机器学习
第8节

涉及知识点：
贝叶斯网络、概率图模型中的条件独立、马尔可夫网络简介、马尔可夫网络应用示例、马尔可夫网络vs.贝叶斯网络、链式模型推断、树图模型推断

概率图模型

张伟楠 – 上海交通大学

本节部分素材取自Chris Bishop PRML第8章

贝叶斯网络

张伟楠 – 上海交通大学

目录

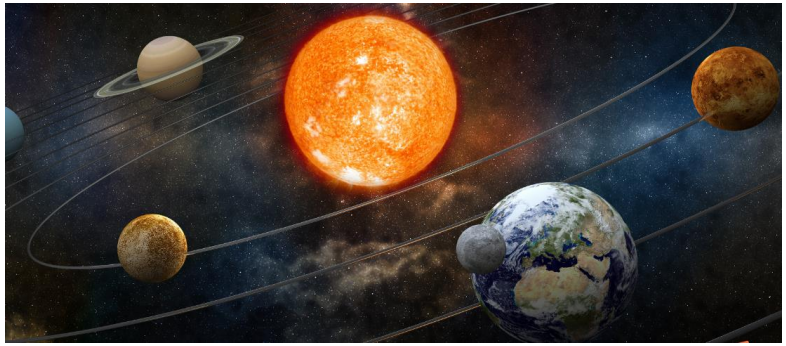
Contents

01
数据科学回顾

什么是数据科学?

物理学

- 目标：探究世界的基本原理



- 解决方法：从观测结果中构建世界模型

$$F = G\frac{m_1 m_2}{r^2}$$

数据科学

- 目标：发现数据的基本原理

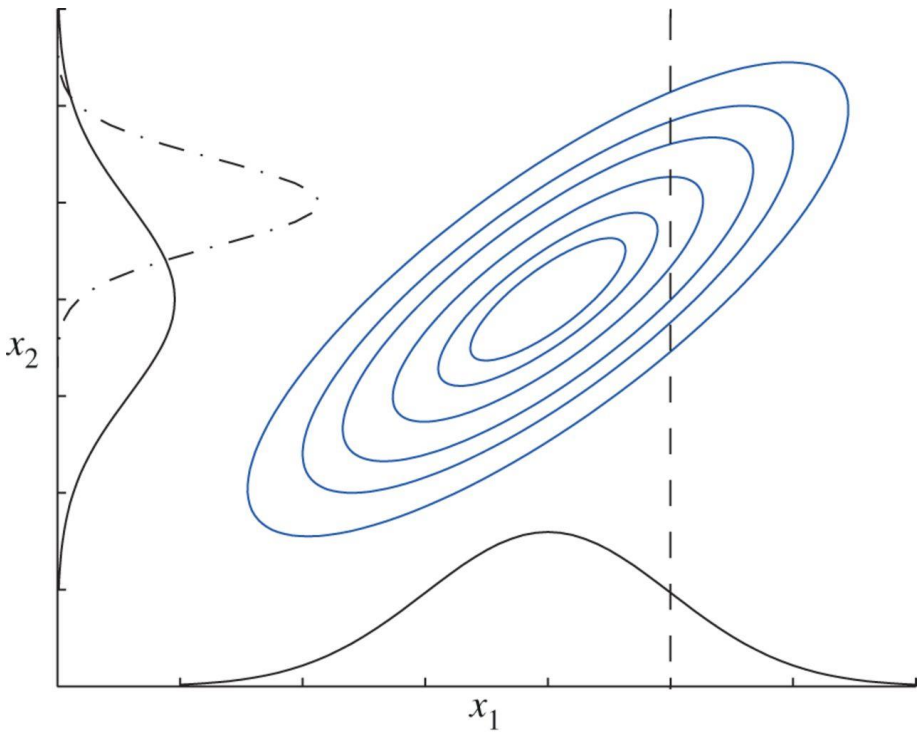
- 解决方法：从观测结果中构建数据模型

$$p(x) = \frac{e^{f(x)}}{\sum_{x'} e^{f(x')}}$$

数据科学

数学上

- 找到联合数据分布$p(x)$
- 找到条件分布$p(x_2|x_1)$



高斯分布

- 多元

$$p(x) = \frac{e^{-(x-\mu)^T \Sigma^{-1}(x-\mu)}}{\sqrt{|2\pi\Sigma|}}$$

- 一元

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

一个用户行为建模的简单例子

Interest	Gender	Age	BBC Sports	PubMed	Bloomberg Business	Spotify
Finance	Male	29	Yes	No	Yes	No
Sports	Male	21	Yes	No	No	Yes
Medicine	Female	32	No	Yes	No	No
Music	Female	25	No	No	No	Yes
Medicine	Male	40	Yes	Yes	Yes	No

- 联合数据分布

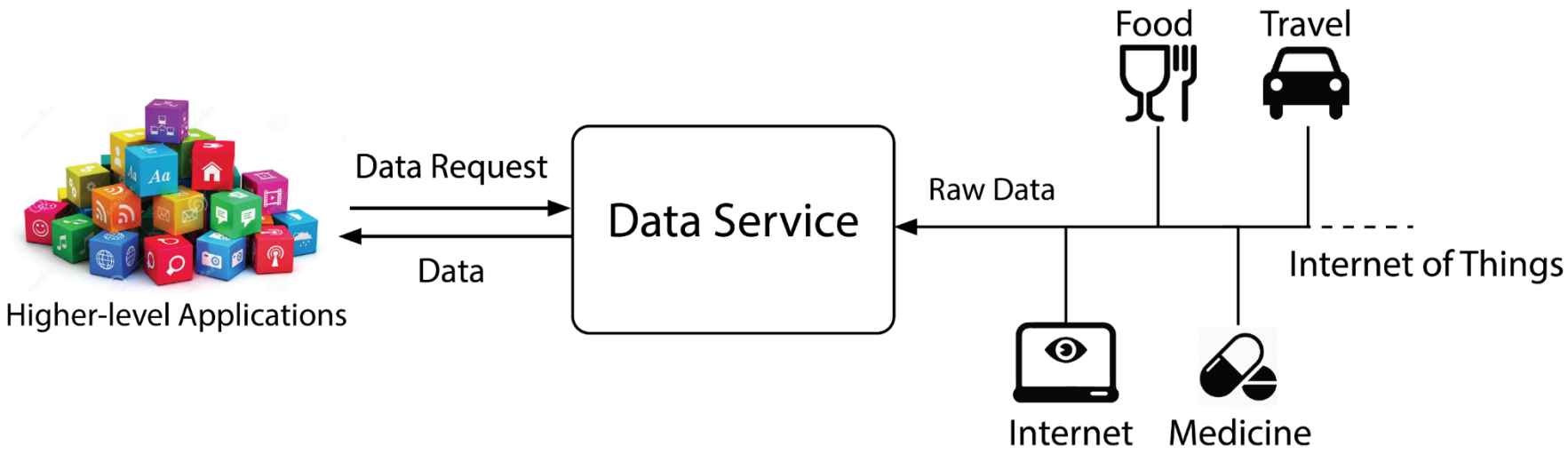
p(Interest=Finance, Gender=Male, Age=29, Browsing=BBC Sports,Bloomberg Business)

- 条件数据分布

p(Interest=Finance | Browsing=BBC Sports,Bloomberg Business)

p(Gender=Male | Browsing=BBC Sports,Bloomberg Business)

数据处理技术


Food
Travel
Data Request
Data Service
Raw Data
Data
Internet of Things
Higher-level Applications
Internet
Medicine

02

概率图模型

数据科学的关键问题

- 如何构建数据模型?

- 具体来说，如何建模联合数据分布 $p(x)$?
  - 比如：温度和人的穿衣数据

Temperature	Cloth	Probability
Hot	Shirt	48%
Hot	Coat	12%
Cold	Shirt	8%
Cold	Coat	32%

数据概率建模

Temperature	Cloth	Probability
Hot	Shirt	48%
Hot	Coat	12%
Cold	Shirt	8%
Cold	Coat	32%

- 从表中，我们可以直接建立联合分布模型

$$
\begin{aligned}
&\mathrm{P}(\mathrm{temperature} = \mathrm{hot}, \mathrm{cloth} = \mathrm{shirt}) = 48\% \\
&\mathrm{P}(\mathrm{temperature} = \mathrm{hot}, \mathrm{cloth} = \mathrm{coat}) = 12\% \\
&\mathrm{P}(\mathrm{temperature} = \mathrm{cold}, \mathrm{cloth} = \mathrm{shirt}) = 8\% \\
&\mathrm{P}(\mathrm{temperature} = \mathrm{cold}, \mathrm{cloth} = \mathrm{coat}) = 32\%
\end{aligned}
$$

  - 需要估计和维护$2 \times 2 = 4$个概率值

数据概率建模

- 如果我们有高维数据怎么办?

Temperature	Cloth	Gender	Weekday	Probability
Hot	Shirt	Male	Monday	2.4%
Hot	Coat	Female	Friday	1.2%
Cold	Shirt	Female	Sunday	3.8%
Cold	Coat	Male	Thursday	3.1%

…

- 直接建立联合分布模型需要估计和维护$2 \times 2 \times 2 \times 7 = 56$个概率值
  - 指数级复杂度
- 我们需要找到一种更好的方法建模数据分布

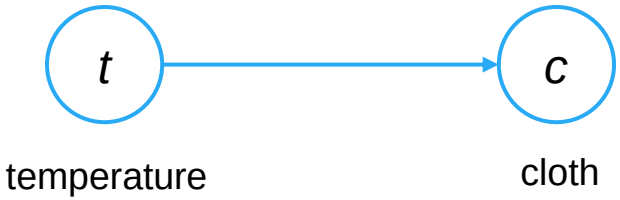
领域知识(Domain Knowledge)

Temperature	Cloth	Probability
Hot	Shirt	48%
Hot	Coat	12%
Cold	Shirt	8%
Cold	Coat	32%

- 利用领域知识构建数据依赖
  - 人们根据温度选择穿衣
  - 因此穿衣变量取决于温度变量

$$p(t,c) = p(t)p(c|t)$$

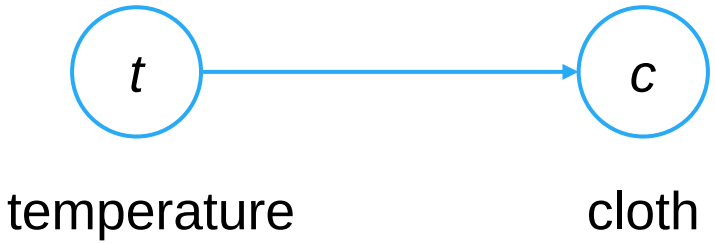
P(temperature=hot) = 60%
P(temperature=cold) = 40%



$\mathrm{P(cloth = shirt|temperature = hot)} = 80\%$
$\mathrm{P(cloth = coat|temperature = hot)} = 20\%$
$\mathrm{P(cloth = shirt|temperature = cold)} = 20\%$
$\mathrm{P(cloth = coat|temperature = cold)} = 80\%$

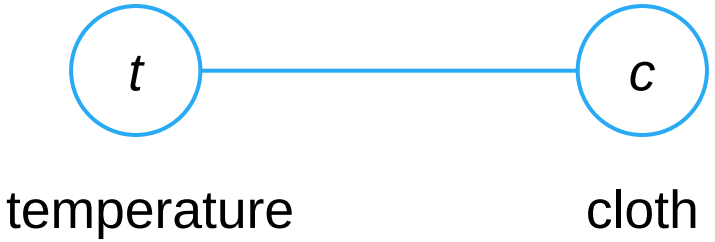
图模型

- 图模型是一种形式化任意领域知识数据依赖的方法
  - 贝叶斯网络(有向图)



$$p(t,c) = p(t)p(c|t)$$

  - 马尔可夫网络(无向图)



$$p(t,c) = \frac{e^{\phi(t,c)}}{\sum_{t',c'} e^{\phi(t',c')}}$$

03

贝叶斯网络

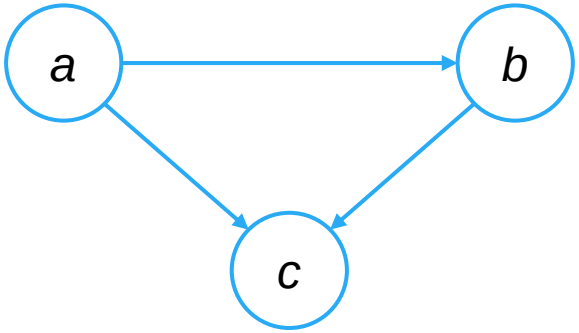
一个简单的贝叶斯网络

- 考虑一个任意的联合分布$p(a,b,c)$
- 运用概率的乘积规则

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

关于a,b,c对称

关于a,b,c不对称



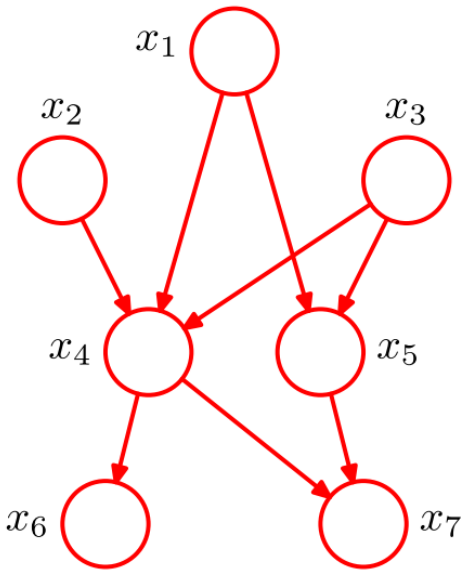
- 图模型一个强大的功能是一个给定概率图可以为广泛的分布类型生成概率描述

- 我们说这个图是全连接的，因为每对节点之间都有一个连接

一个更加复杂的贝叶斯网络

- 一个7维的数据分布

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



- 具有K个节点的图的联合分布

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

x_k的父节点集合

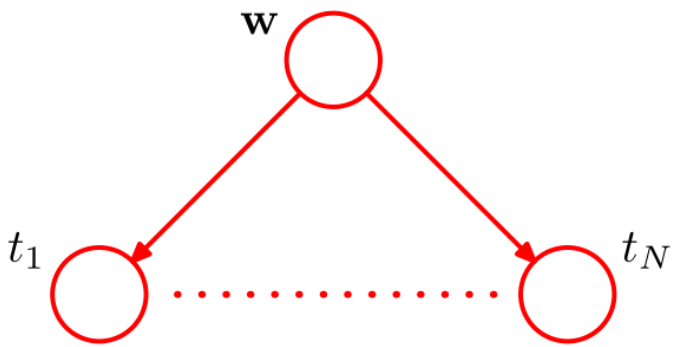
- 一个重要的限制:
  有向无环图(DAG)

