|Y=c_k)P(Y=c_k)}$
- 代入上式：

$$P(Y=c_k|X=x)=\frac{P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_k P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}$$

基本方法

- 贝叶斯分类器：

$$y = f(x) = \arg\max_{c_k} \frac{P(Y = c_k)\prod_j P(X^{(j)} = x^{(j)} \mid Y = c_k)}{\sum_k P(Y = c_k)\prod_j P(X^{(j)} = x^{(j)} \mid Y = c_k)}$$

- 分母对所有c_k都相同：

$$y = \arg\max_{c_k} P(Y = c_k)\prod_j P(X^{(j)} = x^{(j)} \mid Y = c_k)$$

后验概率最大化的含义

- 朴素贝叶斯法将实例分到后验概率最大的类中，等价于期望风险最小化，
- 假设选择0-1损失函数：f(X)为决策函数

$$L(Y, f(X)) = \begin{cases} 1, & Y \neq f(X) \\ 0, & Y = f(X) \end{cases}$$

- 期望风险函数：

$$R_{\exp}(f) = E[L(Y, f(X))]$$

- 取条件期望：

$$R_{\exp}(f) = E_X \sum_{k=1}^{K} [L(c_k, f(X))] P(c_k \mid X)$$

后验概率最大化的含义

- 只需对X=x逐个极小化，得：

$$
\begin{aligned}
f(x) &= \arg\min_{y\in\mathcal{Y}} \sum_{k=1}^{K} L(c_k, y) P(c_k \mid X = x) \\
&= \arg\min_{y\in\mathcal{Y}} \sum_{k=1}^{K} P(y \neq c_k \mid X = x) \\
&= \arg\min_{y\in\mathcal{Y}} (1 - P(y = c_k \mid X = x)) \\
&= \arg\max_{y\in\mathcal{Y}} P(y = c_k \mid X = x)
\end{aligned}
$$

- 推导出后验概率最大化准则：

$$
f(x) = \arg\max_{c_k} P(c_k \mid X = x)
$$

二、朴素贝叶斯法的参数估计

- 应用极大似然估计法估计相应的概率：
- 先验概率$P(Y=c_k)$的极大似然估计是：

$$P(Y=c_k)=\frac{\sum_{i=1}^{N} I(y_i=c_k)}{N},\quad k=1,2,\cdots,K$$

- 设第j个特征$x^{(j)}$可能取值的集合为：$\{a_{j1},a_{j2},\cdots,a_{jS_j}\}$
- 条件概率的极大似然估计：

$$P(X^{(j)}=a_{jl}\mid Y=c_k)=\frac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl},y_i=c_k)}{\sum_{i=1}^{N} I(y_i=c_k)}$$

$$j=1,2,\cdots,n;\quad l=1,2,\cdots,S_j;\quad k=1,2,\cdots,K$$

朴素贝叶斯法的参数估计

- 学习与分类算法Naïve Bayes Algorithm：
- 输入：
- 训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$
- 第i个样本的第j个特征 $x_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)})^{\mathrm{T}}$
- 第j个特征可能取的第l个值 $x_i^{(j)}\in\{a_{j1},a_{j2},\cdots,a_{jS_j}\}$
- 输出：
- x的分类 $y_i\in\{c_1,c_2,\cdots,c_K\}$

朴素贝叶斯法的参数估计

- 步骤
  - 1、计算先验概率和条件概率

$$P(Y=c_k)=\frac{\sum_{i=1}^{N} I(y_i=c_k)}{N},\quad k=1,2,\cdots,K$$

$$P(X^{(j)}=a_{jl}\mid Y=c_k)=\frac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl},y_i=c_k)}{\sum_{i=1}^{N} I(y_i=c_k)}$$

$$j=1,2,\cdots,n;\quad l=1,2,\cdots,S_j;\quad k=1,2,\cdots,K$$

朴素贝叶斯法的参数估计

- 步骤
- 2、对于给定的实例 $x=(x^{(1)},x^{(2)},\cdots,x^{(n)})^T$
- 计算

$$P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}\mid Y=c_k),\quad k=1,2,\cdots,K$$

