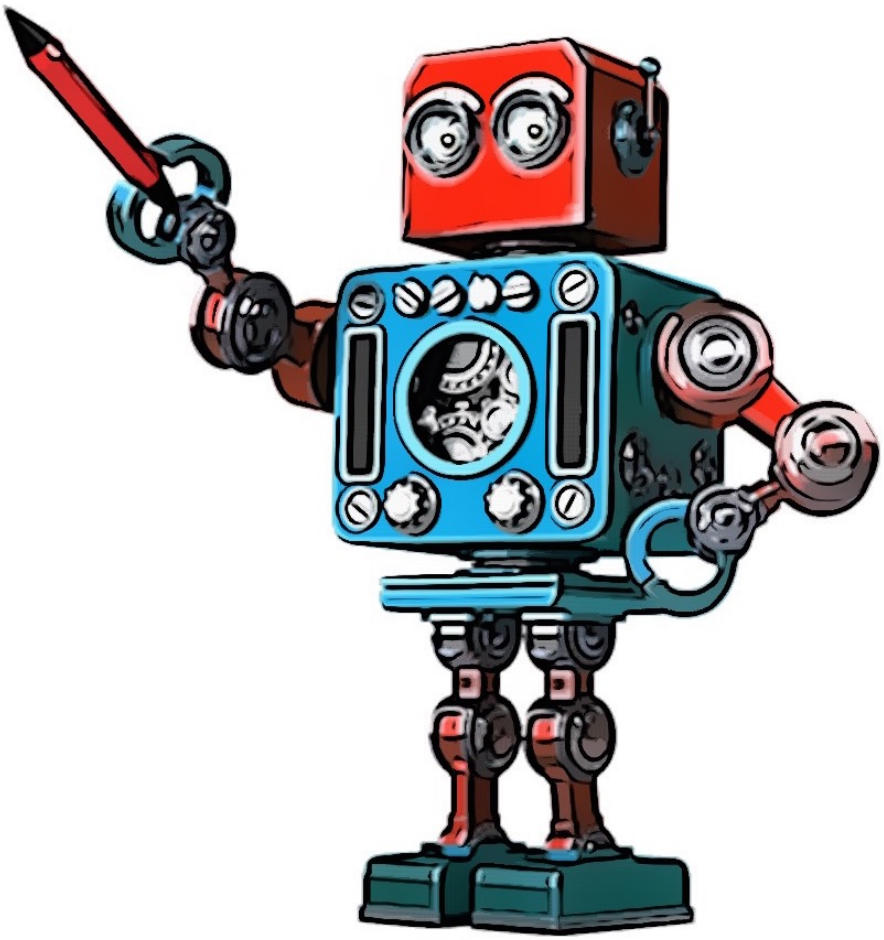
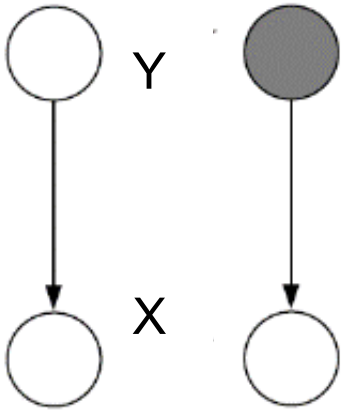
复习回顾：

- 概率图模型基础



suli@ucas.ac.cn

补充复习



- Product Rule

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

- Chain Rule 多个随机变量的联合概率

$$p(x_1,x_2,x_3) = p(x_1 \mid x_2,x_3)P(x_2,x_3)$$

$$= p(x_1 \mid x_2,x_3)p(x_2 \mid x_3)p(x_3)$$

$$p(x_1,\dots,x_D) = \prod_{j=1}^{D} p(x_j \mid x_{j+1},\dots,x_D)$$

补充复习

- Sum Rule

边际化 $$p(x) = \sum_{y \in Y} p(x, y)$$

- 条件概率

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{1}{z} p(x, y)$$

归一化 $$z = p(x) = \sum_{y} p(x, y)$$

补充复习

- 贝叶斯公式

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

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

补充复习

- 独立：$X \perp Y$

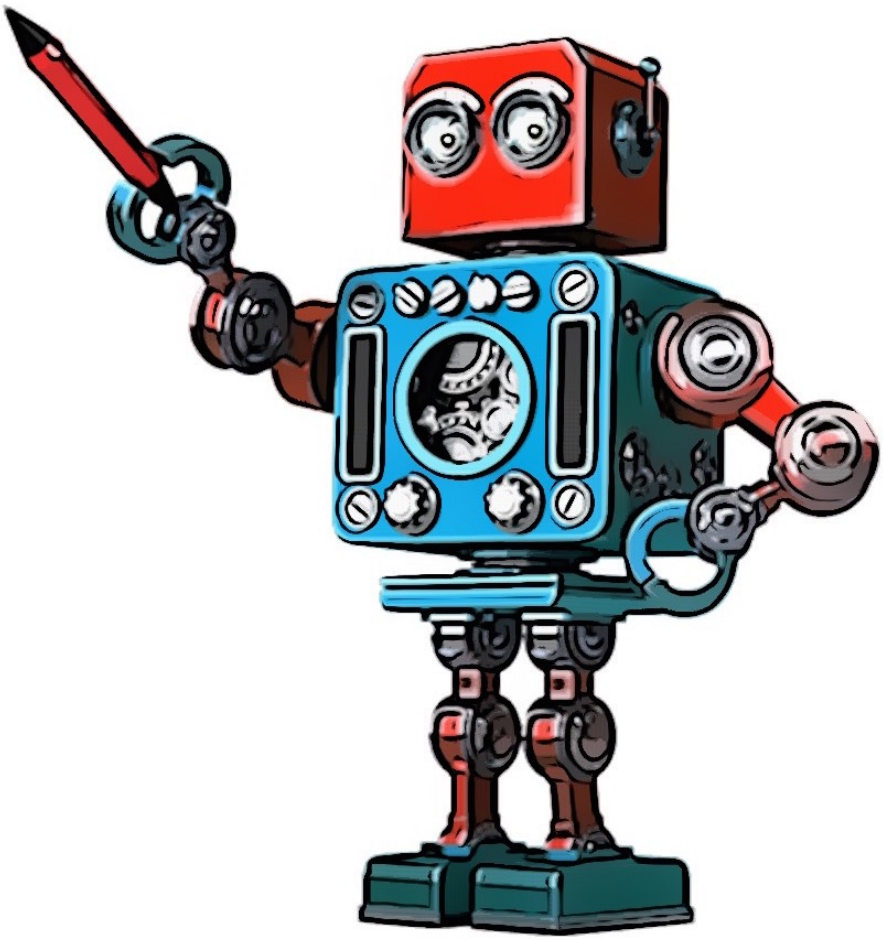
$$p(x,y) = p(x)p(y)$$
$$p(x \mid y) = p(x)$$
$$p(y \mid x) = p(y)$$

- 条件独立：$X \perp Y|Z$

$$p(x,y \mid z) = p(x \mid z)p(y \mid z)$$
$$p(x \mid y,z) = p(x \mid z)$$
$$p(Y \mid x,z) = p(y \mid z)$$

简介

- 两类概率图模型
  - 有向概率图模型
  - 无向概率图模型
- 学习和推断
- 典型的概率图模型



概率图模型
（Probabilistic Graphical Models）

- 描述多元随机变量的条件独立性的概率模型
- 结构预测：元素具有依赖约束的序列预测
- 三大基本问题：
  - 表示
  - 推断
  - 学习

三大基本问题

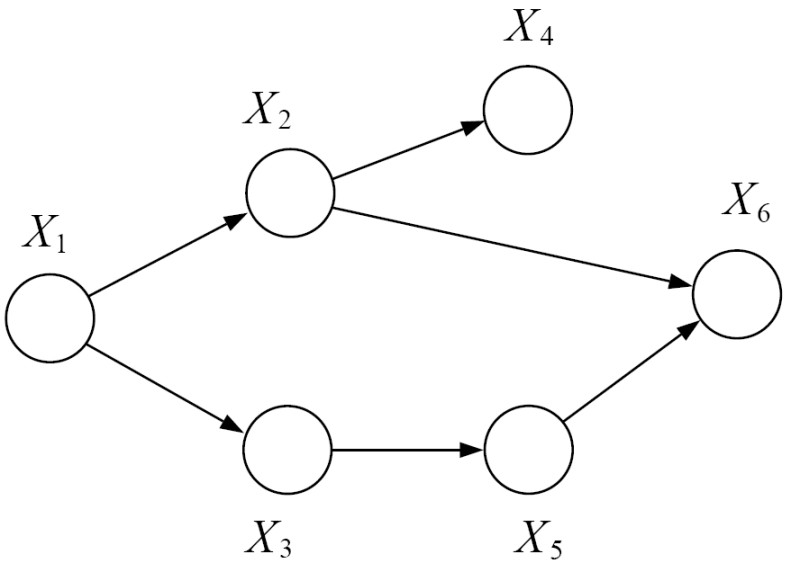
1. 表示：能够用模型去描述随机变量之间依赖关系
- 联合概率 ： $P(X)=P(x_1, x_2, \ldots, x_D)= P(X_O, X_H)$
- 条件独立性： $\{x_i\} \perp \{x_j\} | \{x_n\}$

2. 推断：给定观测数据，逆向推理，回答非确定性问题
- 条件概率：用已观测变量推测未知变量分布$P(X_H | X_O)$

3. 学习：给定观测数据，学习最佳模型（结构，参数）
- 联合概率最大化时的M参数：

$$\Theta^* = \arg\max_{\theta} P(X \mid \theta)$$

有向概率图模型 (贝叶斯网)