回归模型例子

- 训练数据 $D = \{(x_i, t_i)\}$
- 我们建立了一个带有观察高斯噪声的线性预测模型

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{i=1}^{N} p(t_i|\mathbf{w})$$

先验分布

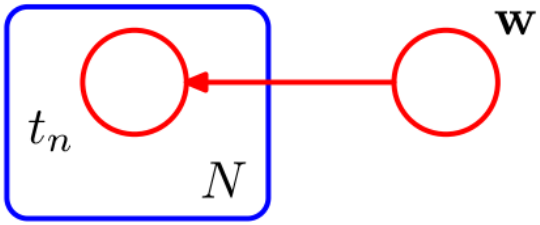


- 进一步来说

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha)$$

$$p(t_i|x_i, \mathbf{w}, \sigma^2) = \mathcal{N}(t_i|\mathbf{w}^\top x_i, \sigma^2)$$

$$p(\mathbf{t}, \mathbf{w}|\mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{i=1}^{N} p(t_i|x_i, \mathbf{w}, \sigma^2)$$



图的另一种更紧凑的表示形式

回归模型例子

- 训练数据 $D = \{(x_i, t_i)\}$
- 我们建立了一个带有观察高斯噪声的线性预测模型

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{i=1}^{N} p(t_i|\mathbf{w})$$

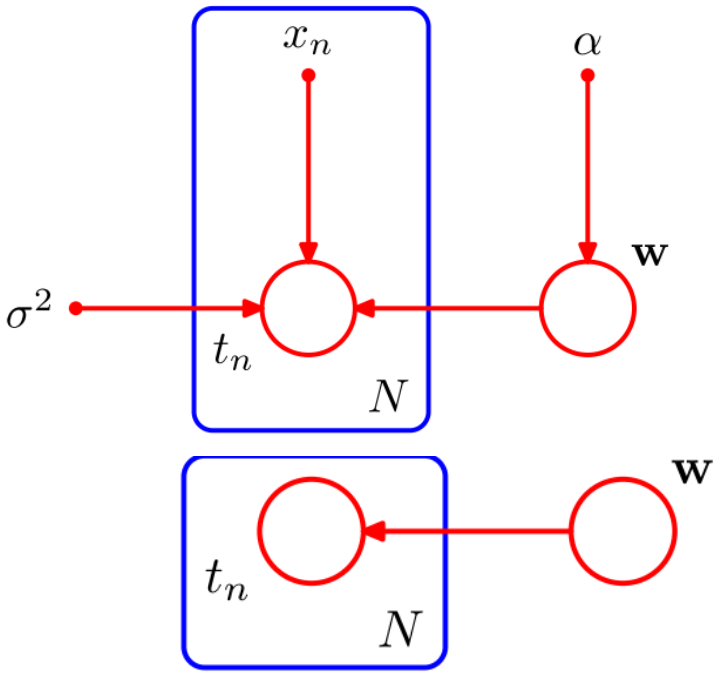
先验分布

- 进一步来说

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha)$$

$$p(t_i|x_i, \mathbf{w}, \sigma^2) = \mathcal{N}(t_i|\mathbf{w}^\top x_i, \sigma^2)$$

$$p(\mathbf{t}, \mathbf{w}|\mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{i=1}^{N} p(t_i|x_i, \mathbf{w}, \sigma^2)$$



图的另一种更紧凑的表示形式

后验分布

- 当观测到$\{t_n\}$，我们可以估计参数**w**的后验分布

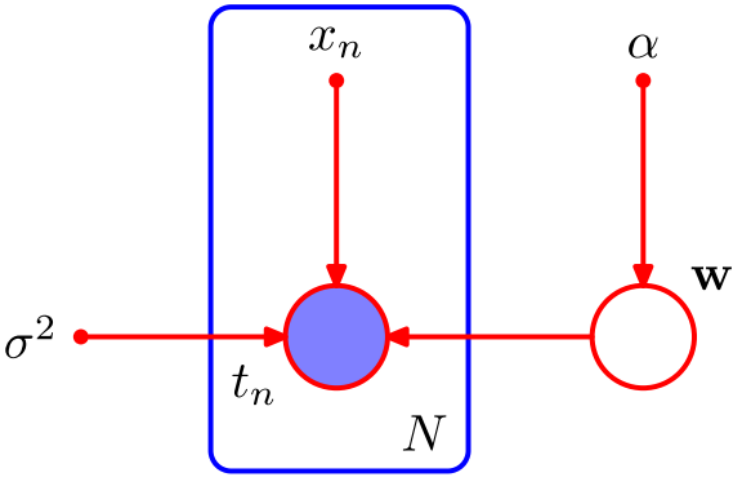
$$p(\mathbf{w}|\mathbf{t}) = \frac{p(\mathbf{w})p(\mathbf{t}|\mathbf{w})}{p(\mathbf{t})}$$

$$\propto p(\mathbf{w}) \prod_{i=1}^{N} p(t_i|\mathbf{w})$$

后验分布
Posterior

先验分布
Prior

数据似然
Likelihood



最大后验估计

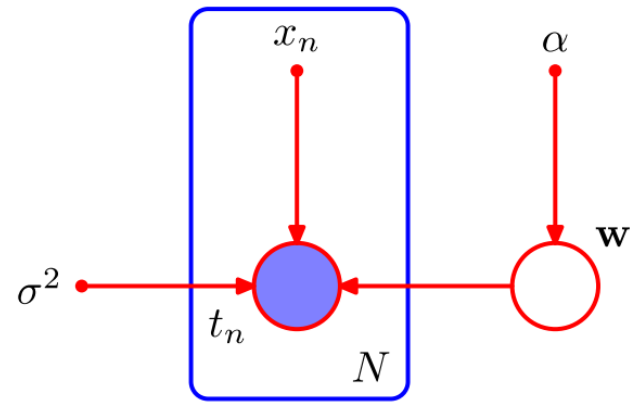
- 模型参数w最大后验(Maximum A Posteriori, MAP)估计

$$\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{t}) = \max_{\mathbf{w}} p(\mathbf{w},\mathbf{t}) = \max_{\mathbf{w}} p(\mathbf{w})p(\mathbf{t}|\mathbf{w})$$

$$p(\mathbf{w})p(\mathbf{t}|\mathbf{w}) = p(\mathbf{w}|\alpha)\prod_{i=1}^{N} p(t_i|x_i,\mathbf{w},\sigma^2)$$

$$= \mathcal{N}(\mathbf{w}|\mathbf{0},\alpha)\prod_{i=1}^{N} \mathcal{N}\left(t_i|\mathbf{w}^\top x_i,\sigma^2\right)$$

$$= \frac{1}{\sqrt{(2\pi\alpha)^d}}\exp\left(-\frac{\mathbf{w}^\top\mathbf{w}}{2\alpha^d}\right)\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(t_i-\mathbf{w}^\top x_i)^2}{2\sigma^2}\right)$$



$$\log p(\mathbf{w})p(\mathbf{t}|\mathbf{w}) = -\frac{\mathbf{w}^\top\mathbf{w}}{2\alpha^d} - \sum_{i=1}^{N}\frac{(t_i-\mathbf{w}^\top x_i)^2}{2\sigma^2} + \text{const}$$

等价于 $\min_{\mathbf{w}} \frac{\sigma^2}{\alpha^d}\|\mathbf{w}\|^2 + \sum_{i=1}^{N}(t_i-\mathbf{w}^\top x_i)^2$ 即平方损失的岭回归

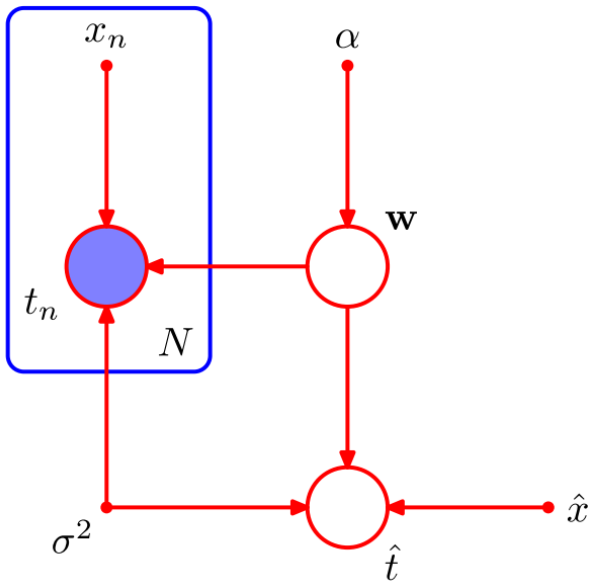
预测新样本

- 给定新的输入值$\hat{x}$，预测其标签$\hat{t}$的对应概率分布
  - 随机变量的联合分布

$$p(\hat{t}, \mathbf{t}, \mathbf{w}|\hat{x}, \mathbf{x}, \alpha, \sigma^2) = \left[\prod_{i=1}^{N} p(t_i|x_i, \mathbf{w}, \sigma^2)\right] p(\mathbf{w}|\alpha) p(\hat{t}|\hat{x}, \mathbf{w}, \sigma^2)$$

  - 边缘化参数**w** (marginalize out **w**)

$$\begin{aligned} p(\hat{t}|\hat{x}, \mathbf{x}, \mathbf{t}, \alpha, \sigma^2) &= \frac{p(\hat{t}, \mathbf{t}|\hat{x}, \mathbf{x}, \alpha, \sigma^2)}{p(\mathbf{t})} \\ &\propto p(\hat{t}, \mathbf{t}|\hat{x}, \mathbf{x}, \alpha, \sigma^2) \\ &= \int p(\hat{t}, \mathbf{t}, \mathbf{w}|\hat{x}, \mathbf{x}, \alpha, \sigma^2) d\mathbf{w} \end{aligned}$$



概率图模型中的条件独立

张伟楠 – 上海交通大学

目录

Contents

01

条件独立

条件独立

定义

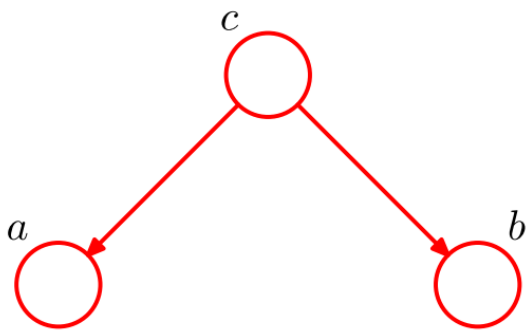
- 考虑三个变量a, b, c
- 假设给定b和c的情况下，a的条件分布不依赖于b的值

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

- 则在给定c的情况下a条件独立于b

等价表达形式

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



表示符号

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

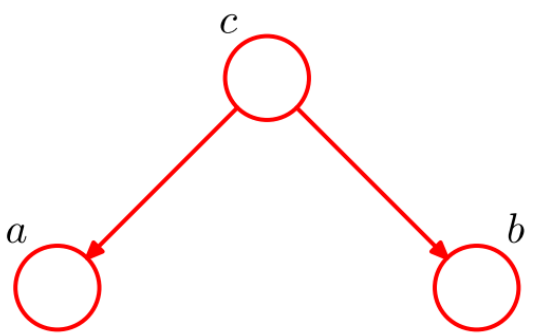
图模型中的条件独立

- 可以直接从图中获取联合分布的条件独立属性

- 例1：tail-to-tail
  - c未被观测到

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

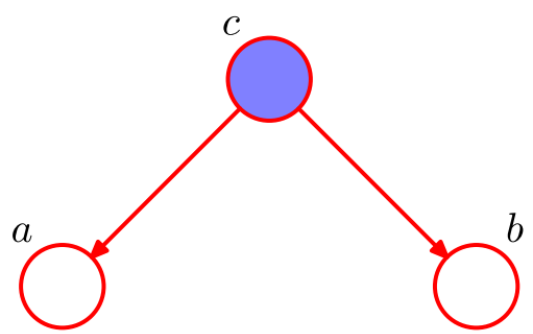
    不条件独立 $a \not\perp\!\!\!\perp b \mid \emptyset$



  - c被观测到

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

    条件独立 $a \perp\!\!\!\perp b \mid c$

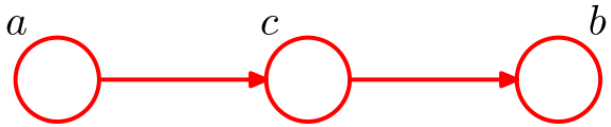


图模型中的条件独立

- 例2：head-to-tail
  - c未被观测到

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

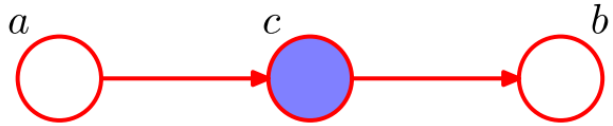
    非条件独立 $a \not\perp\!\!\!\perp b \mid \emptyset$