- 3、确定x的类别

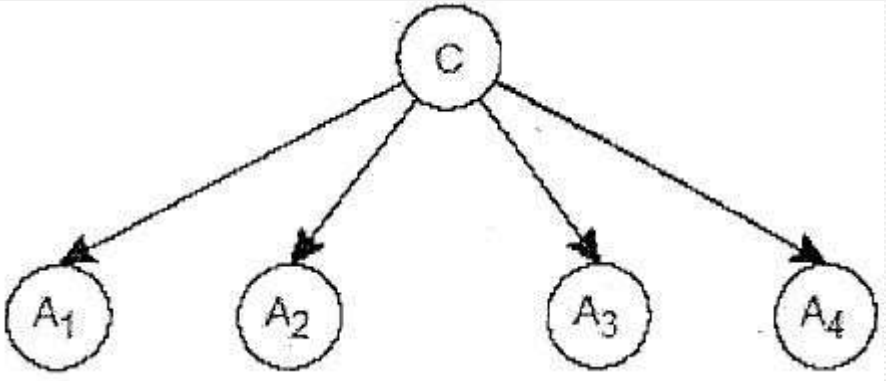
$$y=\arg\max_{c_k}P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}\mid Y=c_k)$$

例子

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

例子

<Outlook=sunny, Temperature=cool, Humidity=high, Wind=strong>



$$c(x) = \arg\max_{c \in \{yes,no\}} P(c)P(sunny \mid c)P(cool \mid c)P(high \mid c)P(strong \mid c)$$

例子

$P(yes)=(9+1)/(14+2)=10/16$

$P(sunny \mid yes)=(2+1)/(9+3)=3/12$

$P(cool \mid yes)=(3+1)/(9+3)=4/12$

$P(high \mid yes)=(3+1)/(9+2)=4/11$

$P(strong \mid yes)=(3+1)/(9+2)=4/11$

$P(no)=(5+1)/(14+2)=6/16$

$P(sunny \mid no)=(3+1)/(5+3)=4/8$

$P(cool \mid no)=(1+1)/(5+3)=2/8$

$P(high \mid no)=(4+1)/(5+2)=5/7$

$P(strong \mid no)=(3+1)/(5+2)=4/7$

$P(yes)P(sunny|yes)P(cool|yes)P(high|yes)P(strong|yes)=0.0069$

$P(no)P(sunny|no)P(cool|no)P(high|no)P(strong|no)=0.0191$

贝叶斯估计

- 考虑：用极大似然估计可能会出现所要估计的概率值为 o的情况，这时会影响到后 验概率的计算结果，使分类 产生偏差.解决这一问题的方法是采用贝叶斯估计。
- 条件概率的贝叶斯估计：

$$P_\lambda(X^{(j)}=a_{jl}\mid Y=c_k)=\frac{\sum_{i=1}^{N}I(x_i^{(j)}=a_{jl},y_i=c_k)+\lambda}{\sum_{i=1}^{N}I(y_i=c_k)+S_j\lambda}$$

- 先验概率的贝叶斯估计：

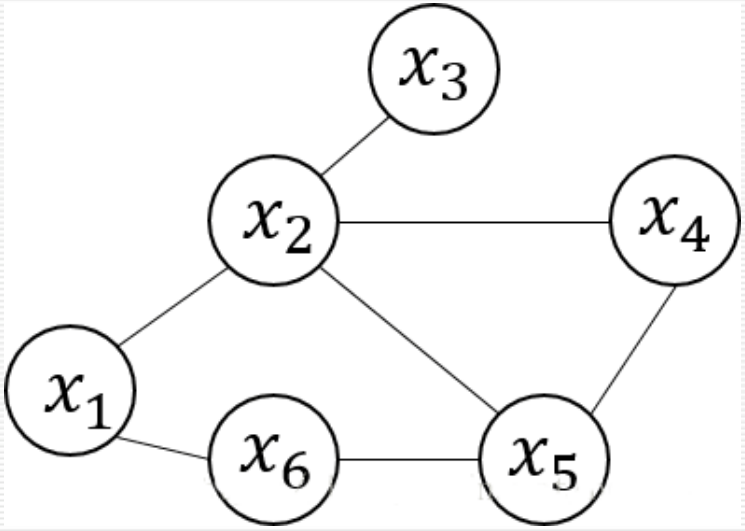
$$P_\lambda(Y=c_k)=\frac{\sum_{i=1}^{N}I(y_i=c_k)+\lambda}{N+K\lambda}$$

马尔科夫随机场

现实生活中，许多任务涉及多个因素（变量），并且因素之间存在依赖关系。概率图模型（Probabilistic Graphical Model，PGM）为表示、学习这种依赖关系提供了一个强大的框架，概率图模型在形式上由图结构组成，一个节点（node）表示一个或一组随机变量，节点之间的边（edge）表示变量之间的关系。

根据图是有向还是无向，概率图模型可以分为两类：
第一类使用有向无环图表示变量之间的因果关系，称为有向图模型或贝叶斯网络（Bayesian network）；
另一类使用无向图表示变量之间的相关关系，称为无向图模型或马尔可夫网（Markov network），马尔可夫随机场（Markov Random Field）。