没有环

1.概率分布 → 用于查询/推断

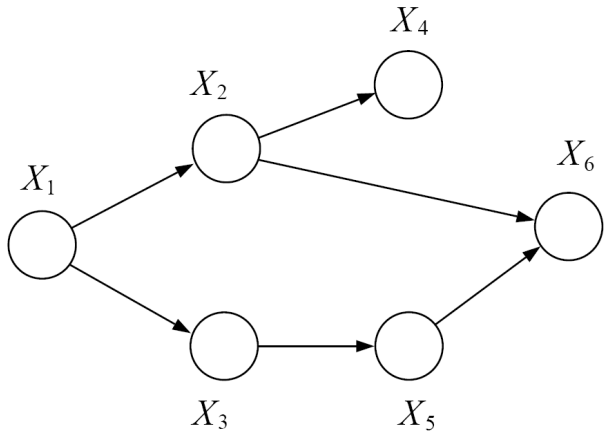
2.表示 → 具体实现

3.条件独立 → 模型的解释

1. 概率分布

一个概率图模型对应着一族概率分布（a family of probability distribution ）

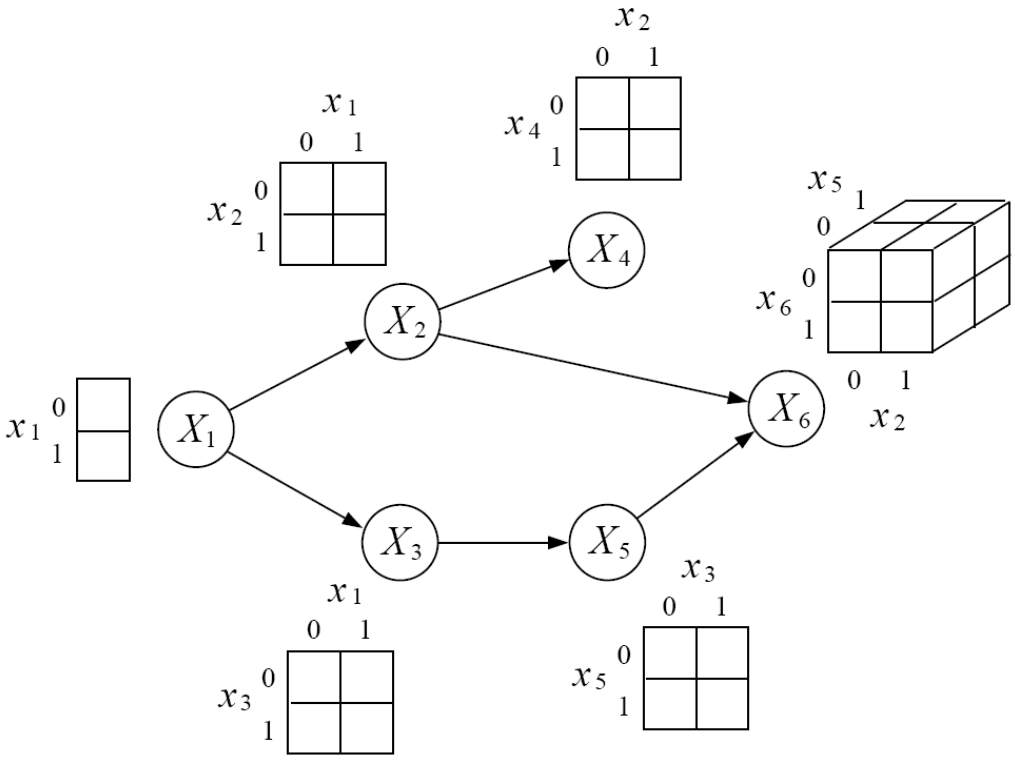
每个节点对应一个条件概率分布$p\left(x_j\middle|x_{\pi_j}\right)$，其中$\pi_j$表示节点$j$的父亲节点集合。联合概率分布可以表示为：$p(x_1, x_2, \dots x_D) = \prod_{j=1}^{D} p(x_j|x_{\pi_j})$



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

2. 表示

贝叶斯网使用一系列变量间的“局部”关系“紧凑”地表示联合概率分布



条件概率表CPT $p(x_6\|x_2,x_5)$	x_6=0	x_6=1
x_2=0， x_5=0		
x_2=0， x_5=1		
x_2=1， x_5=0		
x_2=1， x_5=1		

$$O(2^D) \rightarrow O(D\times 2^k)$$

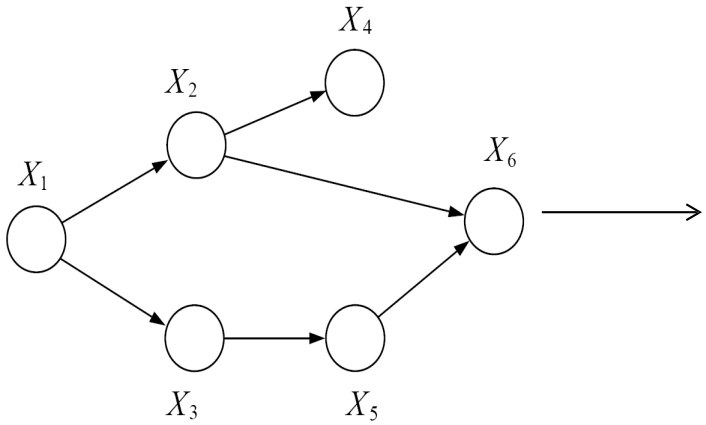
通常D>>k

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

3. 条件独立

利用条件独立解释缺少的边

定义图G中节点的顺序I，如果对每个节点$i \in V$，他的父节点都在这个顺序中出现在它之前，那么我们称I为拓扑排序. 例如$I = \{1,2,3,4,5,6\}$是图的一种拓扑排序。对于节点j，给定图G的拓扑排序I，假设v_j表示在I中除了π_j之外所有出现在j个节点之前的节点，我们将这样一组条件独立性陈述$\{X_j \perp X_{v_j} | X_{\pi_j}\}$和图关联起来



断言：给定一个节点的父节点，该节点和其祖先条件独立

补充复习

- 独立

$$P(X,Y) = P(X)P(Y)$$

等价 $P(X) = P(X \mid Y)$

独立 推不出 条件独立：

有两枚正反概率均为 50% 的硬币，设事件 A 为第一枚硬币为正面，事件 B 为第二枚硬币为正面，事件 C 为两枚硬币同面。A 和 B 显然独立，但如果 C 已经发生，即已知两枚硬币同面，那么 A 和 B 就不条件独立了。

- 条件独立

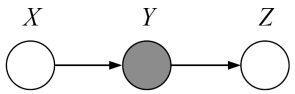
$$P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

等价 $P(X \mid Z) = P(X \mid Y,Z)$

条件独立 推不出 独立：

一枚硬币正面概率为99%，另一枚反面概率为99%，随机拿出一枚投掷两次，事件A为第一次正面，事件B为第二次正面，事件C为拿出的是第一枚硬币。可算出 P(B) = 0.5 但 P(B|A) = 0.9802，说明A和B不独立。但如果C已发生，则A和B条件独立。

三种经典图



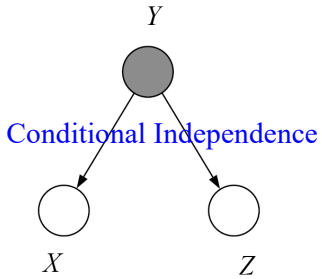
$$p(x,y,z) = p(x)p(y|x)p(z|y)$$

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

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

$$X \perp\!\!\!\perp Z \mid Y$$

经典的马尔科夫链
“过去”，“现在”，“未来”



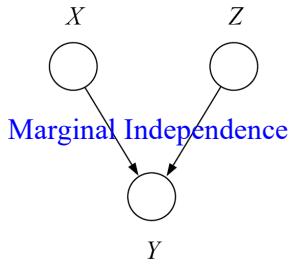
$$p(x,y,z) = p(y)p(x|y)p(z|y)$$

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

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

$$X \perp\!\!\!\perp Z \mid Y$$

共同的起因（Common Cause）
Y “解释” X 和 Z之间所有的依赖



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

$$p(x,z) = p(x)p(z)$$

$$X \perp\!\!\!\perp Z$$

共同效应（Common Effect）
多个相互竞争的解释

条件独立(快速检验)

贝叶斯球算法(规则)：假设在贝叶斯网络中，有一个按一定规则运动的球。**已知中间节点（或节点集合）Z，如果球不能由节点X出发到达节点Y（或者由Y到X），则称X和Y关于Z（条件）独立。**

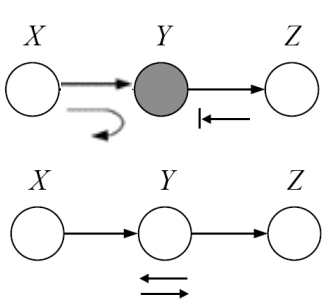
规则：

- **未知节点**：总能使贝叶斯球通过，同时还可以反弹从其子节点方向来的球。（父 -> 子）|（子 -> 父/子）

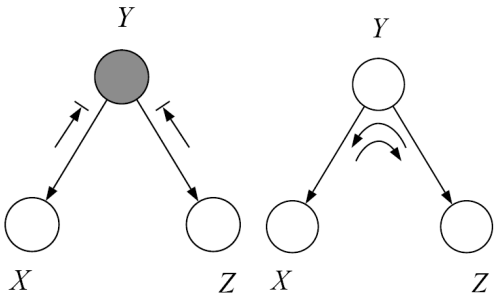
- **已知节点**：反弹从其父节点方向过来的球，截止从其子节点方向过来的球。（父 -> 父）|（子 -> “截止”）

条件独立(快速检验)

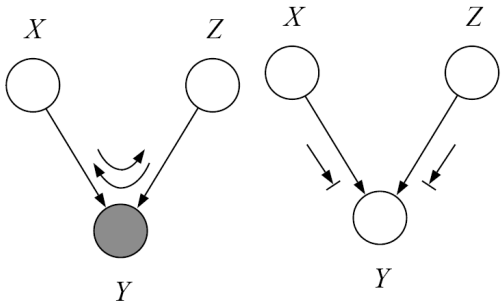
- 未知节点：总能使贝叶斯球通过，同时还可以反弹从其子节点方向来的球
- 已知节点：反弹从其父节点方向过来的球，截止从其子节点方向过来的球