  - c被观测到

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

    条件独立 $a \perp\!\!\!\perp b \mid c$



图模型中的条件独立

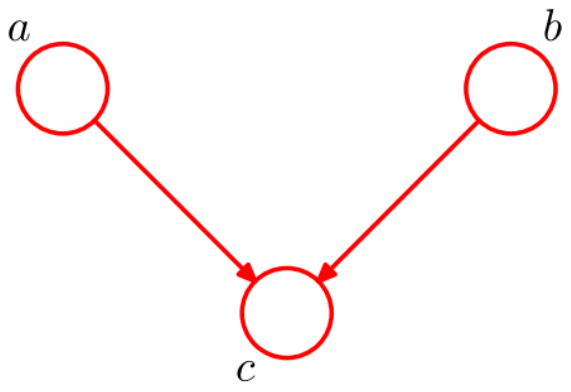
- 例3：head-to-head
  - c未被观测到

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

    两边分别边缘化c

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

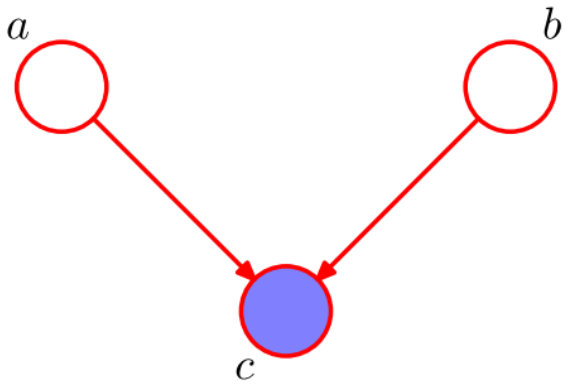
    条件独立 $a \perp\!\!\!\perp b \mid \emptyset$



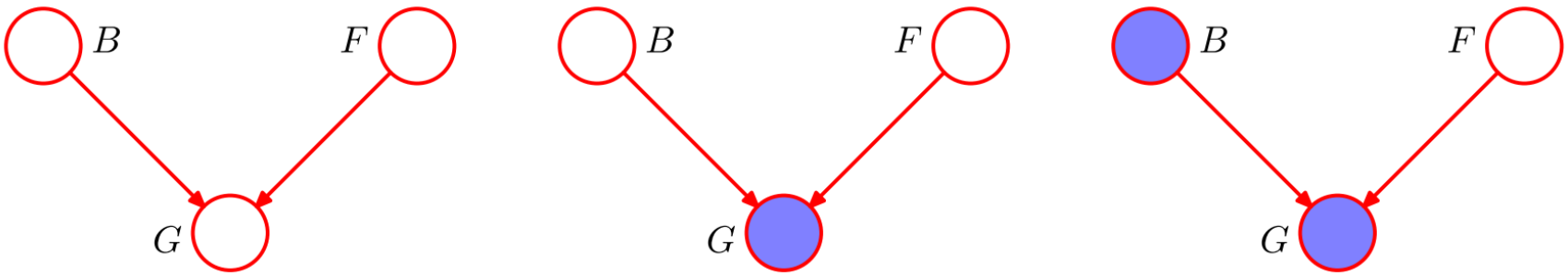
  - c被观测到

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

    非条件独立 $a \not\perp\!\!\!\perp b \mid c$



理解head-to-head的例子



- 变量
  - B: 电池状态，有电(B=1)或没电(B=0)
  - F: 油箱状态，有油(F=1)或没油(F=0)
  - G: 燃油表，有示数(G=1)或读数为空(G=0)

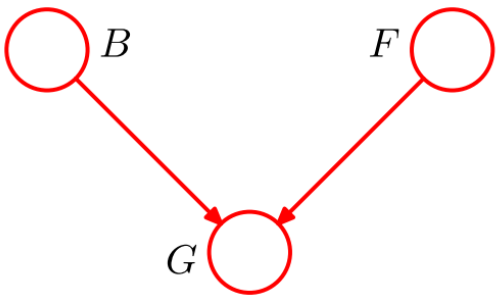
- （条件）概率

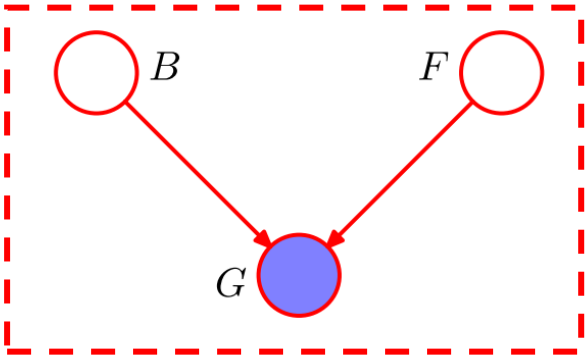
$$p(B=1)=0.9$$
$$p(F=1)=0.9$$

所有剩余概率由概率总和为1的条件确定

$$p(G=1|B=1,F=1)=0.8$$
$$p(G=1|B=1,F=0)=0.2$$
$$p(G=1|B=0,F=1)=0.2$$
$$p(G=1|B=0,F=0)=0.1$$

理解head-to-head的例子





- 如果我们观测到燃油表读数为空，即G=0

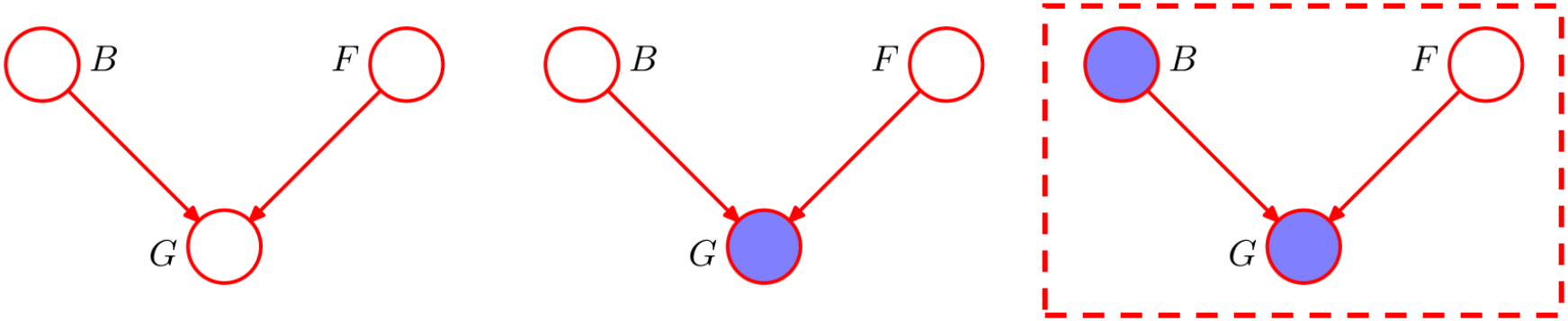
$$p(G=0)=\sum_{B\in\{0,1\}}\sum_{F\in\{0,1\}}p(G=0|B,F)p(B)p(F)=0.315$$

$$p(G=0|F=0)=\sum_{B\in\{0,1\}}p(G=0|B,F=0)p(B)=0.81$$

$$p(F=0|G=0)=\frac{p(G=0|F=0)p(F=0)}{p(G=0)}\simeq 0.257>p(F=0)=0.1$$

因此，观察到油表读数为空，则油箱更可能是空的

理解head-to-head的例子



- 如果我们观测到燃油表读数为空，且电池也没电了，即G=0, B=0

$$p(F=0|G=0)=\frac{p(G=0|F=0)p(F=0)}{p(G=0)}\simeq 0.257>p(F=0)=0.1$$

$$p(F=0|G=0,B=0)=\frac{p(G=0|B=0,F=0)p(F=0)}{\sum_{F\in\{0,1\}}p(G=0|B=0,F)p(F)}\simeq 0.111>p(F=0)=0.1$$

- 由于观测到电池的状态，油箱没油的概率从0.257降到了0.111
- 解释：电池没电可以解释燃油表读数为空的观察结果（explaining away）
- 但是注意，explaining away并不能把油箱没油的概率完全降到先验值（0.111>0.1）

02

有向分离

有向分离(D-separation)

- 考虑一般有向图，其中A，B和C是任意不相交节点集。

- 任何一条路径上的节点满足以下条件，则说这条路径被阻断

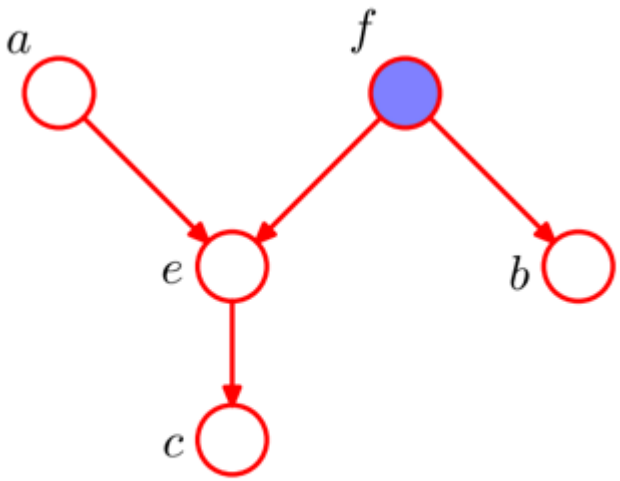
  a) 路径上的箭头在节点head-to-tail或tail-to-tail交汇，而此节点在集合C中

  b) 路径上的箭头在节点head-to-head交汇，而此节点及其后代节点都不在集合C中

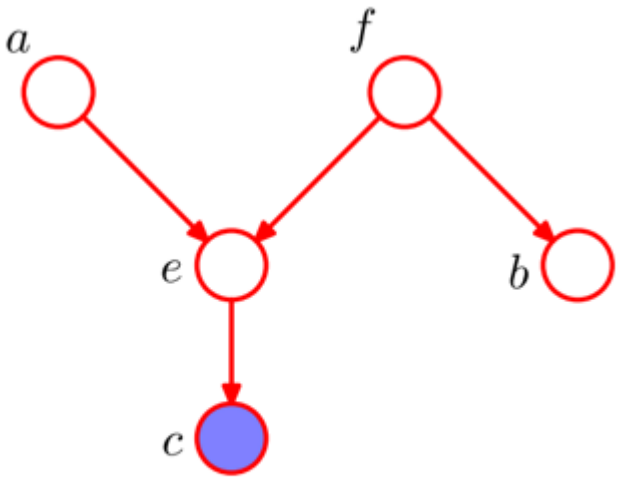
- 如果所有路径都被阻断，则可以说A和B关于C有向分离，并且图中所有变量的联合分布将满足

$$A \perp\!\!\!\perp B \mid C$$

有向分离图示



$a \perp\!\!\!\perp b \mid f$



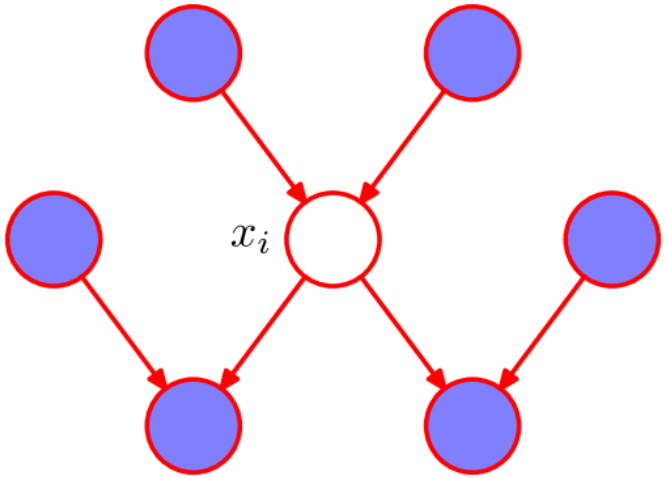
$a \not\perp\!\!\!\perp b \mid c$

- A, B, C满足 $A \perp\!\!\!\perp B \mid C$ ，当且仅当
  a) 路径上的箭头在节点head-to-tail或tail-to-tail交汇，而此节点在集合C中
  b) 路径上的箭头在节点head-to-head交汇，而此节点及其后代节点都不在集合C中

贝叶斯网络中的马尔可夫毯（Markov Blanket）

- 节点x_i的马尔可夫毯包括父母、孩子和孩子的共同父母的集合

- 它具有以下性质：x_i以图中所有剩余变量为条件的条件分布，仅依赖于马尔可夫毯中的变量。



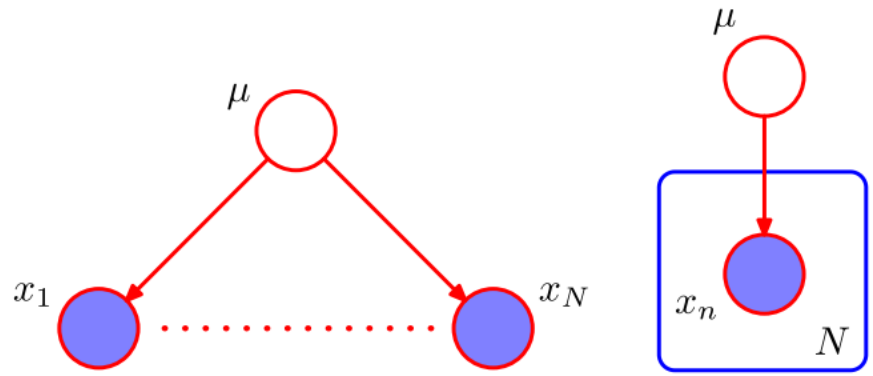
03

独立同分布

独立同分布(i.i.d.)的例子

- 目标：给定**x**观测，推断μ

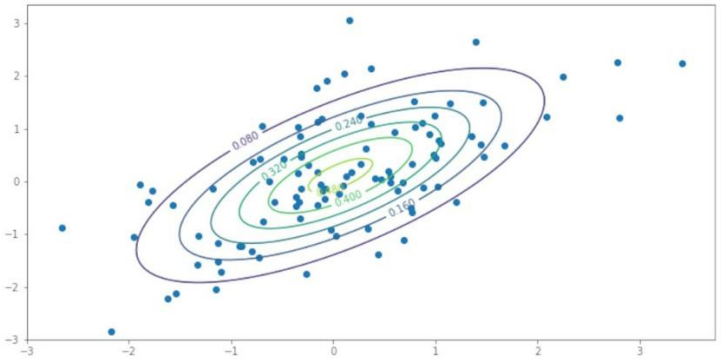
$$p(\mu|D) \propto p(\mu)p(D|\mu)$$
$$= p(\mu)\prod_{i=1}^{N} p(x_i|\mu)$$