马尔科夫随机场



马尔科夫随机场

图中的边表示节点之间具有相互关系，这种关系是双向的、对称的。如：x_2和x_3之间有边相连，则x_2和x_3具有相关关系，这种**相关关系采用势函数进行度量**。例如，可以定义如下势函数：

$$\psi(x_2, x_3) = \begin{cases} 1.5 & \text{if } x_2 = x_3; \\ 0.1 & \text{if } otherwise. \end{cases}$$

则说明该模型偏好变量x_2与x_3拥有相同的取值，换言之，在该模型中，x_2与x_3的取值正相关。势函数刻画了局部变量之间的相关关系，它应该是非负的函数。为了满足非负性，指数函数常被用于定义势函数：

$$\psi(x) = e^{-H(x)}$$

$H(x)$是一个定义在变量x上的实值函数，常见形式为：

$$H(x) = \sum_{u,v \in x, u \neq v} \alpha_{uv} x_u x_v + \sum_{v \in x} \beta_v x_v$$

其中α_{uv}和β_v是需要学习的参数，称为参数估计。

马尔科夫随机场

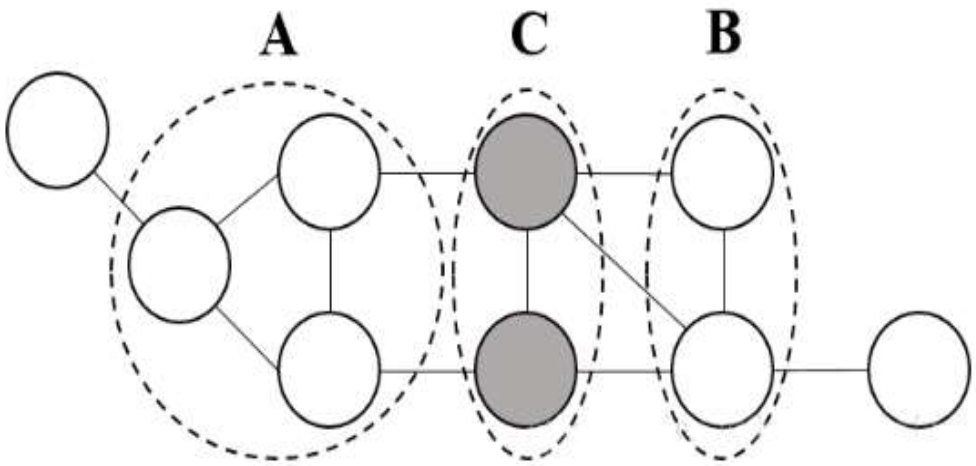
MRF的马尔可夫性

马尔可夫随机场是生成式模型，生成式模型最关心的是变量的联合概率分布。假设我们有n个取值为二值随机变量（$x_1, x_2, \cdots, x_n$），其取值分布将包含2^n种可能，因此确定联合概率分布$p(x_1, x_2, \cdots, x_n)$需要$2^n - 1$个参数，这个复杂度通常是我们不能接受的；而另一种极端情况是，当所有变量都相互独立时，$p(x_1, x_2, \cdots, x_n) = p(x_1)p(x_2)\cdots p(x_n)$只需要$n$个参数。因此，我们可能会思考，能不能将联合概率分布分解为一组子集概率分布的乘积呢？那么应该怎么划分子图呢？应该遵循怎样的原则？首先定义马尔可夫随机场中随机变量之间的全局马尔可夫性、局部马尔可夫性和成对马尔可夫性。

马尔科夫随机场

- **全局马尔可夫性（global Markov property）**：设节点集合A，B是在无向图G中被节点集C分开的任意节点集合，如下图所示。全局马尔可夫性是指在给定x_C的条件下，x_A和x_B条件独立，记为$x_A \perp x_B | x_C$。

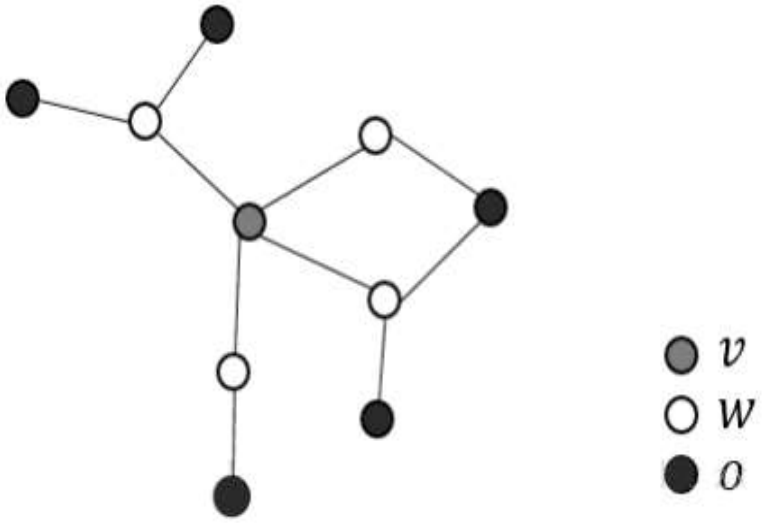
$$p(x_A, x_B | x_C) = p(x_A | x_C) p(x_B | x_C)$$