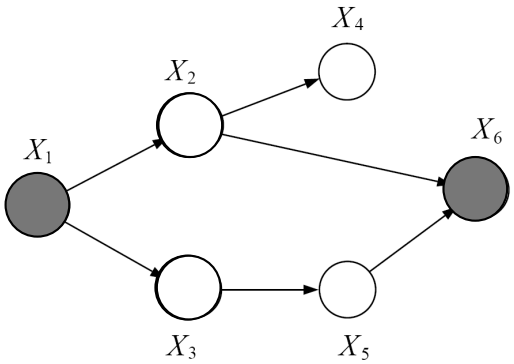
一个输入箭头和
一个输出箭头



两个输出箭头



两个输入箭头

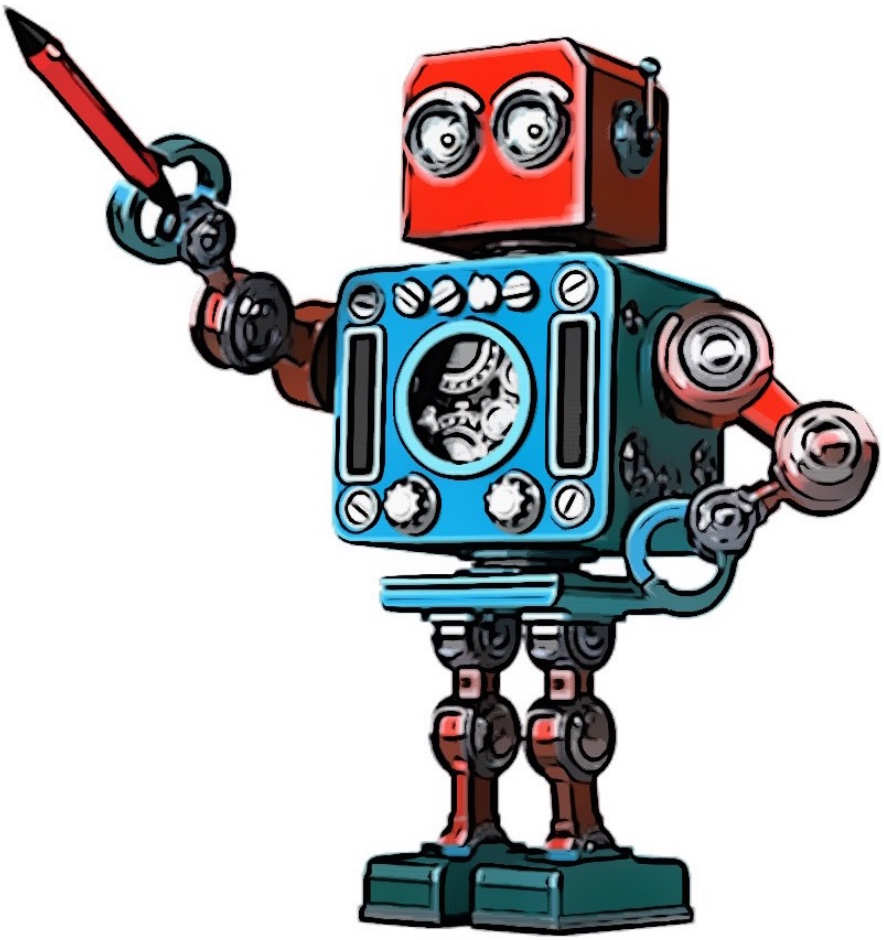


检查通过可达性

$X_1 \perp\!\!\!\perp X_6 \mid \{X_2, X_3\}$ ✓

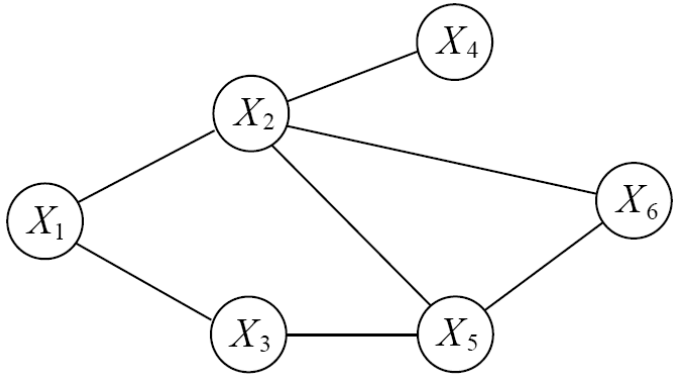
$X_2 \perp\!\!\!\perp X_3 \mid \{X_1, X_6\}$ ✗

- 简介
- 两类概率图模型
  - 有向概率图模型
  - 无向概率图模型
- 学习和推断
- 典型的概率图模型



无向图模型 (马尔科夫随机场)