- 如果我们对μ求积分，则观察结果通常是相关的

$$p(D) = \int p(D|\mu)p(\mu)d\mu \neq \prod_{i=1}^{N} p(x_i)$$

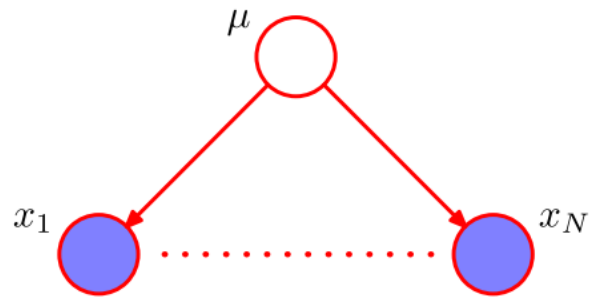
- 我们可以说这些数据样本是联合分布的

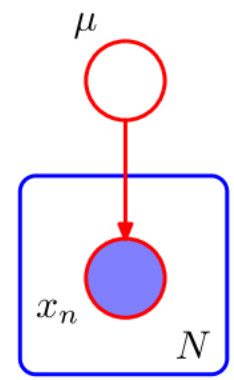


独立同分布(i.i.d.)的例子

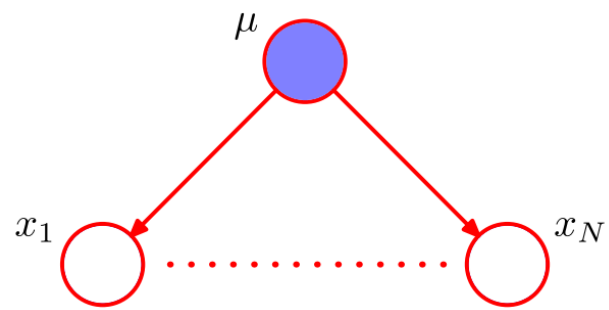
- 目标：给定**x**观测，推断μ

$$p(\mu|D) \propto p(\mu)p(D|\mu)$$
$$= p(\mu)\prod_{i=1}^{N} p(x_i|\mu)$$





- 如果我们以μ为给定条件，考虑观测**x**的联合分布
  - 从x_i到x_j只有单一路径
  - 路径关于μ是tail-to-tail
  - 因此给定观测μ的情况下，路径被阻断



- 数据样本对于模型是条件独立的

朴素贝叶斯分类模型

文本分类的K分类问题

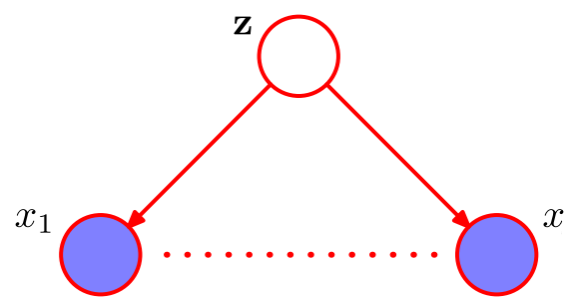
- 每个文档的类$\mathbf{z}$用独热编码的向量表示
  - 多项式先验 $p(\mathbf{z}|\boldsymbol{\mu})$
  - μ_k是C_k类的先验概率

- 每个数据样本(比如一段文本)被表示成一个d维向量$\mathbf{x}$(每个维度表示一个单词)
  - 以$\mathbf{z}$为条件$\mathbf{x}$的生成概率 $p(\mathbf{x}|\mathbf{z})$
  - 朴素贝叶斯的原则是所有x_j条件独立

$$p(\mathbf{x}|\mathbf{z}) = \prod_{j=1}^{d} p(x_j|\mathbf{z})$$

- 类标签推断

$$p(\mathbf{z}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{z})p(\mathbf{z}|\boldsymbol{\mu})$$



多项式朴素贝叶斯

- 每个类y被建模为单词的直方图
  - $y = y(\mathbf{z})$表示**z**向量中1的索引

$$\theta_y = (\theta_{y1}, \theta_{y2}, \cdots, \theta_{yn})$$

- 参数ϑ_y被估计成

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha d}$$

  - N_{yi}表示y类别训练文档中单词i出现的次数
  - N_y表示y类别训练文档的总单词数

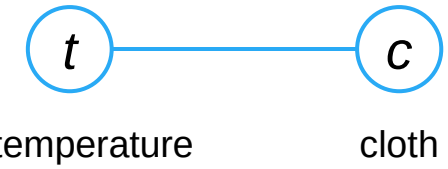
$$p(\mathbf{z}|\mathbf{x}) \propto p(\mathbf{z}|\boldsymbol{\mu})p(\mathbf{x}|\mathbf{z}) = p(\mathbf{z}|\boldsymbol{\mu}) \prod_{i=1}^{d} \theta_{y(\mathbf{z})i}$$

马尔可夫网络简介

张伟楠 – 上海交通大学

马尔可夫随机场（Markov Random Field）

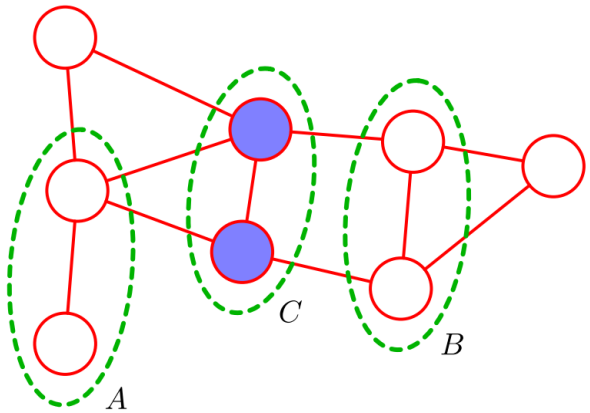
- 无向网络，也称马尔可夫网络（Markov network）



$$p(t,c) = \frac{e^{\phi(t,c)}}{\sum_{t',c'} e^{\phi(t',c')}}$$

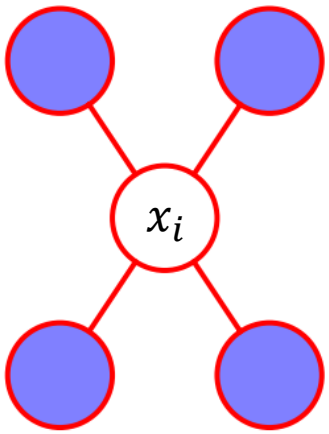
- 与贝叶斯网络相比，确定马尔可夫网络中的条件独立性更为直接：
  - 如果连接A和B中任何节点的所有路径都被C中的节点阻塞，那么

$$A \perp\!\!\!\perp B \mid C$$



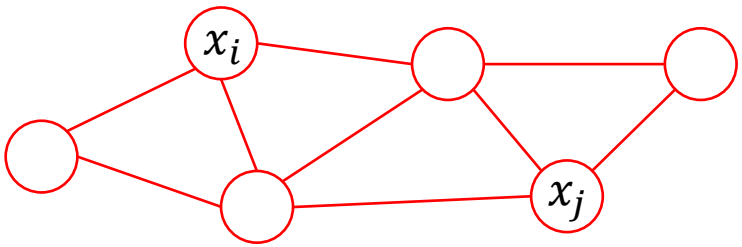
马尔可夫网络中的马尔可夫毯(Markov Blanket)

- 对于无向图，节点x_i的马尔可夫毯由一组相邻节点组成。



- 它具有以下特性：以图中所有剩余变量为条件x_i的条件分布，仅依赖于马尔可夫毯中的变量。

马尔可夫的条件独立性

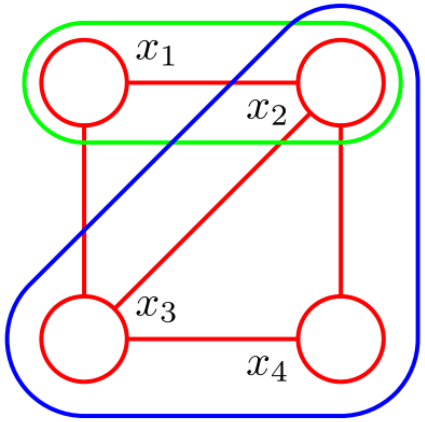


- 考虑两个节点x_i和x_j没有直接连接，当给定图中的其他节点，它们必须是条件独立的，

$$p(x_i, x_j | \boldsymbol{x}_{\setminus\{i,j\}}) = p(x_i | \boldsymbol{x}_{\setminus\{i,j\}}) p(x_j | \boldsymbol{x}_{\setminus\{i,j\}})$$

- 因此，联合分布的因子分解必须使得x_i和x_j不出现在相同的因子(factor)中

马尔可夫网络中的团



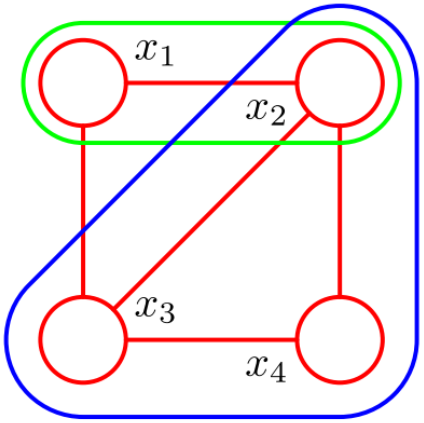
- 团：图中节点的一个子集，节点完全连接在一起
- 一个有四个节点的马尔可夫网络$\{x_1, x_2, x_3, x_4\}$
  - 5个双节点团$\{x_1, x_2\}, \{x_2, x_3\}, \{x_3, x_4\}, \{x_2, x_4\}, \{x_1, x_3\}$
  - 两个最大团$\{x_1, x_2, x_3\}, \{x_2, x_3, x_4\}$
  - 注意$\{x_1, x_2, x_3, x_4\}$不是一个团

联合分布分解

- 将联合分布分解中的因子定义为团中变量的函数
- 设C表示一个团，其中的变量集合为x_C

$$p(\boldsymbol{x}) = \frac{1}{Z}\prod_C \psi_C(\boldsymbol{x}_C)$$

势函数



$$p(\boldsymbol{x}) = \frac{1}{Z}\psi_{\{2,3,4\}}(x_2, x_3, x_4)\psi_{\{1,2,3\}}(x_1, x_2, x_3)$$

- 数量Z，也称作配分函数(partition function)，是一个标准化因子

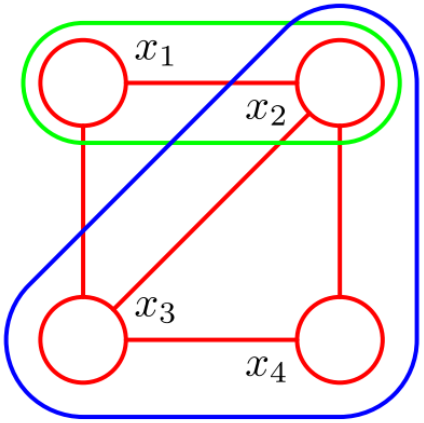
$$Z = \sum_{\boldsymbol{x}}\prod_C \psi_C(\boldsymbol{x}_C)$$

联合分布分解

- 将联合分布分解中的因子定义为团中变量的函数
- 设C表示一个团，其中的变量集合为x_C

$$p(\boldsymbol{x}) = \frac{1}{Z}\prod_C \psi_C(\boldsymbol{x}_C)$$

势函数



- 势函数需要满足$\psi_C(\boldsymbol{x}_C) \geq 0$去确保概率为非负数
- 可以用领域知识定义势函数

势能的能量函数

- 如果我们将势函数定义为严格正值，即

$$\psi_C(\boldsymbol{x}_C) > 0$$

- 将势函数方便地用指数形式表示

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

- $E(\boldsymbol{x}_C)$被称作能量函数
- 有了这样的指数表示，分布$p(\boldsymbol{x})$被称作玻尔兹曼分布

$$p(\boldsymbol{x}) = \frac{1}{Z}\prod_C \psi_C(\boldsymbol{x}_C) = \frac{1}{Z}\exp\{-\sum_C E(\boldsymbol{x}_C)\}$$

玻尔兹曼分布（Boltzmann Distribution）

- 玻尔兹曼分布是系统中各种可能状态下的粒子的概率分布、概率测量或频率分布

$$p(s) = \frac{e^{-E(s)/kT}}{\sum_{s'} e^{-E(s')/kT}}$$

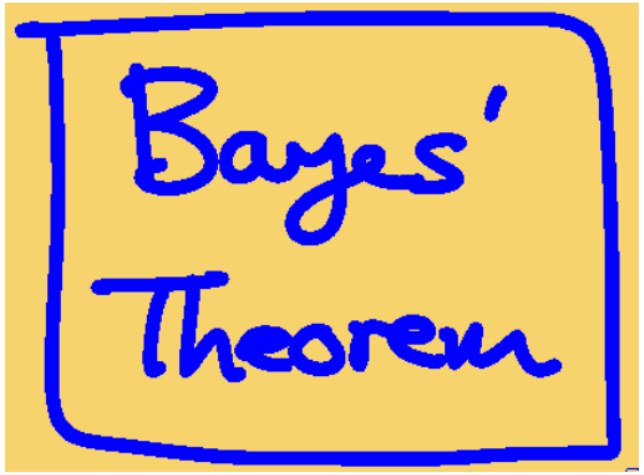
  - s表示特定的状态
  - $E(s)$是状态能量
  - $k = 1.381 \times 10^{-23}\ J/K$是玻尔兹曼常数
  - T是热力学温度
- 低能状态更稳定，即具有更高的概率存在