马尔科夫随机场

- **局部马尔可夫性（local Markov property）**：给定变量v的所有邻接变量w，则该变量v条件独立于其他变量o。即在给定某个变量的邻接变量的取值条件下，该变量的取值将于其他变量无关。

$$p(x_v, x_v | x_w) = p(x_v | x_w) p(x_o | x_w)$$

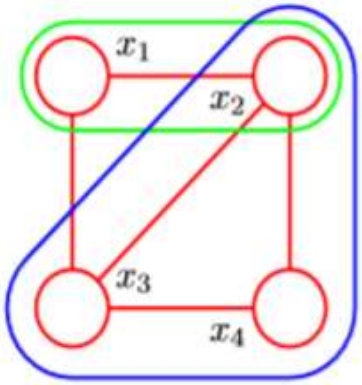


马尔科夫随机场

- **成对马尔可夫性（pairwise Markov property）**：给定所有其他变量，两个非邻接变量条件独立。这是因为两个节点没有直接路径，并且所有其他路径上都有确定的观测节点，因此这些路径也将被阻隔。

$$p(x_i, x_j | x_{\setminus\{i,j\}}) = p(x_i | x_{\setminus\{i,j\}}) p(x_j | x_{\setminus\{i,j\}})$$

其中$x_{\setminus\{i,j\}}$表示所有变量x去除x_i和x_j的集合。于是，**联合概率分布的分解一定要让x_i和x_j不出现在同一个划分中，从而让属于这个图的所有可能概率分布都满足条件独立性质。**让非邻接变量不出现在同一个划分中，即每一个划分中节点都是全连接的。这将我们引向了图的一个概念，团（clique）。它被定义为图中节点的一个子集，并且这个子集中任意两节点间都有边相连。若在一个团中加入其他任何节点都不再形成团，则称该团为极大团。下图给出了团和极大团的一个示意：



图中，绿色圆圈是一个团，蓝色圆圈是一个极大团。显然，最简单的团就是两个节点以及一条边，而我们最开始就针对两节点之间的相关关系（每条边）定义了势函数。**因此，马尔可夫随机场中，多个变量的联合概率分布能基于团分解为多个势函数的乘积，每一个团对应一个势函数。**

$$p(x)=\frac{1}{Z}\prod_{C}\psi_C(x_C)$$

其中，如果C是一个团，ψ_C为团C对应的势函数。$Z=\sum_x\prod_C\psi_C(x_C)$是归一化因子，以确保$p(x)$是正确定义的概率。对图中每一条边都定义一个势函数$\psi$，将导致模型的势函数过多，带来计算负担。例如，最上面的图中x_2、x_4、x_5分别定义需要定义三个势函数，但是x_2、x_4、x_5两两相关，x_2、x_4与x_5的取值将相互影响，因此可以整体定义一个势函数$\psi(x_2,x_4,x_5)$表示三者取值的偏好。所以可以**将联合概率分布分解为其极大团上的势函数的乘积**：

$$p(x)=\frac{1}{Z^*}\prod_{Q\in C^*}\psi_Q(x_Q)$$

其中C^*是极大团构成的集合，$Z^*=\sum_x\prod_{Q\in C^*}\psi_Q(x_Q)$。例如，最上面的图中$x=\{x_1,x_2,\cdots,x_6\}$，联合概率分布$p(x)$定义为

$$p(x)=\frac{1}{Z}\psi_{12}(x_1,x_2)\psi_{16}(x_1,x_6)\psi_{23}(x_2,x_3)\psi_{56}(x_5,x_6)\psi_{245}(x_2,x_4,x_5)$$

马尔科夫随机场应用

图像去噪

令观测的噪声图像通过一个二值像素值 $y_i \in \{-1, +1\}$ 组成的数组来描述,其中下标 $i = 1, \ldots, D$ 覆盖了所有的像素。我们假设图像通过下面的方式获得:取一张未知的无噪声图像,这幅图像由二值像素值 $x_i \in \{-1, +1\}$ 描述,然后以一个较小的概率随机翻转像素值的符号。