定义 一个无向图$G = (V, E)$包含节点集合V和边的集合E，边由点对组成



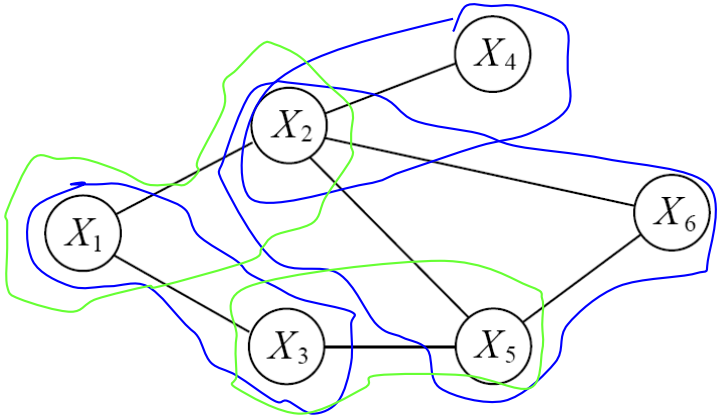
1.概率分布 → 用于查询/推断

2.表示 → 具体实现

3.条件独立 → 模型的解释

1. 概率分布

极大团



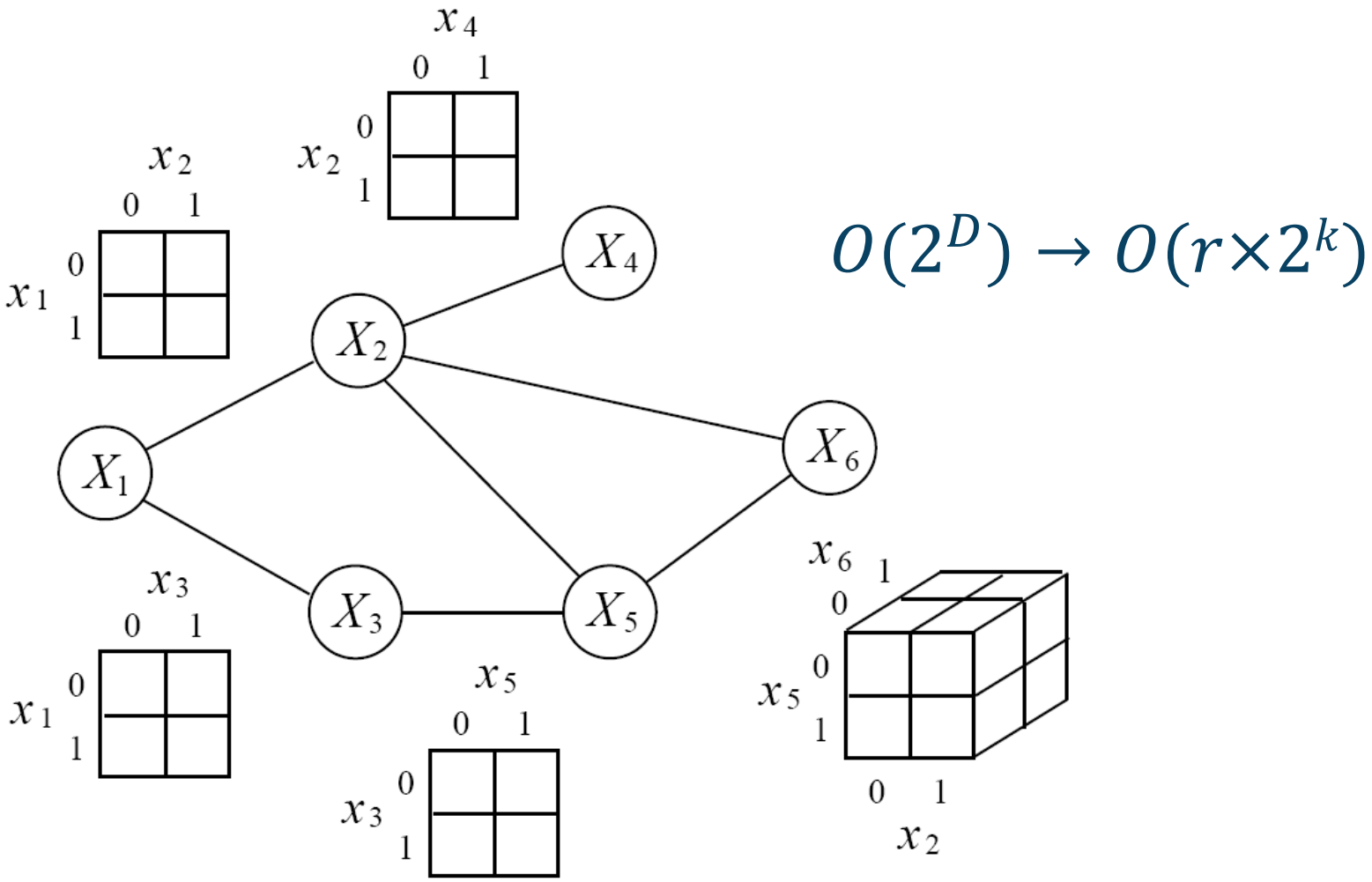
- 联合概率分布

$$P(x) = \frac{1}{Z}\prod_{C\in\mathcal{C}}\psi_{X_C}(x_C)$$

- 归一化因子

$$Z = \sum_{x}\prod_{C\in\mathcal{C}}\psi_{X_C}(x_C)$$

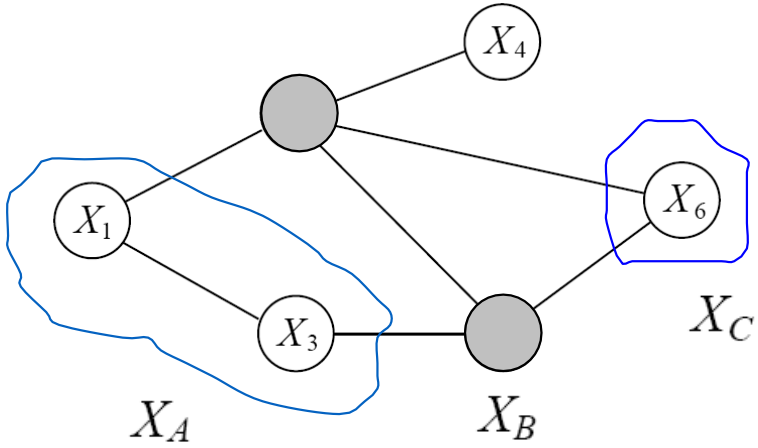
2. 表示



$$O(2^D) \rightarrow O(r \times 2^k)$$

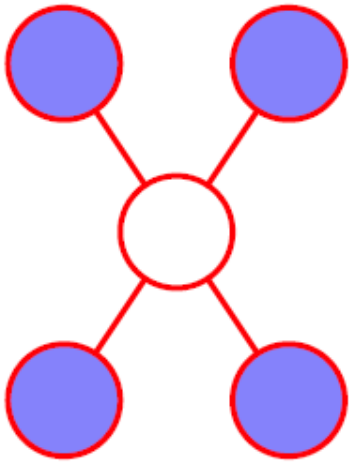
3. 条件独立

朴素图论分割



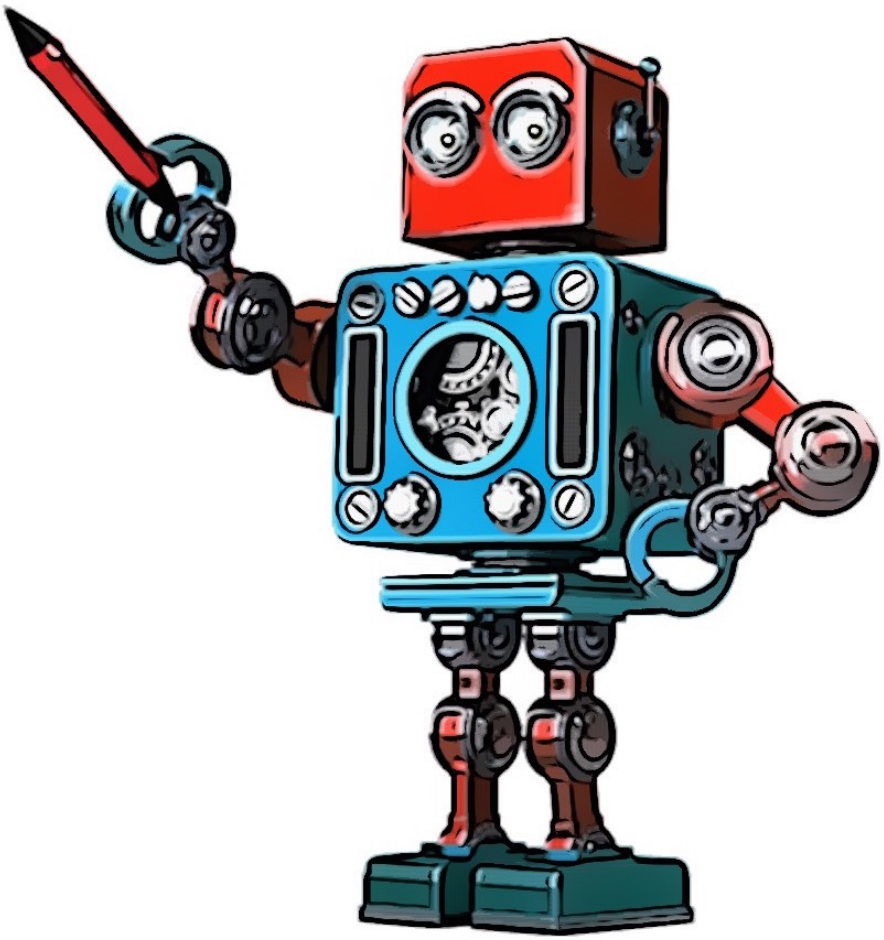
$X_A \perp X_C | X_B$

图论中的“可达性”问题



对于一个无向图, 一个节点所有邻居节点，构成该节点的马尔科夫包裹(blanket)

- 简介
- 两类概率图模型
  - 有向概率图模型
  - 无向概率图模型
- **学习和推断**
- 典型的概率图模型

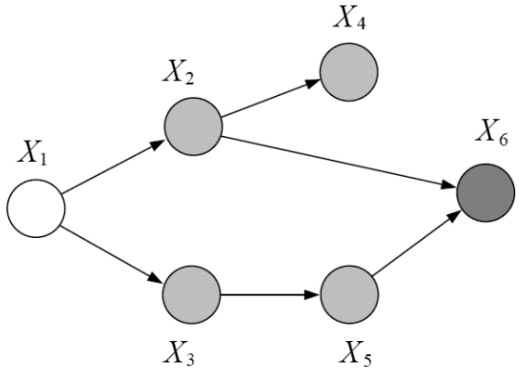


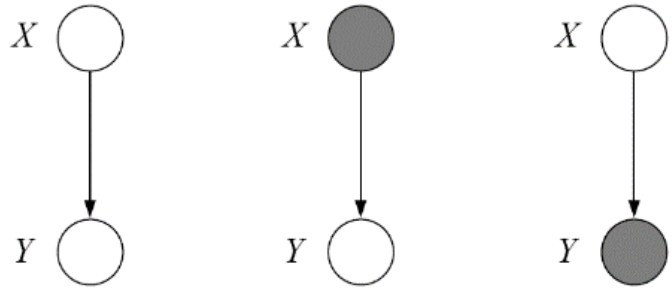
概率推断和学习

- 我们现在有了紧凑的概率分布表示: 概率图模型
- 概率图M 描述了唯一的概率分布P
- 典型任务:
  - 任务1: 我们如何回答关于P_M的查询, 例如$P_M(X|Y)$?
    - 我们使用**推断**表示计算上述问题答案的过程
  - 任务2: 我们如何基于数据$\mathcal{D}$估计**合理的模型M**?
    1. 我们使用**学习**来命名获得M的点估计过程。
    2. 对于**贝叶斯学派**，寻找p(M|D) 实际上是一个**推断**问题。
    3. 当不是所有的变量都是可观察时, 即使是计算M的点估计，也需要进行**推断**处理隐含变量。

1. 推断

- 可能性推断





边际概率： 例如 $p(X_6)$

后验概率： 例如 $p(X_2|X_6 = 1)$

后验概率 $p(x|y) = \dfrac{p(y|x)p(x)}{p(y)}$

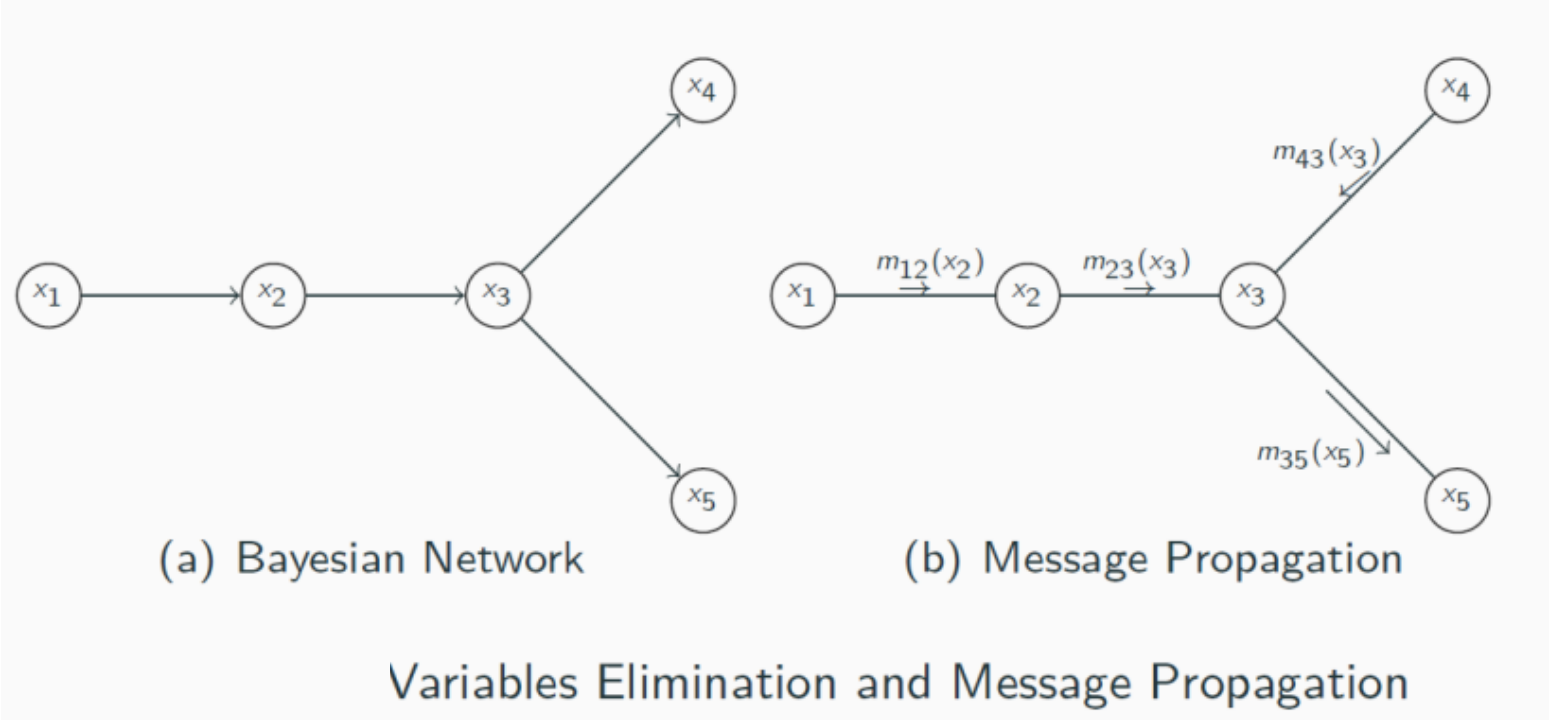
边际概率 $p(y) = \sum_x p(y|x)p(x)$

变量消去法

- 给定联合分布 $p(x_1, \dots, x_5)$, 要计算边缘分布$p(x_5)$:

$$p(x_5) = \sum_{x_1,x_2,x_3,x_4} p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_3)p(x_5|x_3)$$