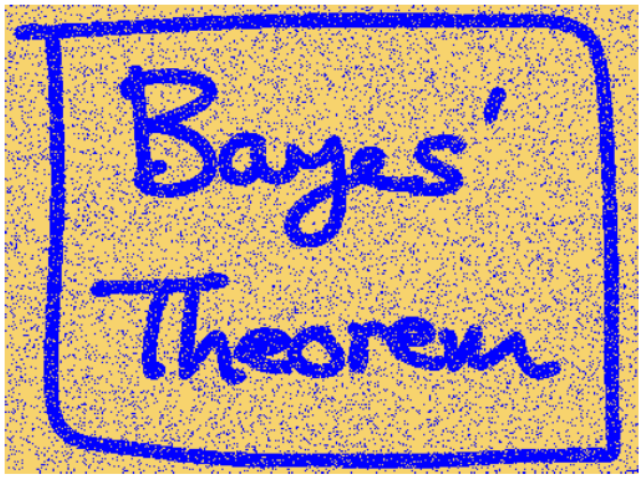
马尔可夫网络应用示例

张伟楠 – 上海交通大学

马尔可夫网络例子：图像去噪

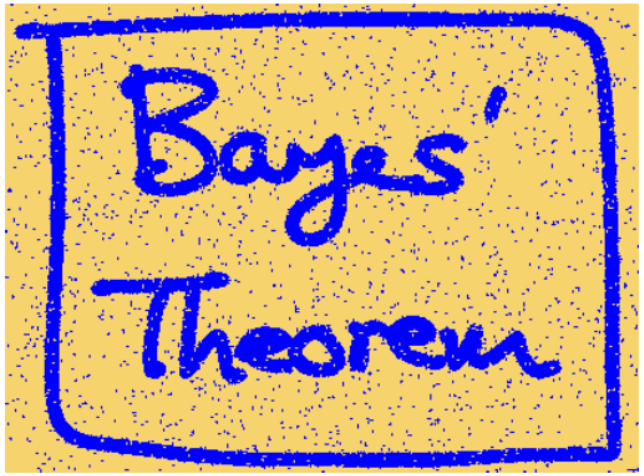
源图像

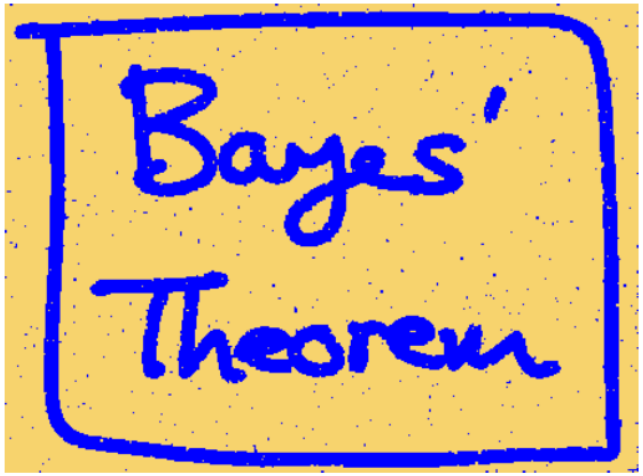




损坏
图像

由ICM
去噪





按图
切割
去噪

马尔可夫网络例子：图像去噪

- 噪点图像由二进制像素值数组描述

$$y_i \in \{-1, +1\}, i = 1, \dots, d \quad \text{所有像素}$$

- 假设真实的无噪点图像

$$x_i \in \{-1, +1\}, i = 1, \dots, d$$

- 噪点生成：以比较小的概率（例如10%）随机翻转一些像素的符号

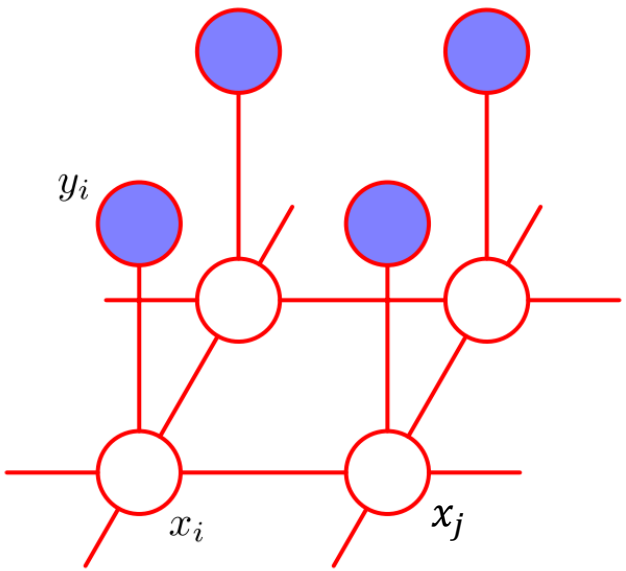
模型假设

- x_i和y_i之间有很强的相关性
- 相邻像素x_i和x_j之间存在强相关性

用于图像去噪的马尔可夫网络

模型假设

- x_i和y_i之间有很强的相关性
- 相邻像素x_i和x_j之间存在强相关性



模型

- 对于团$\{x_i, y_i\}$

$$E(\{x_i, y_i\}) = -\eta x_i y_i$$

- 对于团$\{x_i, x_j\}$

$$E(\{x_i, x_j\}) = -\beta x_i x_j$$

- 更进一步，对每个$\{x_i\}$

$$E(\{x_i\}) = h x_i$$

- 完整的能量函数

$$E(\boldsymbol{x}, \boldsymbol{y}) = h \sum_i x_i - \beta \sum_{i,j} x_i x_j - \eta \sum_i x_i y_i \qquad p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \exp\{-E(\boldsymbol{x}, \boldsymbol{y})\}$$

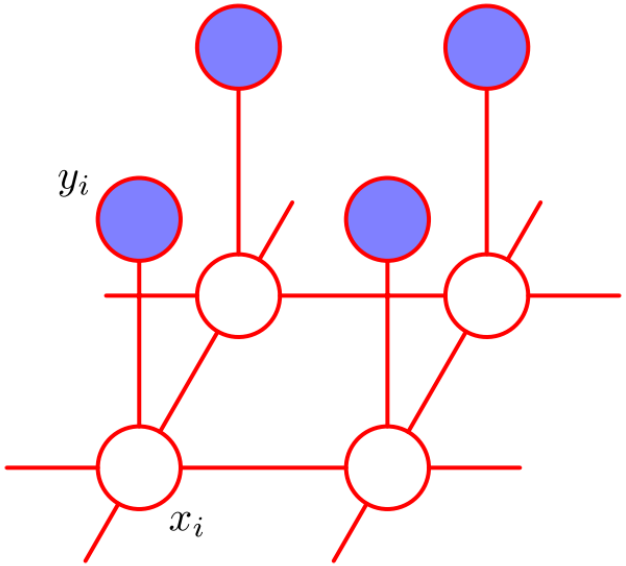
解决方法：迭代条件模式（ICM）

目标

$$\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}) = \max_{\boldsymbol{x}} p(\boldsymbol{x},\boldsymbol{y})$$ 在$\boldsymbol{y}$被观察到的条件下

想法

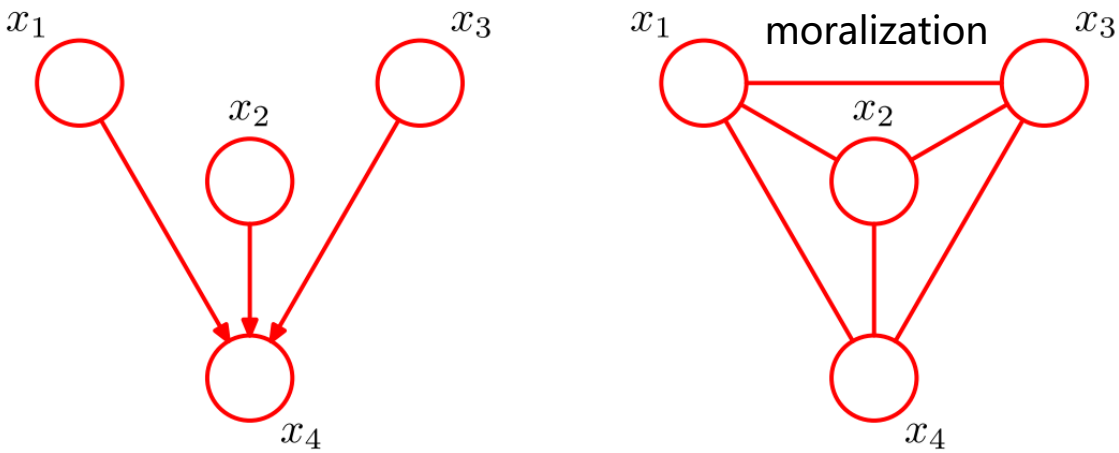
- 坐标梯度上升
  - 对每个节点x_j，检查哪个$x_j = +1$或-1导致比较低的$E(\boldsymbol{x},\boldsymbol{y})$
  - 令$\beta = 1.0$，$\eta = 2.1$，$h = 0$



- 能量函数

$$E(\boldsymbol{x},\boldsymbol{y}) = h\sum_i x_i - \beta\sum_{i,j} x_i x_j - \eta\sum_i x_i y_i \qquad p(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{Z}\exp\{-E(\boldsymbol{x},\boldsymbol{y})\}$$

有向图 vs. 无向图



- 将有向图转换成无向图
  - 有向图

$$p(\boldsymbol{x}) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)$$

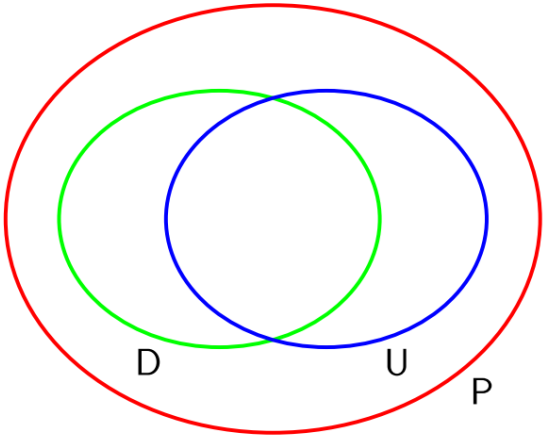
  - 无向图

$$p(\boldsymbol{x}) = \frac{1}{Z}\psi_{1,2,3,4}(x_1, x_2, x_3, x_4)$$

Moralization：父节点联姻

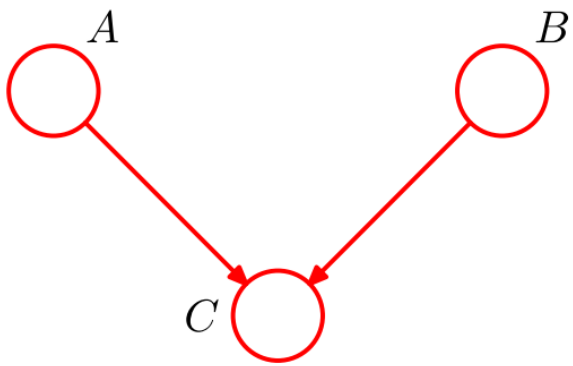
有向图 vs. 无向图

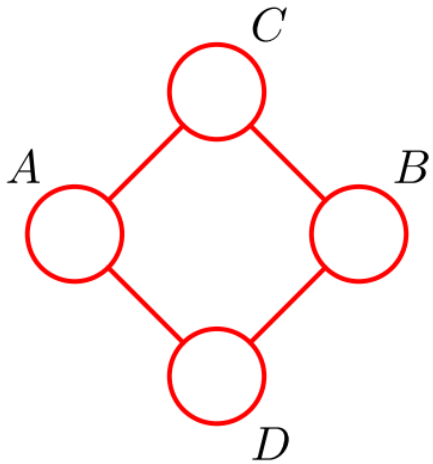
- 尽管每个有向图都可以转换为无向图
  - 一种强制解决方案是使用全连接的无向图
- 有向图和无向图可以表达不同的条件独立属性



P：所有可能的分布
D / U：可以由有向/无向图表示的分布

有向图 vs. 无向图





一个有向图，其条件独立属性不能用同样三个变量组成的无向图表示

- 有向图

$$A \perp\!\!\!\perp B \mid \emptyset$$

$$A \not\perp\!\!\!\perp B \mid C$$

一个无向图，其条件独立属性不能用同样四个变量组成的有向图表示

- 无向图

$$A \not\perp\!\!\!\perp B \mid \emptyset,\ C \perp\!\!\!\perp D \mid A \cup B$$

$$A \perp\!\!\!\perp B \mid C \cup D$$

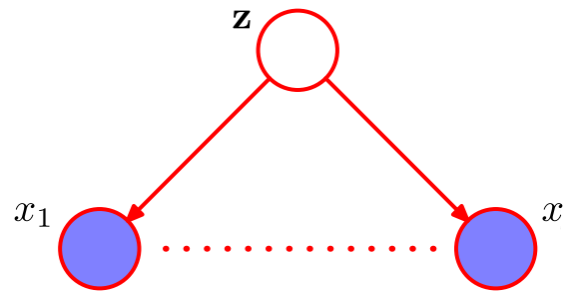
链式模型推断

张伟楠 – 上海交通大学

变量推断和参数估计

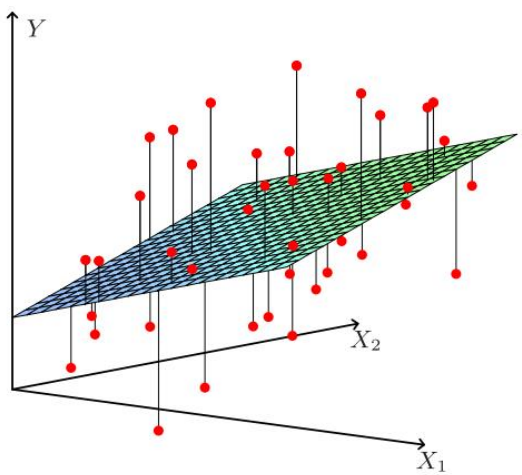
- 随机变量推断
  - 根据先前和观察到的数据推断随机变量的后验分布

    $$p(\boldsymbol{z}|\boldsymbol{x}) \propto p(\boldsymbol{z}|\boldsymbol{\mu})p(\boldsymbol{x}|\boldsymbol{z})$$



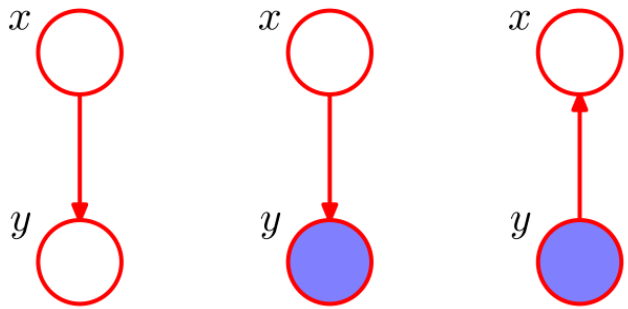
- 参数估计
  - 找到目标函数的最佳参数值，例如最小损失或最大似然

    $$\hat{\theta} = \underset{\theta}{argmin}\mathcal{L}(D;\theta)$$



Tzikas, Dimitris G., Aristidis C. Likas, and Nikolaos P. Galatsanos. "The variational approximation for Bayesian inference." IEEE Signal Processing Magazine 25.6 (2008): 131-146.