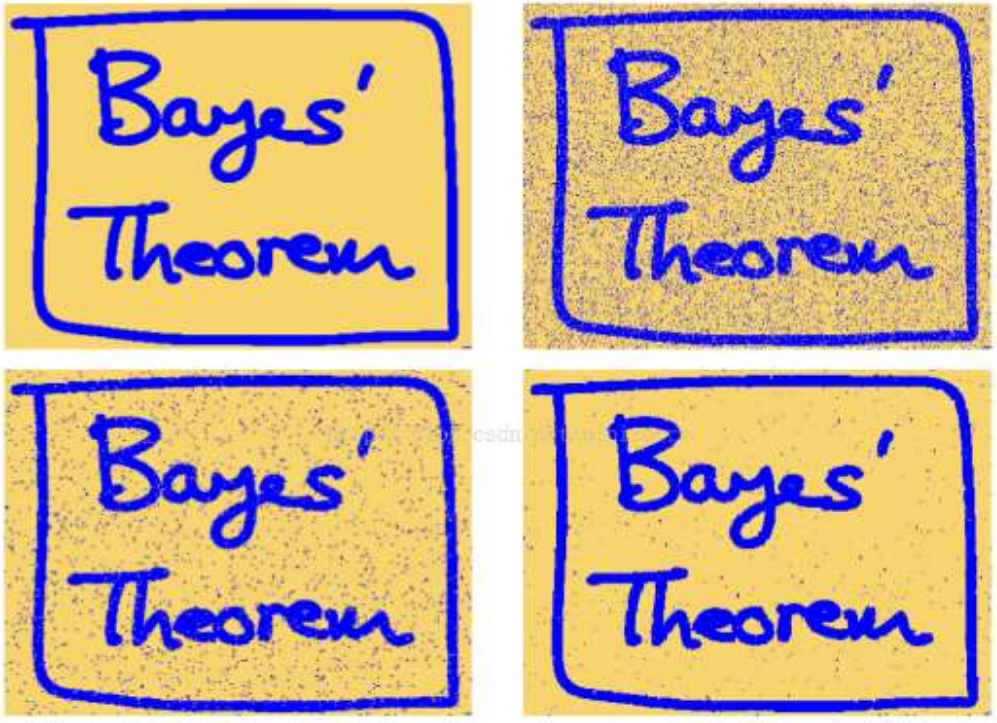


图 8.30: 使用马尔科夫随机场进行图像去噪的例子。上面一行中，左侧是原始的二值图像，右侧是随机改变10%的像素后得到的带有噪声的图像。下面一行中，左图是使用迭代条件模型（ICM）恢复的图像，右图是使用最大割算法得到的图像。ICM产生的图像中，96%的像素与原始图像相符，而最大割算法产生的图像中，这个比例为99%。

马尔科夫随机场应用

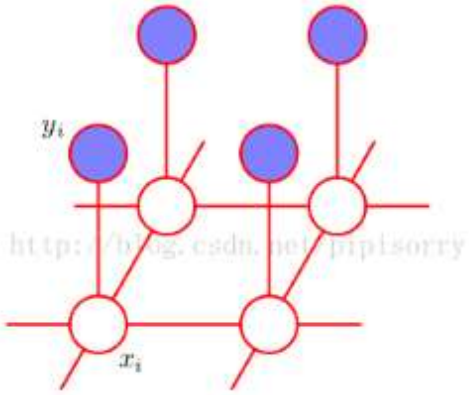


图 8.31: 一个无向图模型，表示图像去噪的马尔科夫随机场，其中x_i是一个二值变量，表示像素i在一个未知的无噪声的图像中的状态，y_i表示在观测到的噪声图像中，像素i的对应值。

由于噪声等级比较小,因此我们知道 x i 和 y i 之间有着强烈的相关性。我们还知道图像中相邻像素 x i 和 x j 的相关性很强。这种先验知识可以使用马尔科夫随机场模型进行描述,它的无向图如图8.31所示。

图中两种类型的团块

形如 {x i , y i } 的团块有一个关联的能量函数,表达了这些变量之间的相关性。对于这些团块,我们选择一个非常简单的能量函数 −ηx i y i , 其中 η 是一个正的常数。这个能量函数的效果是:当 x i 和 y i 符号相同时,能量函数会给出一个较低的能量(即,较高的概率),而当 x i 和 y i 符号相反时,能量函数会给出一个较高的能量。

马尔科夫随机场应用

由变量 {x i , x j } 组成的团块,其中 i 和 j 是相邻像素的下标。与之前一样,我们希望当两个像素符号相同时能量较低,当两个像素符号相反时能量较高,因此我们选择能量函数 −βx i x j ,其中 β 是一个正的常数。