- Sum Rule消去其他变量$\{x_1, \dots, x_4\}$

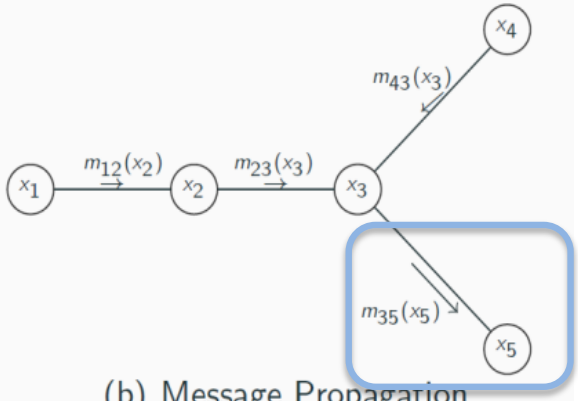


Variables Elimination and Message Propagation

变量消去

- 利用顺序$\{x_1, x_2, x_4, x_3\}$ 消去其他变量 $\{x_1, \ldots, x_4\}$

$$
\begin{aligned}
P(x_5) &= \sum_{x_1,x_2,x_3,x_4} P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_3)P(x_5|x_3) \\
&= \sum_{x_3} P(x_5|x_3) \sum_{x_4} P(x_4|x_3) \sum_{x_2} P(x_3|x_2) \sum_{x_1} P(x_1)P(x_2|x_1) \\
&= \sum_{x_3} P(x_5|x_3) \sum_{x_4} P(x_4|x_3) \sum_{x_2} P(x_3|x_2) m_{12}(x_2) \\
&= \sum_{x_3} P(x_5|x_3) \sum_{x_4} P(x_4|x_3) m_{23}(x_3) \\
&= \sum_{x_3} P(x_5|x_3) m_{43}(x_3) m_{23}(x_3) \\
&= m_{35}(x_5)
\end{aligned}
$$



(b) Message Propagation

变量消去法

- Sum-Product算法: 适用于贝叶斯网络和马尔科夫网络

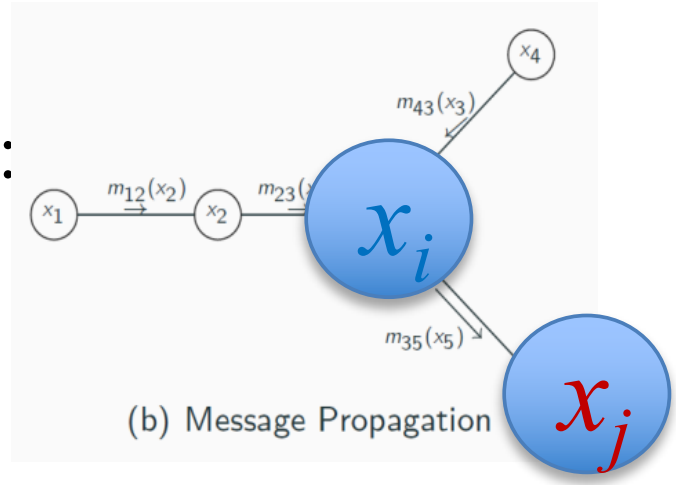
$$m_{i\to j}(x_j) = \sum_{x_i} \psi(x_i, x_j) \prod_{k\in n(i)\backslash \mathrm{j}} m_{ki}(x_i)$$

权重函数

- 计算多个边际概率有很多重复计算
- **信念传播法（Blief Propagation）**：

把$m_{i\to j}(x_j)$ 作为x_i 传递到 x_j的消息

- 边际分布：



$$p(x_i) \propto \prod_{k\in n(i)} m_{ki}(x_i)$$

正比于 当前节点x_i联通的其他所有节点传来的消息的连乘积

信念传播算法

1. 叶子到根：指派一个根节点， 从叶子节点开始传递信息，直到根节点接收到所有邻居节点传来的信息
2. 根到叶子：从根节点开始传播信息，直到所有叶子节点接受到信息

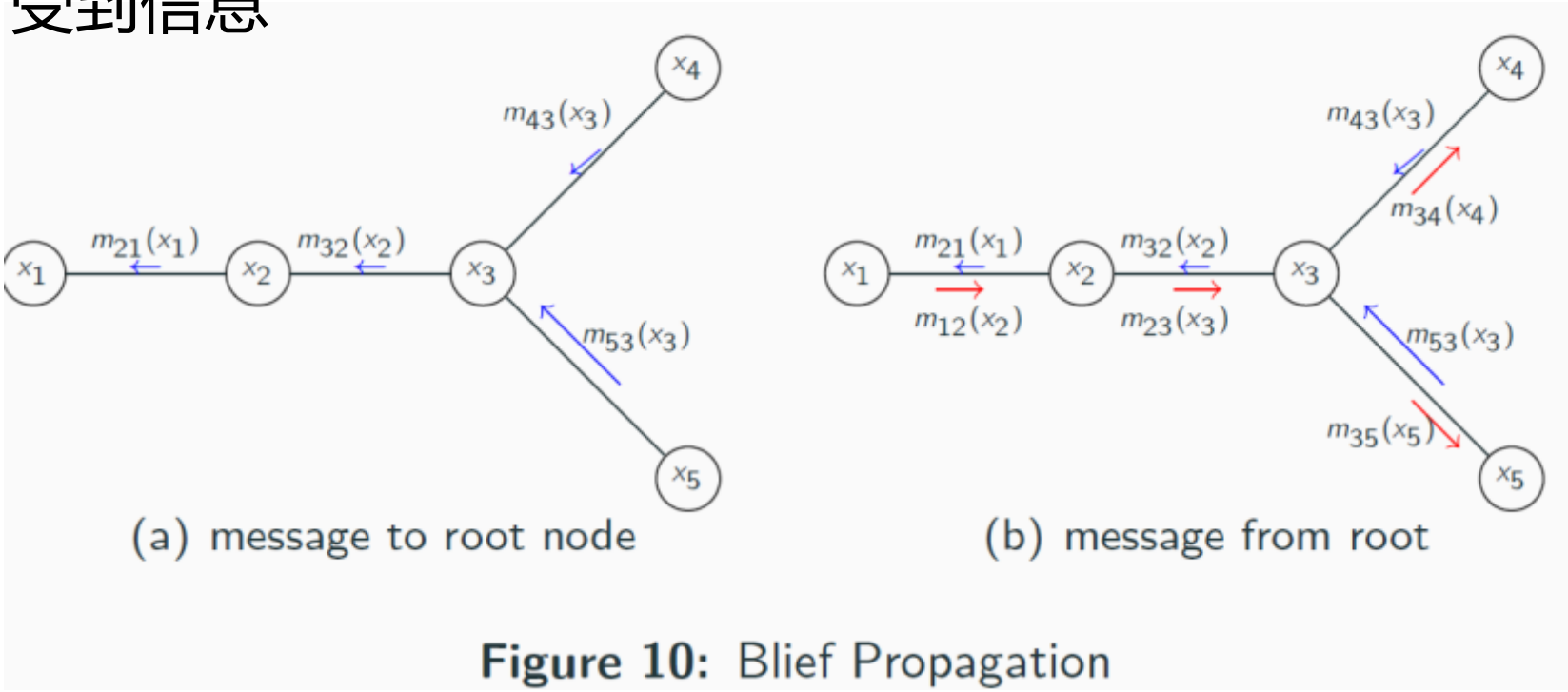
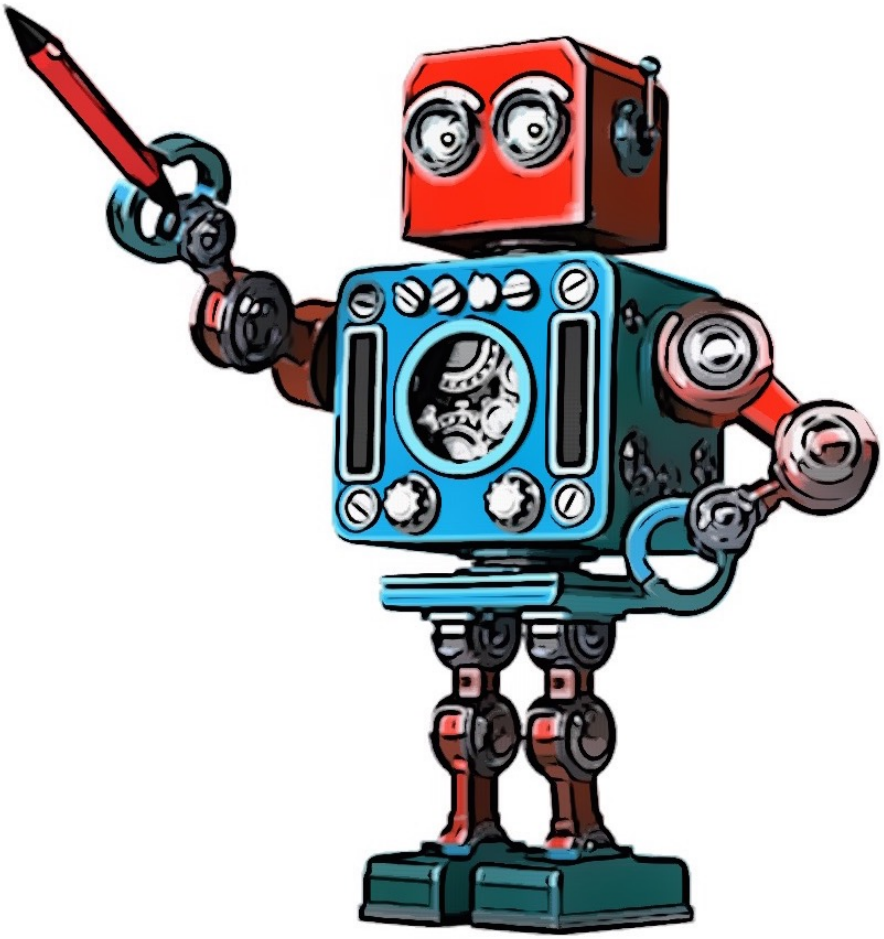
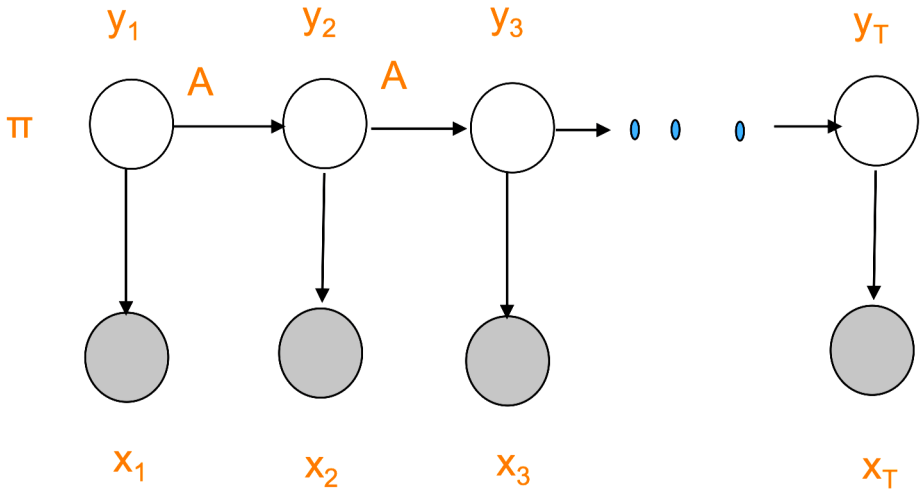