推断的基本案例



- 两个随机变量的联合分布x和y

$$p(x,y) = p(x)p(y|x)$$

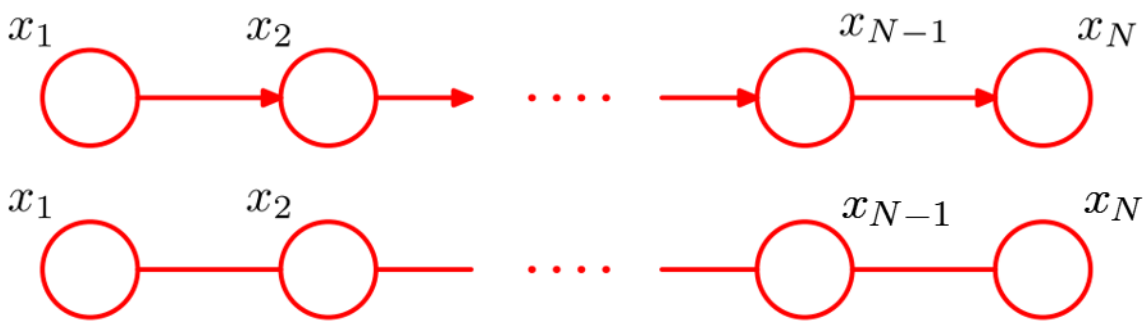
- y的边缘分布

$$p(y) = \sum_{x'} p(x')p(y|x')$$

- 逆条件分布

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

在链上做推断



- 联合分布

$$p(\boldsymbol{x}) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\cdots\psi_{N-1,N}(x_{N-1}, x_N)$$

- 离散变量设置
  - N个节点表示各自具有K个状态的离散变量
  - 每个势函数$\psi_{N-1,N}(x_{N-1}, x_N)$包括一个$K \times K$状态表
  - 因此，联合分布具有$(N-1)K^2$个参数

计算边缘分布

- 寻找边缘分布$p(x)$的推断问题

$$p(x_n) = \sum_{x_1} \cdots \sum_{x_{n-1}} \sum_{x_{n+1}} \cdots \sum_{x_N} p(\boldsymbol{x})$$

- 暴力求解
  - 对K^{N-1}个值加和，会导致指数级复杂度$O(K^{N-1})$
- 一个高效的动态规划解决方案
  - 利用条件独立性

$$p(\boldsymbol{x}) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\cdots\psi_{N-1,N}(x_{N-1}, x_N)$$

通用加乘算法 $ab + ac = a(b + c)$

计算边缘分布的动态规划

$$p(\boldsymbol{x}) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\cdots\psi_{N-1,N}(x_{N-1}, x_N)$$

- 条件独立性
  - 只有势能函数$\psi_{N-1,N}(x_{N-1}, x_N)$依赖于$x_N$

$$\begin{aligned}
p(x_n) &= \sum_{\mathrm{x}_1}\cdots\sum_{\mathrm{x}_{n-1}}\sum_{\mathrm{x}_{n+1}}\cdots\sum_{\mathrm{x}_N} p(\mathrm{x}) \\
&= \frac{1}{Z}\sum_{\mathrm{x}_1}\cdots\sum_{\mathrm{x}_{n-1}}\sum_{\mathrm{x}_{n+1}}\cdots\sum_{x_N}\psi_{1,2}(x_1, x_2)\cdots\psi_{N-1,N}(x_{N-1}, x_N) \\
&= \frac{1}{Z}\sum_{\mathrm{x}_1}\cdots\sum_{\mathrm{x}_{n-1}}\sum_{\mathrm{x}_{n+1}}\cdots\sum_{x_{N-1}}\psi_{1,2}(x_1, x_2)\cdots\psi_{N-2,N-1}(x_{N-2}, x_{N-1})\sum_{x_N}\psi_{N-1,N}(x_{N-1}, x_N)
\end{aligned}$$

通用加乘算法 $ab + ac = a(b + c)$

计算边缘分布的动态规划

$$p(\boldsymbol{x}) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\cdots\psi_{N-1,N}(x_{N-1}, x_N)$$

- 条件独立性
  - 只有势能函数$\psi_{1,2}(x_1, x_2)$依赖于x_1

$$\begin{aligned}
p(x_n) &= \sum_{\mathrm{x}_1}\cdots\sum_{\mathrm{x}_{n-1}}\sum_{\mathrm{x}_{n+1}}\cdots\sum_{\mathrm{x}_N} p(\mathrm{x}) \\
&= \frac{1}{Z}\sum_{\mathrm{x}_N}\cdots\sum_{\mathrm{x}_{n+1}}\sum_{\mathrm{x}_{n-1}}\cdots\sum_{x_1}\psi_{N-1,N}(x_{N-1}, x_N)\cdots\psi_{1,2}(x_1, x_2) \\
&= \frac{1}{Z}\sum_{\mathrm{x}_N}\cdots\sum_{\mathrm{x}_{n+1}}\sum_{\mathrm{x}_{n-1}}\cdots\sum_{x_2}\psi_{N-1,N}(x_{N-1}, x_N)\cdots\psi_{2,3}(x_2, x_3)\sum_{x_1}\psi_{1,2}(x_1, x_2)
\end{aligned}$$

计算边缘分布的动态规划

$$p(\boldsymbol{x}) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\cdots\psi_{N-1,N}(x_{N-1}, x_N)$$

- 条件独立性
  - 势能$\psi_{1,2}(x_1, x_2)$是唯一取决于x_1的因子

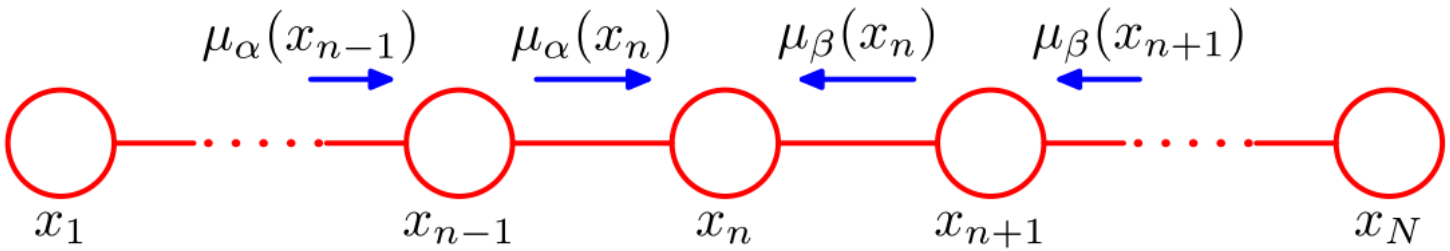
$$p(x_n) = \frac{1}{Z}\underbrace{\left[\sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n)\cdots\left[\sum_{x_2} \psi_{2,3}(x_2, x_3)\left[\sum_{x_1} \psi_{1,2}(x_1, x_2)\right]\right]\cdots\right]}_{\mu_\alpha(x_n)}$$

$$\underbrace{\left[\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1})\cdots\left[\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)\right]\cdots\right]}_{\mu_\beta(x_n)}$$

复杂度$O(NK^2)$

解释：消息传递

- 在图上传递局部消息



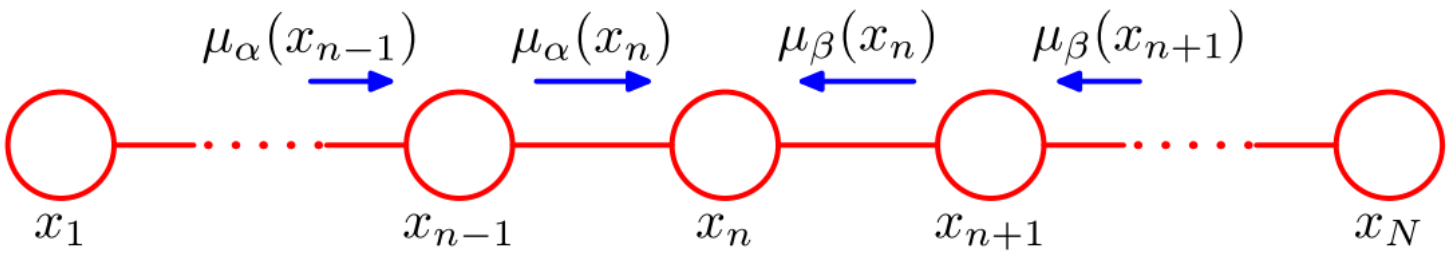
$$p(x_n) = \frac{1}{Z}\underbrace{\left[\sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n) \cdots \left[\sum_{x_2} \psi_{2,3}(x_2, x_3) \left[\sum_{x_1} \psi_{1,2}(x_1, x_2)\right]\right] \cdots\right]}_{\mu_\alpha(x_n)}$$

$$\underbrace{\left[\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) \cdots \left[\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)\right] \cdots\right]}_{\mu_\beta(x_n)}$$

$$p(x_n) = \frac{1}{Z}\mu_\alpha(x_n)\mu_\beta(x_n)$$

解释：消息传递

- 在图上传递局部消息



$$p(x_n) = \frac{1}{Z}\mu_\alpha(x_n)\mu_\beta(x_n)$$

- 信息递归传递

$$\mu_\alpha(x_n) = \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n)\left[\sum_{x_{n-2}} \psi_{n-2,n-1}(x_{n-2}, x_{n-1}) \sum_{x_{n-3}} \cdots\right]$$

$$= \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n)\mu_\alpha(x_{n-1})$$

- 递归开始

$$\mu_\alpha(x_2) = \sum_{x_1} \psi_{1,2}(x_1, x_2)$$

树图模型推断

张伟楠 – 上海交通大学

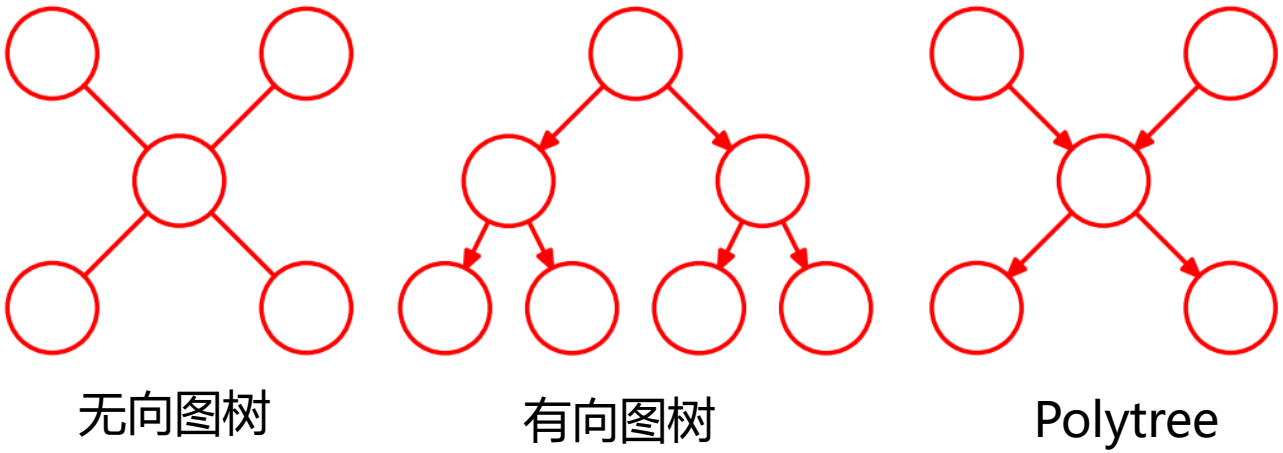
目录

Contents

01

树图模型

树图模型



- **无向图树**：任何节点对之间只有一条路径的图
- **有向图树**：有一个被称作根的节点没有父节点，所有其他节点都有一个父节点
  - 因此，联姻(Moralization)步骤不会增加任何连边
- **Polytree**：有向图中具有多个父节点的节点，但在任意两个节点之间仍然只有一条路径（忽略箭头的方向）
- 在介绍推断算法之前，先讨论一般形式：因子图