我们可以为无噪声图像的每个像素 i 加上一个额外的项 hx i 。这样的项具有下面的效果:将模型进行偏置,使得模型倾向于选择一个特定的符号,而不选择另一个符号。

模型的完整的能量函数的形式为

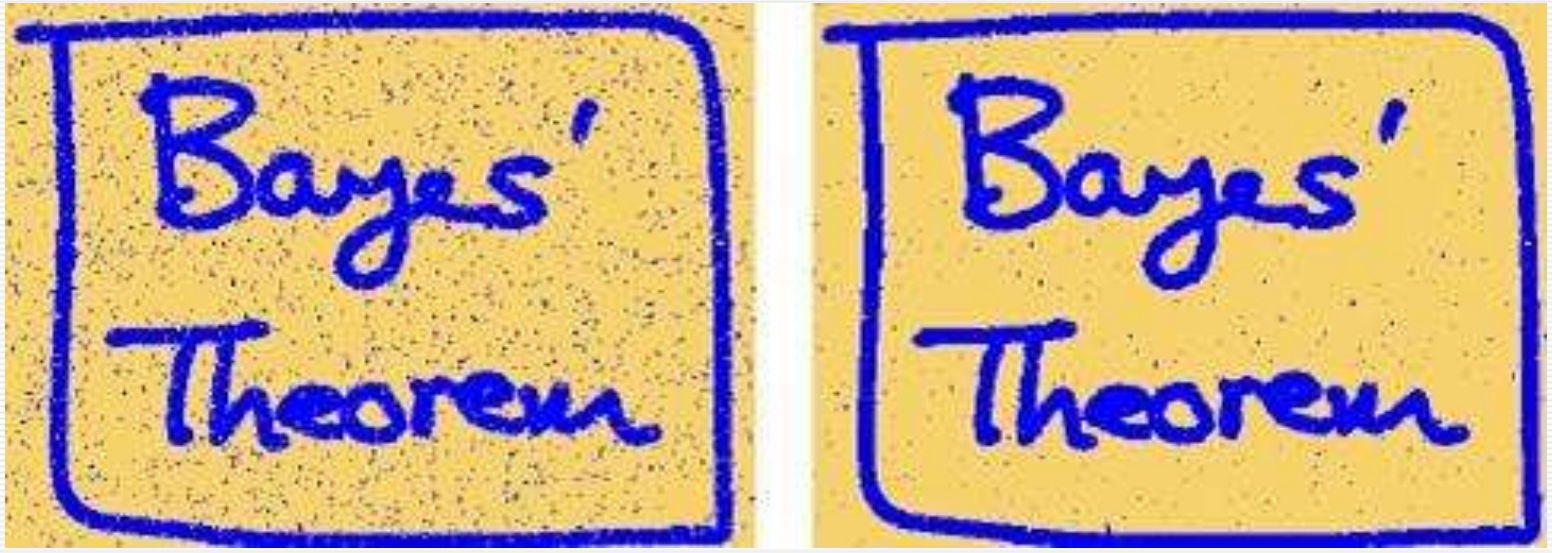
$$E(\boldsymbol{x}, \boldsymbol{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i \qquad (8.42)$$

Note: 令 h = 0 意味着两个状态 x i 的先验概率是相等的。令 β = 0 ,从而去除了相邻像素点之间的联系,那么整体概率最大的解为 x i = y i (对于所有的 i),这对应于观测到的噪声图像。

x 和 y 的联合概率分布,形式为

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

马尔科夫随机场应用



条件随机场

从随机场到马尔科夫随机场

随机场是由若干个位置组成的整体，当给每一个位置中按照某种分布随机赋予一个值之后，其全体就叫做随机场。

举词性标注的例子：假如我们有一个十个词形成的句子需要做词性标注。这十个词每个词的词性可以在我们已知的词性集合（名词，动词...)中去选择。当我们为每个词选择完词性后，这就形成了一个随机场。

马尔科夫随机场是随机场的特例，它假设随机场中某一个位置的赋值仅仅与和它相邻的位置的赋值有关，和与其不相邻的位置的赋值无关。

继续举十个词的句子词性标注的例子：　如果我们假设所有词的词性只和它相邻的词的词性有关时，这个随机场就特化成一个马尔科夫随机场。比如第三个词的词性除了与自己本身的位置有关外，只与第二个词和第四个词的词性有关。

条件随机场

从马尔科夫随机场到条件随机场

理解了马尔科夫随机场，再理解CRF就容易了。CRF是马尔科夫随机场的特例，它假设马尔科夫随机场中只有X和Y两种变量，X一般是给定的，而Y一般是在给定X的条件下我们的输出。这样马尔科夫随机场就特化成了条件随机场。在我们十个词的句子词性标注的例子中，X是词，Y是词性。因此，如果我们假设它是一个马尔科夫随机场，那么它也就是一个CRF。