Figure 10: Blief Propagation

- 简介
- 两类概率图模型
  - 有向概率图模型
  - 无向概率图模型
- 学习和推断
- 典型的概率图模型
  - HMM
  - MEMM
  - CRF



HMM简介



可能的文字

观测到的语音

语音识别

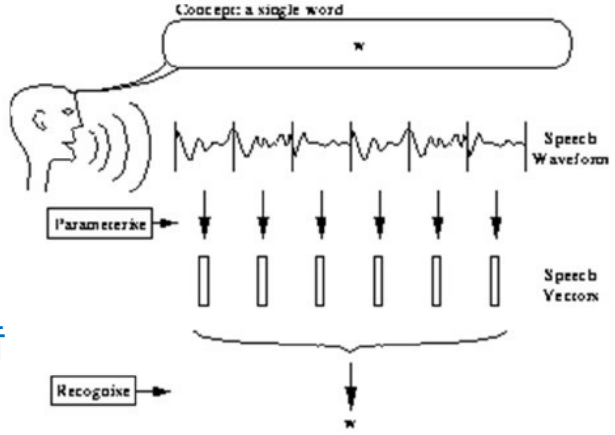


Fig. 1.2 Isolated Word Problem

- 顶层节点表示隐含变量y_t（状态节点）
- 底层节点表示观测变量x_t（输出节点）
- 假设隐含变量y_t的取值范围为状态空间$\{s_1, s_2, \ldots, s_N\}$，观测变量$x_t$的取值范围为$\{o_1, o_2, \ldots, o_M\}$（实际问题中可为连续值）

- 符号约定：

$$y_t^j = 1 \quad \leftrightarrow \quad y_t = s_j$$
$$x_t^j = 1 \quad \leftrightarrow \quad x_t = o_j$$

回答经典问题

1. 状态序列解码（推断）
   问题： 给定
   - 观察序列 $\mathbf{x}$
   - 模型参数$\boldsymbol{\theta}$

●寻找
最优的状态序列$\mathbf{y}$
$\boldsymbol{p(y|x,\theta)}$

语音识别为例：

给定一段语音，语言模型，
对应的文字是？

2. 似然评估问题：给定
   - 观察序列 $\mathbf{x}$
   - 模型参数$\boldsymbol{\theta}$

●计算
似然函数$p(\mathbf{x}|\boldsymbol{\theta})$

按照当前语言模型规则，
观测到的语音是否像句话？

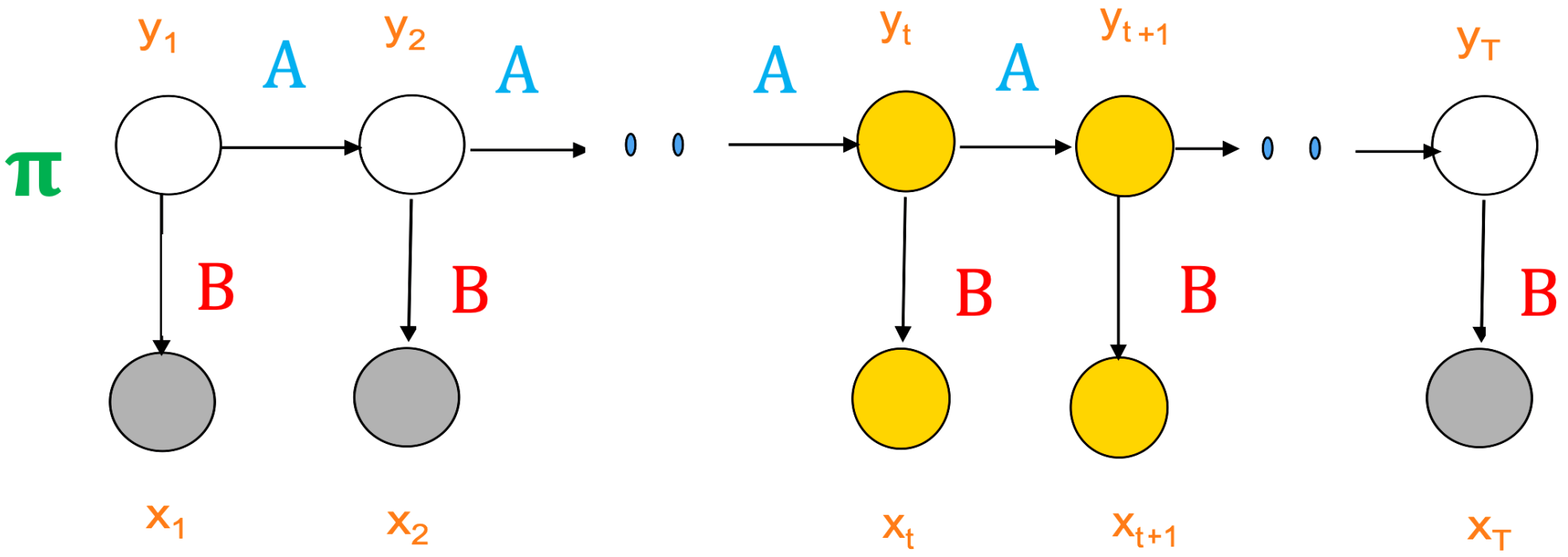
3. 参数估计问题(学习)：
   给定
   - 观察序列$\mathbf{x}$

●寻找
$\boldsymbol{\theta}$的ML估计：
$\boldsymbol{\theta}_{ML} = \text{argmax}\{p(\mathbf{x}|\boldsymbol{\theta})\}$

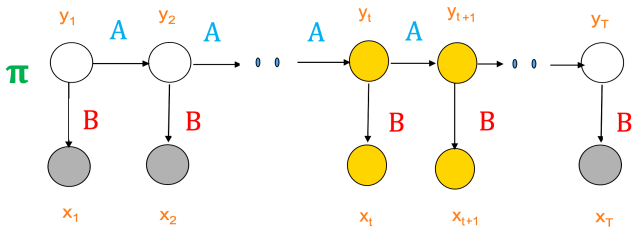
只有观测的语音，
学习当前语言模型规则

HMM条件独立

- 给定状态节点y_t，对于时刻$s < t$ 和$t < u$， y_s独立于y_u。
- 给定状态节点y_t，输出节点x_s和x_u也相互独立。
- 给定输出节点，不带来任何条件独立。



HMM表示



- 第一个状态节点对应一个初始状态概率分布

  $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$： $\pi_i = P(y_1^i = 1)$;

- 状态转移矩阵**A**，其中a_{ij}为转移概率：$a_{ij} = P(y_{t+1}^j = 1 | y_t^i = 1)$, $1 \le i \le N,\ 1 \le j \le N$;

- 每个输出节点有一个状态节点作为父节点，对应发射概率矩阵**B**：$b_{ij} = p(x_t^j = 1 | y_t^i = 1)$, $1 \le i \le N,\ 1 \le j \le M$;

- 对于特定的配置，$(\mathbf{x}, \mathbf{y}) = (x_0, x_1, \cdots, x_T, y_0, y_1, \cdots, y_T)$联合概率可以表示为：

T个时刻的发射概率

$$p(\mathbf{x}, \mathbf{y}) = p(y_1) \prod_{t=1}^{T-1} p(y_{t+1} | y_t) \prod_{t=1}^{T} p(x_t | y_t)$$

初始状态y_1概率　　后续T-1个时刻的状态y的转移概率

HMM推断

- 状态序列解码：给定观测序列$\mathbf{x} = (x_1 \cdots x_T)$为整个可观测的输出，计算隐状态序列$\mathbf{y}$的概率：

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

- 给定输出序列$\mathbf{x}$，计算特定单个隐含状态y_t的分布
- 给定部分输出，计算条件概率
  - 过滤：最后状态$p(y_t|x_1,\cdots,x_t)$
  - 预测：未来状态$p(y_t|x_1,\cdots,x_s)$其中$t > s$
  - 平滑：基于已有和未来的数据计算后验概率

    $p(y_t|x_1,\cdots,x_u)$,其中$t < u$

HMM推断

- 状态序列解码：给定观测序列$\mathbf{x} = (x_1 \cdots x_T)$为整个可观测输出，计算隐状态序列$\mathbf{y}$的概率：$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})}$
- 似然评估：为计算$p(\mathbf{x})$, 需要将所有可能的隐状态值求和

$$p(\mathbf{x}) = \sum_{y_1}\sum_{y_2}\cdots\sum_{y_T}\pi(y_1)\prod_{t=1}^{T-1}a_{y_t,y_{t+1}}\prod_{t=1}^{T}b_{y_t,x_t}$$

  - T个隐状态节点，每个有 N 个可能值，意味着我们需要做N^T次求和
  - 把求和放到乘法里，并形成递归形式，显著减少计算

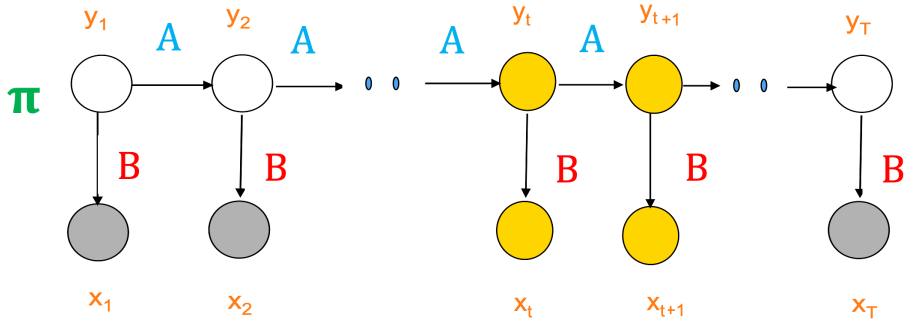
HMM推断：递归算法

$$p(y_t|\mathbf{x}) = \frac{p(\mathbf{x}|y_t)p(y_t)}{p(\mathbf{x})}$$

$$= \frac{p(x_1 \dots x_t|y_t)\boldsymbol{p(x_{t+1} \dots x_T|y_t)}p(y_t)}{p(\mathbf{x})}$$