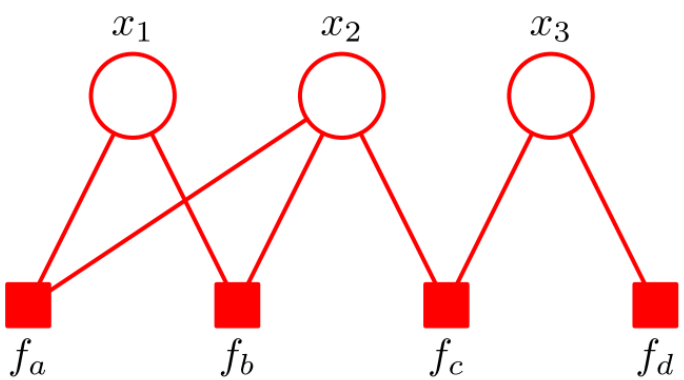
02

因子图

因子图（Factor Graph）

- 观察：有向图和无向图都允许将几个变量的全局函数表示为这些变量子集上的因子的乘积
- 因子图通过为因子引入额外的节点来进行显式分解

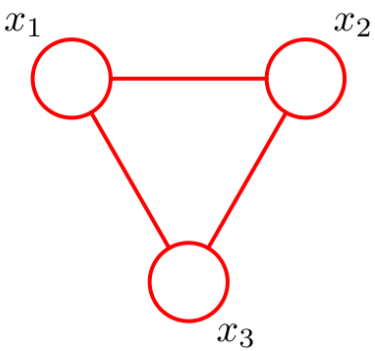
$$p(\boldsymbol{x}) = \prod_s f_s(\boldsymbol{x}_s)$$



因子图是二部图

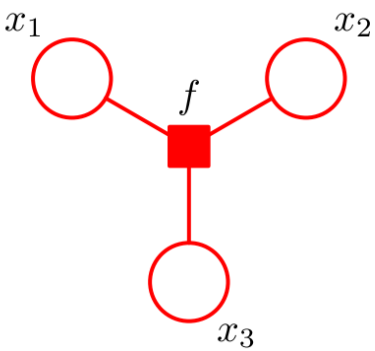
因子图

- 无向图变成因子图



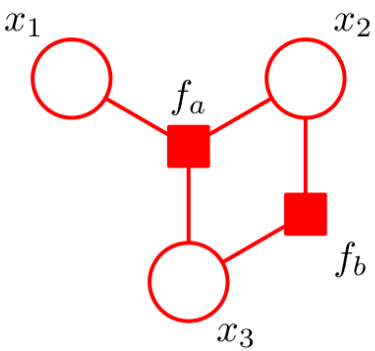
只有单个
团势能的
无向图

$\psi(x_1, x_2, x_3)$



表示因子的
相同分布的
因子图

$f(x_1, x_2, x_3) = \psi(x_1, x_2, x_3)$

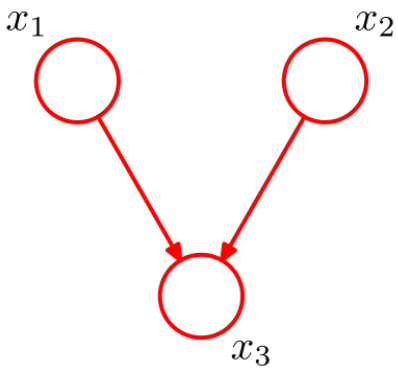


另一个表示
相同分布的
因子图

$$f_a(x_1, x_2, x_3) f_b(x_2, x_3) = \psi(x_1, x_2, x_3)$$

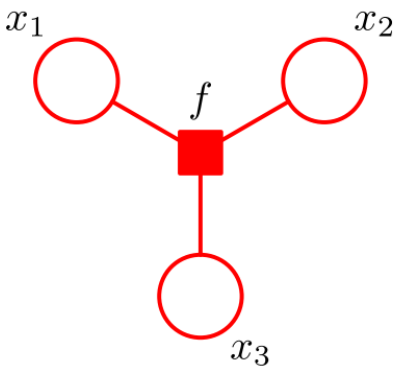
因子图

- 有向图变成因子图



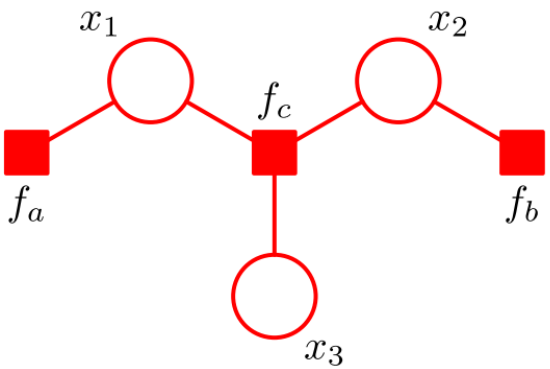
带分解的
有向图

$$p(x_1)p(x_2)p(x_3|x_1,x_2)$$



表示因子的
相同分布的
因子图

$$f(x_1,x_2,x_3) = p(x_1)p(x_2)\,p(x_3|x_1,x_2)$$



另一个表示
相同分布的
因子图

$$f_a(x_1) = p(x_1)$$
$$f_b(x_2) = p(x_2)$$
$$f_c(x_1,x_2,x_3) = p(x_3|x_1,x_2)$$

03

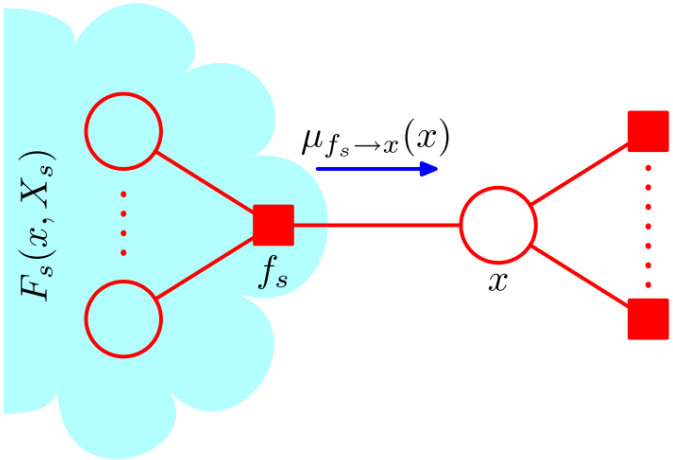
加乘算法

在树上推断：加乘

- 考虑因子图树上特定变量x的边缘分布

$$p(x) = \sum_{\boldsymbol{x}\setminus x} p(\boldsymbol{x}) \qquad\qquad p(x) = \prod_{s\in \mathrm{ne}(x)} \sum_{X_s} F_s(x, X_s)$$

- $\mathrm{ne}(x)$：x的邻居因子集
- X_s：通过因子节点连接到变量节点x的子树中的所有变量的集合
- $F_s(x, X_s)$：与因子f_s相关的集合内所有因子的积

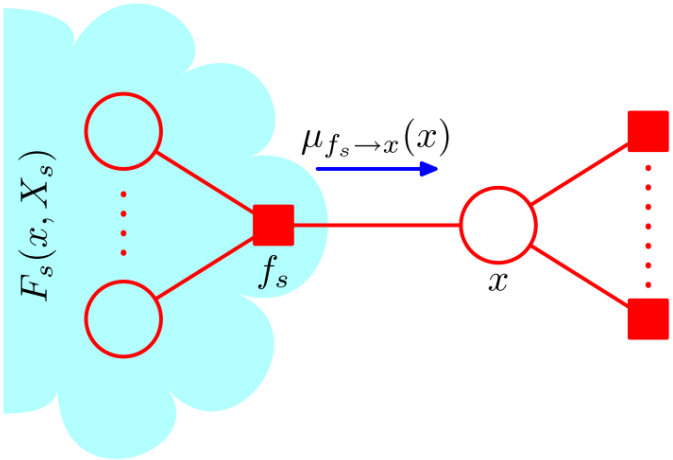


信息传递

- 考虑因子图树上特定变量x的边缘概率

$$p(x) = \prod_{s \in \mathrm{ne}(x)} \left[\sum_{X_s} F_s(x, X_s) \right]$$
$$\equiv \prod_{s \in \mathrm{ne}(x)} \mu_{f_s \to x}(x)$$

从因子节点f_s到变量节点x的信息



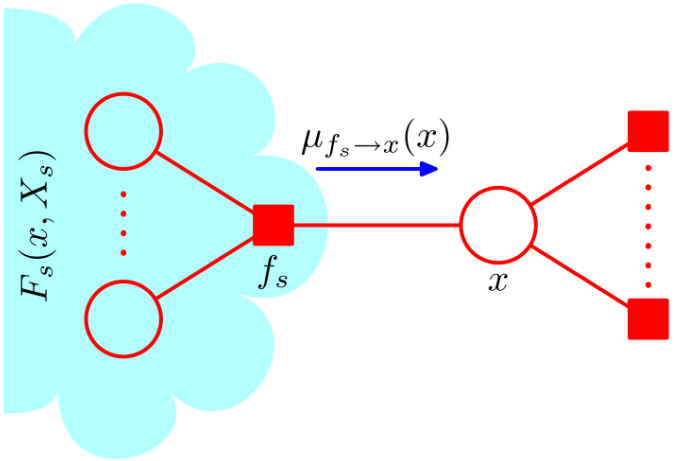
信息传递迭代

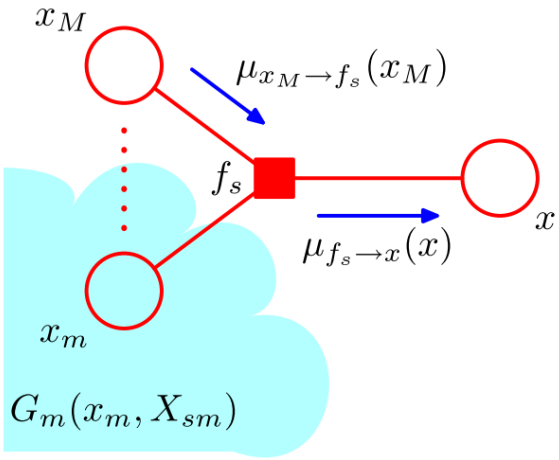
- 将$\{x, x_1, \dots, x_M\}$表示为因子f_s所依赖的变量集，我们可以计算

$$F_s(x, X_s) = f_s(x, x_1, \dots, x_M) G_1(x_1, X_{s1}) \cdots G_M(x_M, X_{sM})$$

$$\mu_{f_s \to x}(x) = \sum_{x_1} \cdots \sum_{x_M} f_s(x, x_1, \dots, x_M) \prod_{m \in \mathrm{ne}(f_s) \setminus x} \left[\sum_{X_{sm}} G_m(x_m, X_{sm}) \right]$$

$$= \sum_{x_1} \cdots \sum_{x_M} f_s(x, x_1, \dots, x_M) \prod_{m \in \mathrm{ne}(f_s) \setminus x} \mu_{x_m \to f_s}(x_m)$$



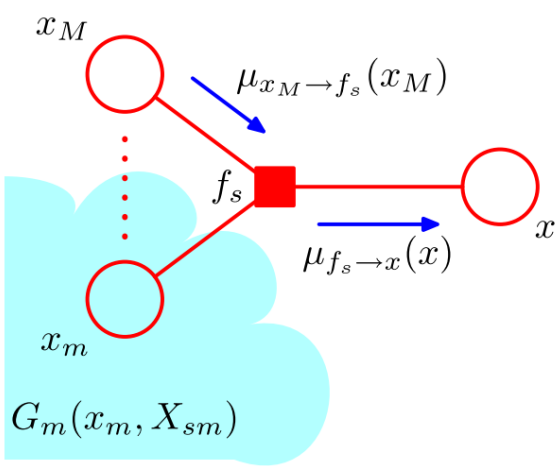


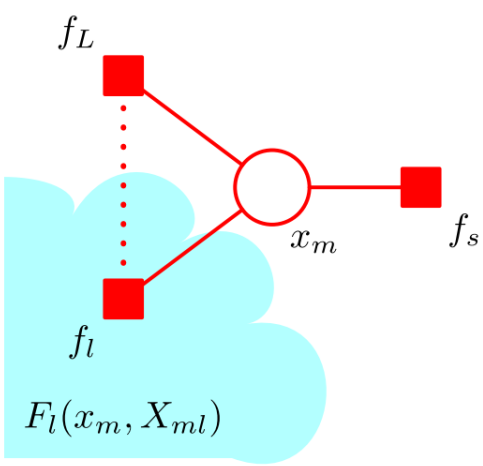
信息传递迭代

- 将$\{x, x_1, \dots, x_M\}$表示为因子f_s所依赖的变量集，我们可以计算

$$F_s(x, X_s) = f_s(x, x_1, \dots, x_M) G_1(x_1, X_{s1}) \cdots G_M(x_M, X_{sM})$$

$$G_m(x_m, X_{sm}) = \prod_{l \in \mathrm{ne}(x_m) \setminus f_s} F_l(x_m, X_{ml})$$





两种信息类型

- 从因子节点到变量节点的信息

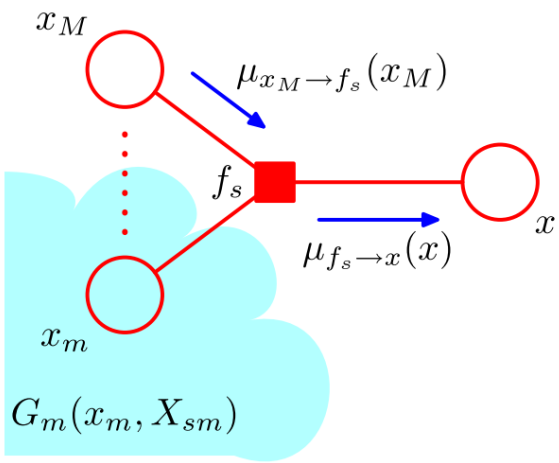
$$\mu_{f_s \to x}(x) = \sum_{X_s} F_s(x, X_s)$$

$$F_s(x, X_s) = f_s(x, x_1, \dots, x_M) G_1(x_1, X_{s1}) \cdots G_M(x_M, X_{sM})$$