对于CRF，我们给出准确的数学语言描述：

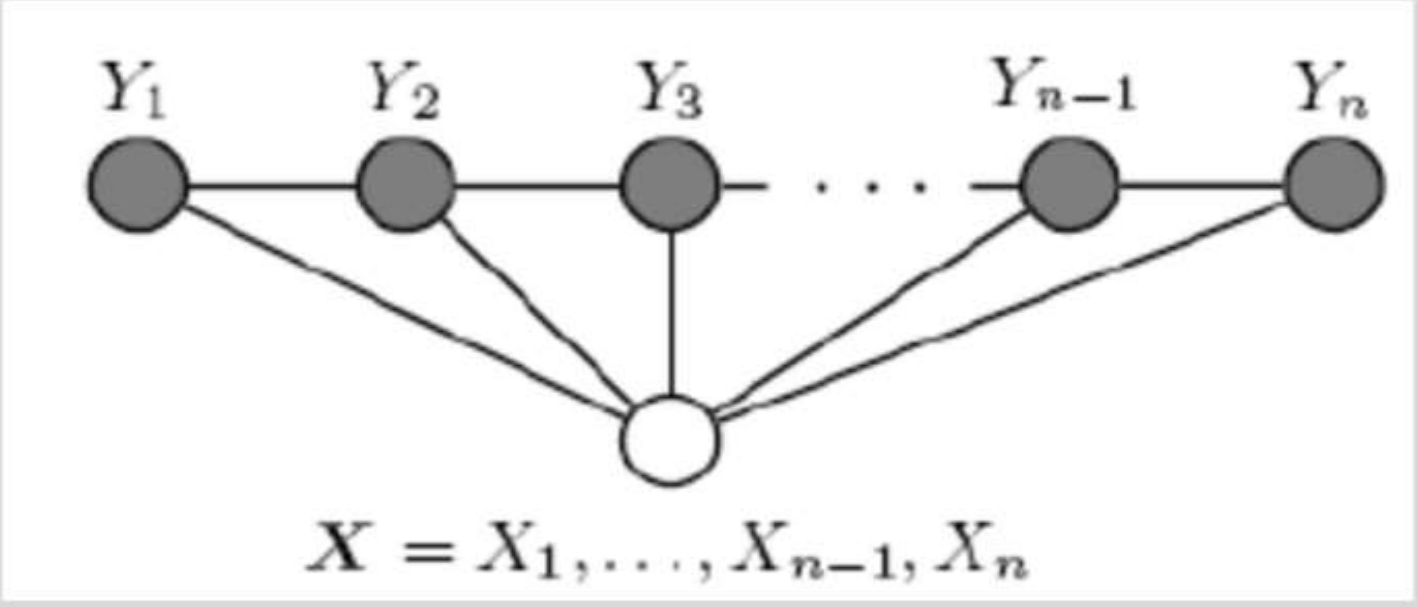
设X与Y是随机变量，$P(Y|X)$是给定X时Y的条件概率分布，若随机变量Y构成的是一个马尔科夫随机场，则称条件概率分布$P(Y|X)$是条件随机场。

从条件随机场到线性链条件随机场

注意在CRF的定义中，我们并没有要求X和Y有相同的结构。而实现中，我们一般都假设X和Y有相同的结构，即：

$$X=(X_1,X_2,\dots X_n),\ Y=(Y_1,Y_2,\dots Y_n)$$

我们一般考虑如下图所示的结构：X和Y有相同的结构的CRF就构成了线性链条件随机场(Linear chain Conditional Random Fields,以下简称 linear-CRF)。



在我们的十个词的句子的词性标记中，词有十个，词性也是十个，因此，如果我们假设它是一个马尔科夫随机场，那么它也就是一个linear-CRF。

我们再来看看 linear-CRF的数学定义：

设$X=(X_1,X_2,\dots X_n),\ Y=(Y_1,Y_2,\dots Y_n)$均为线性链表示的随机变量序列，在给定随机变量序列$X$的情况下，随机变量$Y$的条件概率分布$P(Y|X)$构成条件随机场，即满足马尔科夫性：

$$P(Y_i|X,Y_1,Y_2,\dots Y_n)=P(Y_i|X,Y_{i-1},Y_{i+1})$$

则称$P(Y|X)$为线性链条件随机场。

在linear-CRF中，特征函数分为两类，第一类是定义在Y节点上的节点特征函数，这类特征函数只和当前节点有关，记为：

$$s_l(y_i, x, i), \ \ l = 1, 2, \ldots L$$

其中L是定义在该节点的节点特征函数的总个数，i是当前节点在序列的位置。

第二类是定义在Y上下文的局部特征函数，这类特征函数只和当前节点和上一个节点有关，记为：

$$t_k(y_{i-1}, y_i, x, i), \ \ k = 1, 2, \ldots K$$

其中K是定义在该节点的局部特征函数的总个数，i是当前节点在序列的位置。之所以只有上下文相关的局部特征函数，没有不相邻节点之间的特征函数，是因为我们的linear-CRF满足马尔科夫性。