给定yt，x从1到t时刻的观测输出 和x从t+1到T时刻的观测输出是条件独立的

$$p(y_t|\mathbf{x}) = \frac{\alpha(y_t)\boldsymbol{\beta(y_t)}}{p(\mathbf{x})}$$



$$p(\mathbf{x}) = \sum_{y_t} \alpha(y_t)\boldsymbol{\beta(y_t)}$$

其中$\alpha(y_t)$是产生部分输出序列 $x_1, \cdots, x_t$ ， 并结束于y_t的概率

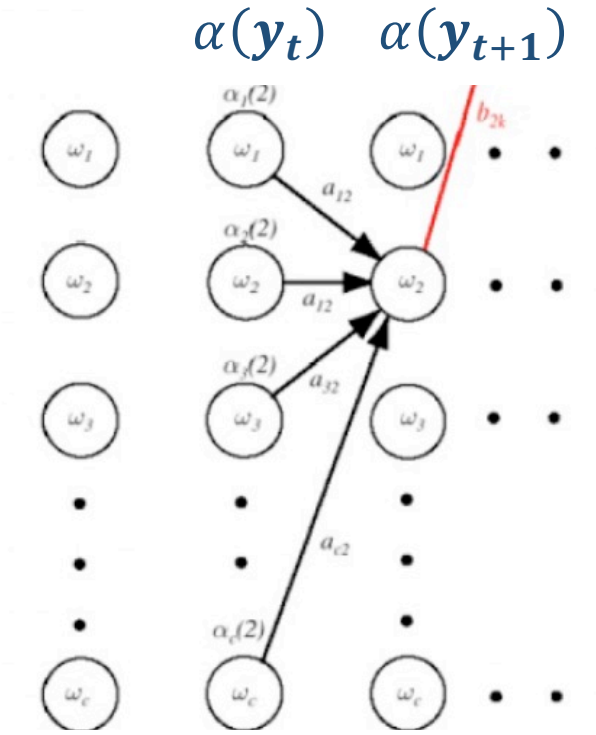
其中$\beta(y_t)$是从y_t状态开始，产生输出序列$x_{t+1}, \cdots, x_T$的概率

α递归计算—前向算法

- 因此可将$p(y_t|\mathbf{x})$转化为计算α，β
- 可以获得$\alpha(y_t)$和 $\alpha(y_{t+1})$的递归关系：

$$\alpha(y_{t+1}) = p(x_1 \dots x_t, x_{t+1}, y_{t+1})$$

$$= \sum_{y_t} \alpha(y_t)\, a_{y_t,y_{t+1}}\, b_{y_{t+1},x_{t+1}}$$



- 初始化：定义α的第一步

$$\alpha(y_1) = p(x_1, y_1)$$
$$= p(x_1|y_1)p(y_1)$$
$$= b_{y_1,x_1}\pi_{y_1}$$

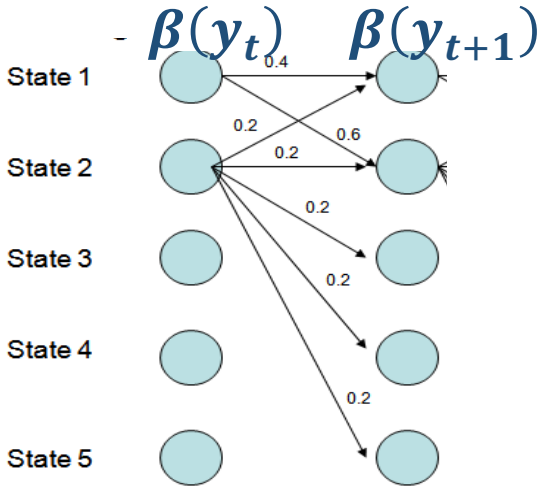
- 终止:

似然函数 $p(x_1 \dots x_T|\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}) = \sum_{y_T} \alpha(y_T)$

β递归计算—后向算法

- 类似地，我们获得后向递归关系

$$\begin{aligned}&\boldsymbol{\beta}(\boldsymbol{y_t})\\&=\boldsymbol{p}(\boldsymbol{x_{t+1}}\ldots\boldsymbol{x_T}|\boldsymbol{y_t})\\&=\sum_{\boldsymbol{y_{t+1}}}\boldsymbol{a_{y_t,y_{t+1}}}b_{y_{t+1},x_{t+1}}\ \boldsymbol{\beta}(\boldsymbol{y_{t+1}})\end{aligned}$$



- 初始化： $\beta(y_T)$ =1
- 终止：

似然函数 $p(x_1 \ldots x_T|\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}) = \sum_{y_1} \boldsymbol{\pi}_{y_1} \boldsymbol{p}(\boldsymbol{x_1}|\boldsymbol{y_1})\beta(y_1)$

状态转移后验概率的推断

- $\alpha-\beta$算法可以扩展到计算状态转移的后验概率$p(y_t, y_{t+1}|\mathbf{x})$。
  - 定义$\boldsymbol{\xi(y_t, y_{t+1}) \equiv p(y_t, y_{t+1}|\mathbf{x})}$
- 我们基于α和β，计算$\xi(y_t, y_{t+1})$

贝叶斯公式展开

$$\xi(y_t, y_{t+1}) \equiv p(y_t, y_{t+1}|\mathbf{x}) = \frac{p(\mathbf{x}|y_t, y_{t+1})p(y_{t+1}, y_t)}{p(\mathbf{x})}$$

$$= \frac{p(\mathbf{x}|y_t, y_{t+1})p(y_{t+1}|y_t)p(y_t)}{p(\mathbf{x})}$$

过去1~t，现在t+1，未来t+2~T 独立

$$= \frac{p(x_1 \dots x_t|y_t)p(x_{t+1}|y_{t+1})p(x_{t+2} \dots x_T|y_{t+1})p(y_{t+1}|y_t)p(y_t)}{p(\mathbf{x})}$$

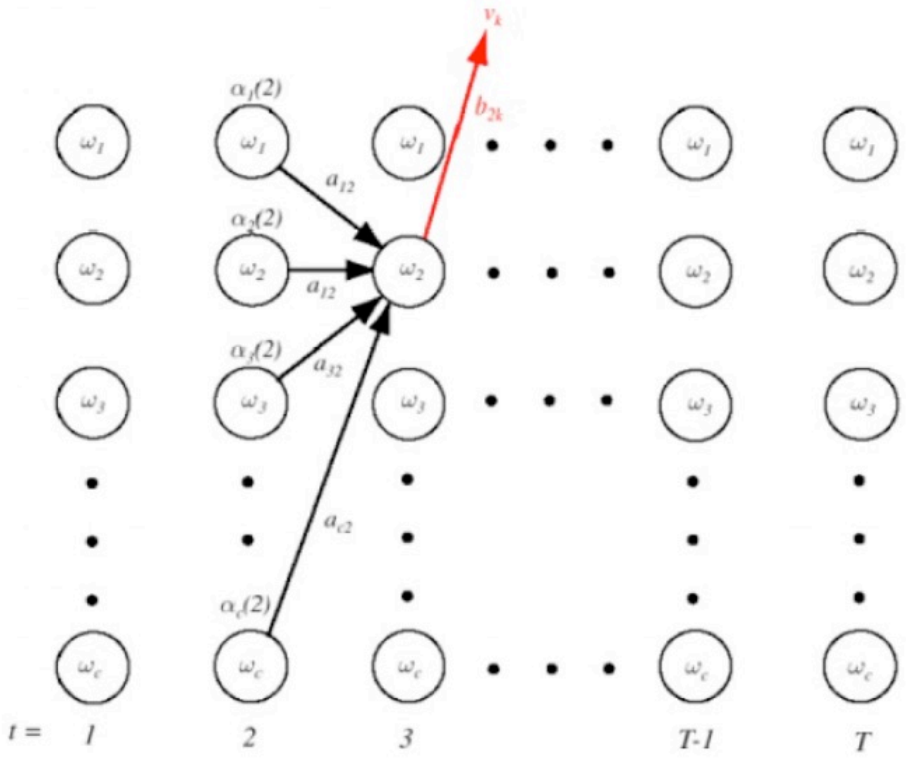
$\alpha(y_t) = p(x_1 \dots x_t|y_t)\, p(y_t)$

$$= \frac{\alpha(y_t)b_{y_{t+1},x_{t+1}}\beta(y_{t+1})a_{y_t,y_{t+1}}}{p(\mathbf{x})}$$

$\beta(y_t) = p(x_{t+1} \dots x_T|y_t)$

HMM推断：小结

- 我们可以递归地计算HMM所有的后验概率
- 给定一个观测序列**x**，前向计算α-递归
- 如果需要似然函数，只需简单地求和最终步的α
- 如果需要状态的后验概率，再使用β-递归



状态序列解码: Viterbi解码

- 我们现在可以计算

$$p(y_t = s_k|\mathbf{x}) = \frac{\alpha(y_t)\beta(y_t)}{p(\mathbf{x})}$$

- 序列$\mathbf{x}$在时刻t的最可能的状态是：

$$k_t^* = argmax\, p(y_t = s_k|\mathbf{x})$$

  - 这是单个隐状态的MAP，如果我们想要整个序列的最大后验？
  - 后验解码：$\{y_t = s_{k_t^*} : t = 1 \ldots T\}$
  - 这是不是整个隐状态序列的MAP？

Viterbi解码：动态规划

- 给定$x = x_1, \ldots, x_T$，我们要找 $y = y_1, \ldots, y_T$，使得$p(\mathbf{y}|\mathbf{x})$最大