- 从变量节点到因子节点的信息

$$\mu_{x_m \to f_s}(x_m) = \sum_{X_{sm}} G_m(x_m, X_{sm})$$

$$G_m(x_m, X_{sm}) = \prod_{l \in \mathrm{ne}(x_m) \setminus f_s} F_l(x_m, X_{ml})$$



两种信息类型

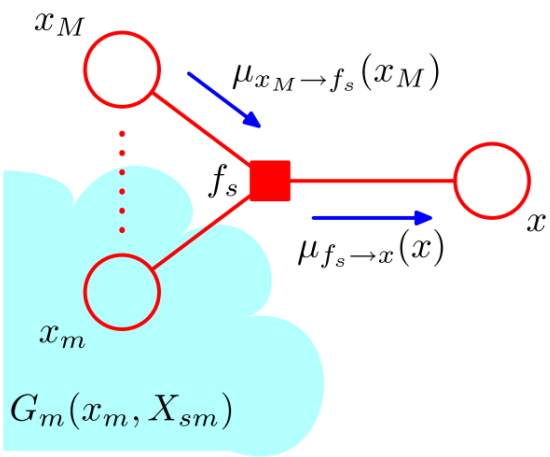
- 两种类型信息的关系

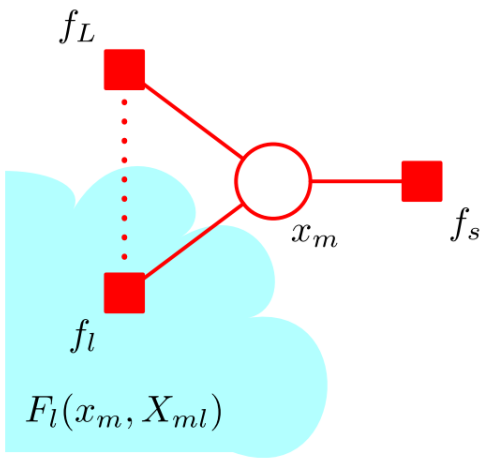
$$\mu_{x_m \to f_s}(x_m) = \sum_{X_{sm}} G_m(x_m, X_{sm})$$

$$= \sum_{X_{sm}} \prod_{l \in \mathrm{ne}(x_m) \setminus f_s} F_l(x_m, X_{ml})$$

(树结构) $$= \prod_{l \in \mathrm{ne}(x_m) \setminus f_s} \left[\sum_{X_{ml}} F_l(x_m, X_{ml}) \right]$$

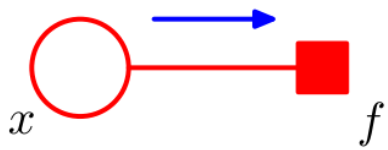
$$= \prod_{l \in \mathrm{ne}(x_m) \setminus f_s} \mu_{f_l \to x_m}(x_m)$$



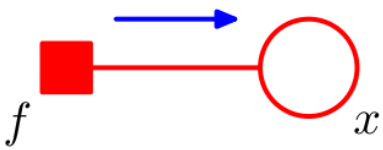


递归的起始过程

$$\mu_{x\to f}(x) = 1$$



$$\mu_{f\to x}(x) = f(x)$$



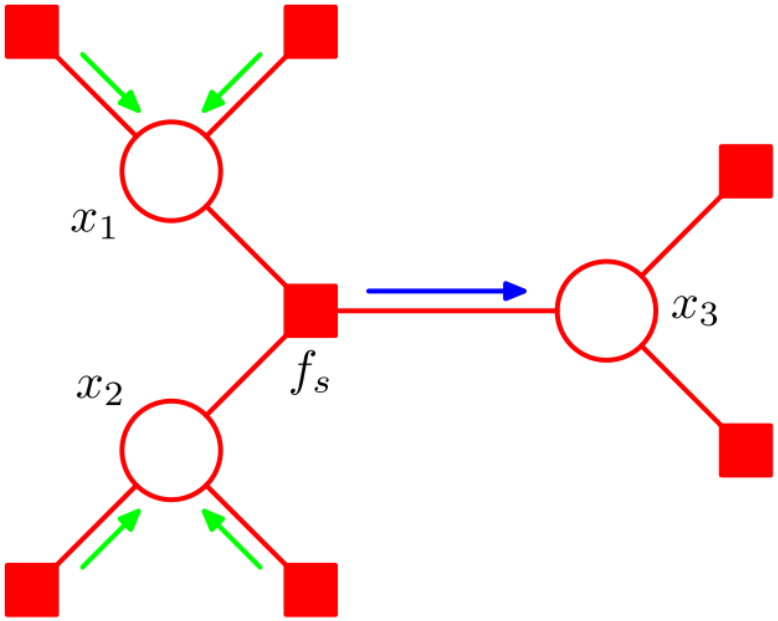
- 从变量节点到因子节点的信息

$$\mu_{x_m\to f_s}(x_m) = \sum_{X_{sm}} G_m(x_m, X_{sm}) \qquad G_m(x_m, X_{sm}) = \prod_{l\in \mathrm{ne}(x_m)\setminus f_s} F_l(x_m, X_{ml})$$

- 从因子节点到变量节点的信息

$$\mu_{f_s\to x}(x) = \sum_{X_s} F_s(x, X_s) \qquad F_s(x, X_s) = f_s(x, x_1, \dots, x_M) G_1(x_1, X_{s1}) \cdots G_M(x_M, X_{sM})$$

因子变量的边缘分布

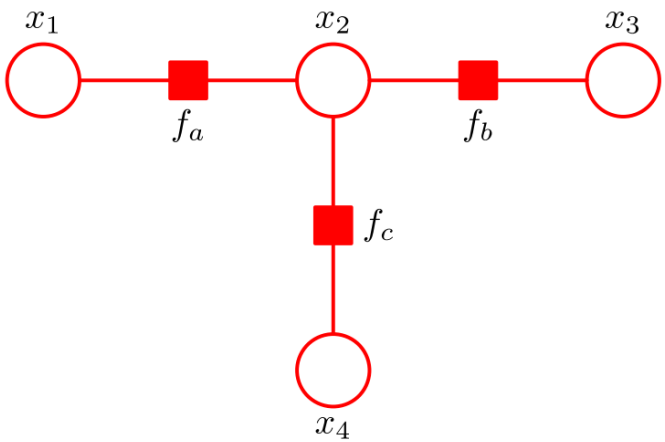


$$p(\boldsymbol{x}_s) = f_s(\boldsymbol{x}_s) \prod_{i \in \mathrm{ne}(f_s)} \mu_{x_i \to f_s}(x_i)$$

一个实际例子

- 非归一化联合分布

$$\tilde{p}(\boldsymbol{x}) = f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4)$$



- 令节点x_3作为根结点，信息传递如下：

$$\mu_{x_1 \to f_a}(x_1) = 1$$

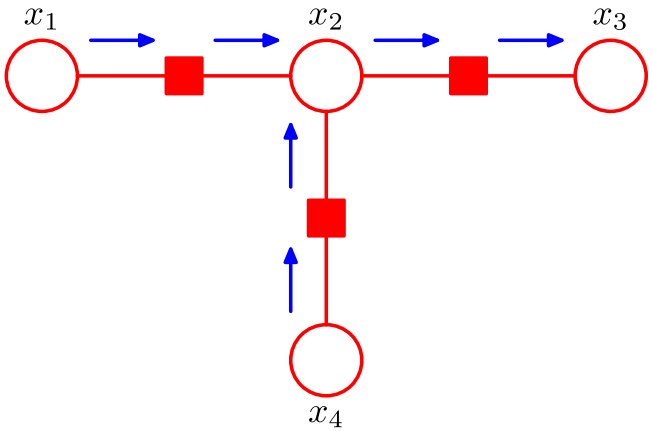
$$\mu_{f_a \to x_2}(x_2) = \sum_{x_1} f_a(x_1, x_2)$$

$$\mu_{x_4 \to f_c}(x_4) = 1$$

$$\mu_{f_c \to x_2}(x_2) = \sum_{x_4} f_c(x_2, x_4)$$

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

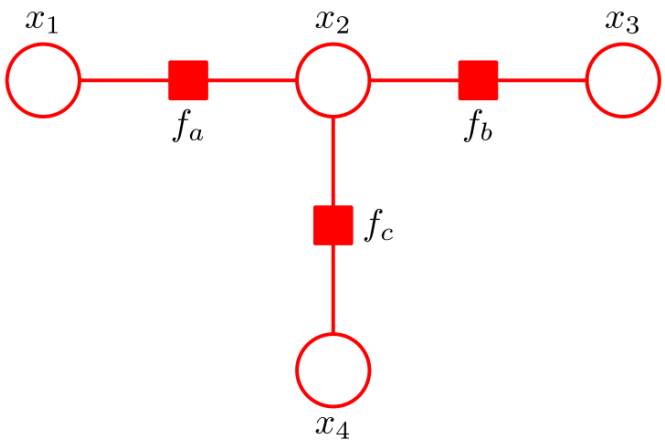
$$\mu_{f_b \to x_3}(x_3) = \sum_{x_2} f_b(x_2, x_3) \mu_{x_2 \to f_b}(x_2)$$



一个实际例子

- 非归一化联合分布

$$\tilde{p}(\boldsymbol{x}) = f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4)$$



- 令节点x_3作为根结点，信息传递如下：

$$\mu_{x_3 \to f_b}(x_3) = 1$$

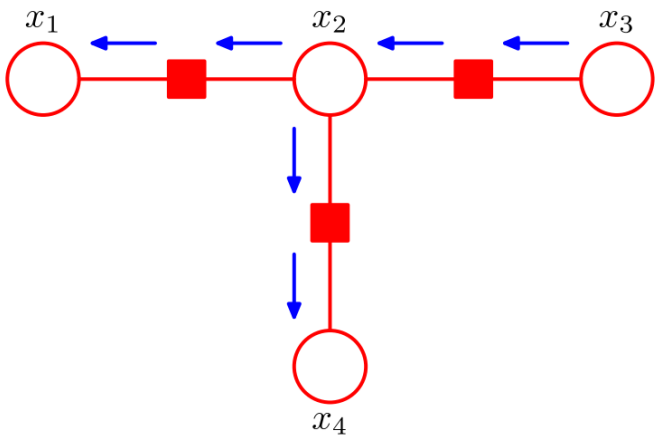
$$\mu_{f_b \to x_2}(x_2) = \sum_{x_3} f_b(x_2, x_3)$$

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

$$\mu_{f_a \to x_1}(x_2) = \sum_{x_2} f_a(x_1, x_2) \mu_{x_2 \to f_a}(x_2)$$

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

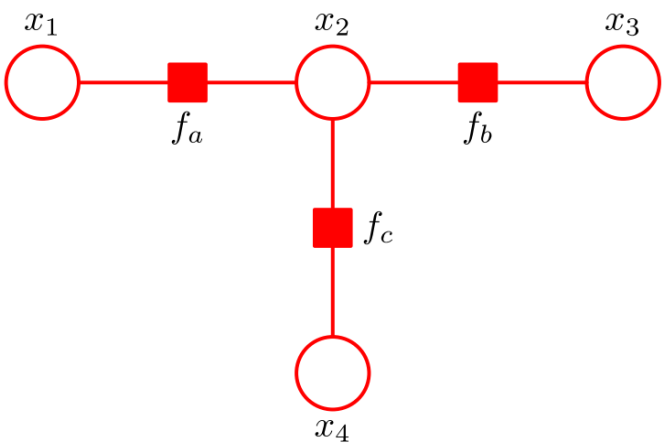
$$\mu_{f_c \to x_4}(x_4) = \sum_{x_2} f_c(x_2, x_4) \mu_{x_2 \to f_c}(x_2)$$



一个实际例子

- 验证边界$p(x_2)$

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



与下式一致

$$
p(x) = \sum_{\boldsymbol{x} \setminus x} p(\boldsymbol{x})
$$

以观测变量为条件

- 假设我们将$\boldsymbol{x}$划分为
  - 隐变量$\boldsymbol{h}$
  - 观测变量$\boldsymbol{v} = \hat{\boldsymbol{v}}$
- 计算 $p(h|\boldsymbol{v} = \hat{\boldsymbol{v}}) = \sum_{\boldsymbol{x}\setminus h} p(\boldsymbol{x})$
  - 我们只需要更新$p(\boldsymbol{x})$

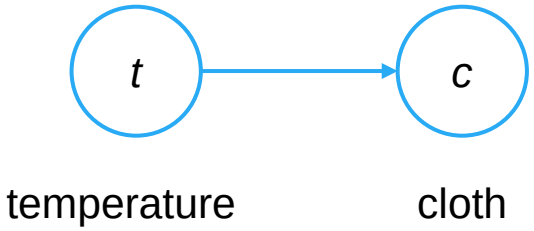
$$p(\boldsymbol{x}) \leftarrow p(\boldsymbol{x}) \prod_i I(v_i = \hat{v}_i)$$

- 加乘算法是高效的

总结概率图模型

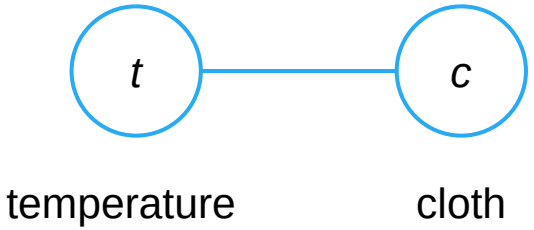
- 概率图模型是一套有效刻画多个随机变量之间的依赖关系，从而高效建模变量的联合分布的数学工作

贝叶斯网络



$$p(t,c) = p(t)p(c|t)$$

马尔可夫网络



$$p(t,c) = \frac{e^{\phi(t,c)}}{\sum_{t',c'} e^{\phi(t',c')}}$$

- 后验概率正比于先验概率乘以数据似然 $p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{w})p(\mathbf{t}|\mathbf{w})$
- 基于概率图模型做目标变量的条件推断可以使用加乘方法快速求解

THANK YOU