无论是节点特征函数还是局部特征函数，它们的取值只能是0或者1。即满足特征条件或者不满足特征条件。同时，我们可以为每个特征函数赋予一个权值，用以表达我们对这个特征函数的信任度。假设t_k的权重系数是λ_k,s_l的权重系数是μ_l,则linear-CRF由我们所有的$t_k, \lambda_k, s_l, \mu_l$共同决定。

此时我们得到了linear-CRF的参数化形式如下：

$$P(y|x) = \frac{1}{Z(x)} exp\Big(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\Big)$$

其中，$Z(x)$为规范化因子：

$$Z(x) = \sum_{y} exp\Big(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\Big)$$

回到特征函数本身，每个特征函数定义了一个linear-CRF的规则，则其系数定义了这个规则的可信度。所有的规则和其可信度一起构成了我们的linear-CRF的最终的条件概率分布。

这里我们给出一个linear-CRF用于词性标注的实例，为了方便，我们简化了词性的种类。假设输入的都是三个词的句子，即$X=(X_1,X_2,X_3)$,输出的词性标记为$Y=(Y_1,Y_2,Y_3)$,其中$Y\in\{1(名词),2(动词)\}$

这里只标记出取值为1的特征函数如下：

$$t_1=t_1(y_{i-1}=1,y_i=2,x,i),i=2,3,\ \lambda_1=1$$

$$t_2=t_2(y_1=1,y_2=1,x,2)\ \lambda_2=0.5$$

$$t_3=t_3(y_2=2,y_3=1,x,3)\ \lambda_3=1$$

$$t_4=t_4(y_1=2,y_2=1,x,2)\ \lambda_4=1$$

$$t_5=t_5(y_2=2,y_3=2,x,3)\ \lambda_5=0.2$$

$$s_1=s_1(y_1=1,x,1)\ \mu_1=1$$

$$s_2=s_2(y_i=2,x,i),i=1,2,\ \mu_2=0.5$$

$$s_3=s_3(y_i=1,x,i),i=2,3,\ \mu_3=0.8$$

$$s_4=s_4(y_3=2,x,3)\ \mu_4=0.5$$

求标记(1,2,2)的非规范化概率。

利用linear-CRF的参数化公式，我们有：

$$P(y|x)\propto exp\Big[\sum_{k=1}^{5}\lambda_k\sum_{i=2}^{3}t_k(y_{i-1},y_i,x,i)+\sum_{l=1}^{4}\mu_l\sum_{i=1}^{3}s_l(y_i,x,i)\Big]$$

带入(1,2,2)我们有：

$$P(y_1=1,y_2=2,y_3=2|x)\propto exp(3.2)$$

假设我们在某一节点我们有K_1个局部特征函数和K_2个节点特征函数，总共有$K=K_1+K_2$个特征函数。我们用一个特征函数$f_k(y_{i-1},y_i,x,i)$来统一表示如下：

$$f_k(y_{i-1},y_i,x,i)=\begin{cases}t_k(y_{i-1},y_i,x,i) & k=1,2,\dots K_1\\ s_l(y_i,x,i) & k=K_1+l,\ l=1,2\dots,K_2\end{cases}$$

对$f_k(y_{i-1},y_i,x,i)$在各个序列位置求和得到：

$$f_k(y,x)=\sum_{i=1}^{n}f_k(y_{i-1},y_i,x,i)$$

同时我们也统一$f_k(y_{i-1},y_i,x,i)$对应的权重系数w_k如下：

$$w_k=\begin{cases}\lambda_k & k=1,2,\dots K_1\\ \mu_l & k=K_1+l,\ l=1,2\dots,K_2\end{cases}$$

这样，我们的linear-CRF的参数化形式简化为：

$$P(y|x)=\frac{1}{Z(x)}exp\sum_{k=1}^{K}w_kf_k(y,x)$$

其中，$Z(x)$为规范化因子：

$$Z(x)=\sum_{y}exp\sum_{k=1}^{K}w_kf_k(y,x)$$

如果将上两式中的w_k与f_k的用向量表示，即：

$$w=(w_1,w_2,\dots w_K)^T\quad F(y,x)=(f_1(y,x),f_2(y,x),\dots f_K(y,x))^T$$

则linear-CRF的参数化形式简化为内积形式如下：

$$P_w(y|x)=\frac{exp(w\bullet F(y,x))}{Z_w(x)}=\frac{exp(w\bullet F(y,x))}{\sum_{y}exp(w\bullet F(y,x))}$$

谢谢