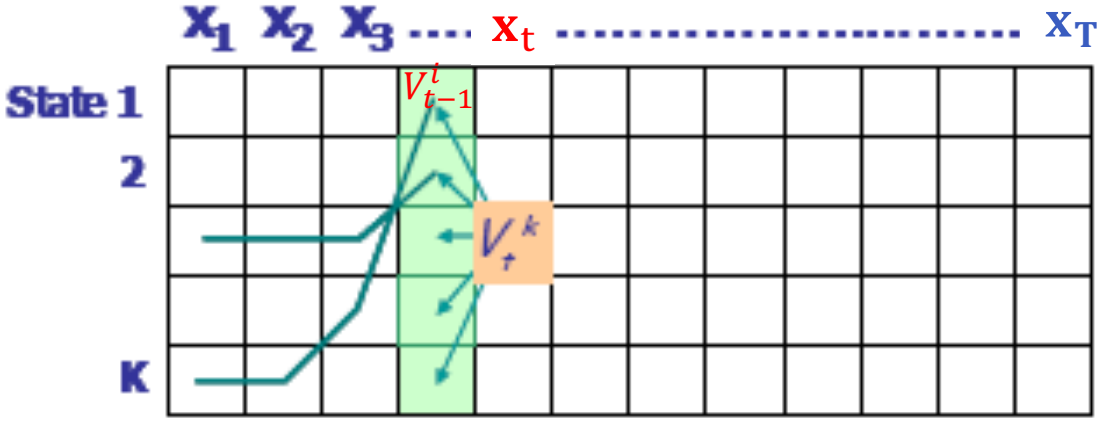
$$\mathbf{y}^* = \underset{\mathbf{y}}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{x}) = \underset{\mathbf{y}}{\operatorname{argmax}}\, p(\mathbf{x}, \mathbf{y})$$

去掉共同的分母$p(\mathbf{x})$

- 令$V_t^k = \max_{\{y_1,\ldots,y_{t-1}\}} p(x_1, \ldots, x_{t-1}, y_1, \ldots, y_{t-1}, x_t, y_t = s_k)$

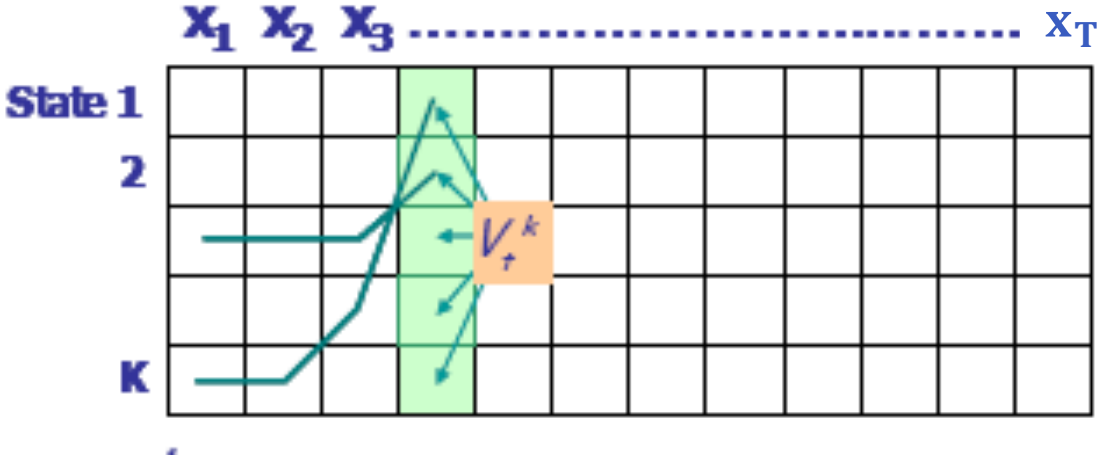
=结尾状态为$y_t = s_k$时，最可能状态序列的概率

- 递归：$V_t^k = p(x_t|y_t^k = 1)\max_i a_{i,k} V_{t-1}^i$



Viterbi解码

- 初始化：$V_1^k = p(x_1, y_1 = s_k) = \pi_k b_{y_1,x_1}$
- 之后就不断迭代：
- $V_t^k = p(x_t|y_t^k = 1)\max_i a_{i,k} V_{t-1}^i$
- 终止：$P^* = \max_{1 \le i \le N} V_T^i$



Viterbi 解码

- 假设有 3 个盒子，分别装有不同数量的苹果(A)和桔子(O)：
- 盒子一:2个A，2个O;
  盒子二:3个A，1个O;
  盒子三:1个A，3个O;
- 每次随机选择一个盒子并从中抽取一个水果，观测并记录看到的水果是哪种。不幸的是，忘记去记录所选的盒子号码，只记录了每次看到的水果是 A 还是 O。
- (1)请用 HMM 模型描述上述过程。
- (2)假如观测到水果序列为x={A，A，O，O，O}，请给出最可能的盒子序列。

参考答案：

（1）初始概率$\boldsymbol{\pi} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$，

盒子间的转移概率矩阵$\mathbf{A} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$，

发放概率矩阵（给定盒子时，选择每种水果的概率）$\mathbf{B} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}$。

当$t=1$时：已知 $x_1 = apple$，所以

初始化

$$V_1^1 = \pi_1 b_{1,x_1} = \frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$$

$$V_1^2 = \pi_2 b_{2,x_1} = \frac{1}{3} \times \frac{3}{4} = \frac{1}{4}$$

$$V_1^3 = \pi_3 b_{3,x_1} = \frac{1}{3} \times \frac{1}{4} = \frac{1}{12}$$

	t=1	t=2	t=3	t=4	t=5	
V_t^1	$\frac{1}{6}$					
V_t^2	$\frac{1}{4}$					
V_t^3	$\frac{1}{12}$					

当$t=2$时：已知$x_2 = apple$，所以

迭代 $V_t^k = p(x_t|y_t^k = 1)\max_i a_{i,k} V_{t-1}^i$

$$V_2^1 = b_{1,1} \max_{\{y_1^i\}} a_{i,1} V_1^i = b_{1,1} \max \left(\frac{1}{3}\times\frac{1}{6}, \frac{1}{3}\times\frac{1}{4}, \frac{1}{3}\times\frac{1}{12}\right) = \frac{1}{2}\times\frac{1}{12} = \frac{1}{24}$$

	t=1	t=2	t=3	t=4	t=5	
V_t^1	$\frac{1}{6}$	$\frac{1}{24}$				
V_t^2	$\frac{1}{4}$					
V_t^3	$\frac{1}{12}$					

当$t=2$时：已知$x_2 = apple$，所以

迭代 $V_t^k = p(x_t|y_t^k = 1)\max_i a_{i,k} V_{t-1}^i$

$$V_2^1 = b_{1,1} \max_{\{y_1^i\}} a_{i,1} V_1^i = b_{1,1} \max \left(\frac{1}{3}\times\frac{1}{6}, \frac{1}{3}\times\frac{1}{4}, \frac{1}{3}\times\frac{1}{12}\right) = \frac{1}{2}\times\frac{1}{12} = \frac{1}{24}$$

$$V_2^2 = b_{2,1} \max \left(\frac{1}{3}\times\frac{1}{6}, \frac{1}{3}\times\frac{1}{4}, \frac{1}{3}\times\frac{1}{12}\right) = \frac{3}{4}\times\frac{1}{12} = \frac{1}{16}$$

	t=1	t=2	t=3	t=4	t=5	
V_t^1	$\frac{1}{6}$	$\frac{1}{24}$				
V_t^2	$\frac{1}{4}$	$\frac{1}{16}$				
V_t^3	$\frac{1}{12}$					

当$t=2$时：已知$x_2 = apple$，所以

迭代 $V_t^k = p(x_t|y_t^k = 1)\max_i a_{i,k} V_{t-1}^i$

$$V_2^1 = b_{1,1} \max_{\{y_1^i\}} a_{i,1} V_1^i = b_{1,1} \max \left(\frac{1}{3}\times\frac{1}{6}, \frac{1}{3}\times\frac{1}{4}, \frac{1}{3}\times\frac{1}{12}\right) = \frac{1}{2}\times\frac{1}{12} = \frac{1}{24}$$

$$V_2^2 = b_{2,1} \max \left(\frac{1}{3}\times\frac{1}{6}, \frac{1}{3}\times\frac{1}{4}, \frac{1}{3}\times\frac{1}{12}\right) = \frac{3}{4}\times\frac{1}{12} = \frac{1}{16}$$

$$V_2^3 = b_{3,1} \max \left(\frac{1}{3}\times\frac{1}{6}, \frac{1}{3}\times\frac{1}{4}, \frac{1}{3}\times\frac{1}{12}\right) = \frac{1}{4}\times\frac{1}{12} = \frac{1}{48}$$

	t=1	t=2	t=3	t=4	t=5	
V_t^1	$\frac{1}{6}$	$\frac{1}{24}$				
V_t^2	$\frac{1}{4}$	$\frac{1}{16}$				
V_t^3	$\frac{1}{12}$	$\frac{1}{48}$				

当$t=3$时：已知$x_3 = orange$，所以

迭代

$$V_3^1 = b_{1,2} \max_{\{y_2^i\}} a_{i,1} V_2^i = b_{1,2} \max \left(\frac{1}{24}\times\frac{1}{3}, \frac{1}{16}\times\frac{1}{3}, \frac{1}{48}\times\frac{1}{3}\right) = \frac{1}{2}\times\frac{1}{48} = \frac{1}{96}$$

$$V_3^2 = b_{2,2} \max_{\{y_2^i\}} a_{i,2} V_2^i = b_{2,2} \max \left(\frac{1}{24}\times\frac{1}{3}, \frac{1}{16}\times\frac{1}{3}, \frac{1}{48}\times\frac{1}{3}\right) = \frac{1}{4}\times\frac{1}{48} = \frac{1}{192}$$

$$V_3^3 = b_{3,2} \max_{\{y_2^i\}} a_{i,3} V_2^i = b_{3,2} \max \left(\frac{1}{24}\times\frac{1}{3}, \frac{1}{16}\times\frac{1}{3}, \frac{1}{48}\times\frac{1}{3}\right) = \frac{3}{4}\times\frac{1}{48} = \frac{1}{64}$$

当$t=4$时：已知$x_4 = orange$，所以

迭代

$$V_4^1 = b_{1,2} \max_{\{y_3^i\}} a_{i,1} V_3^i = \frac{1}{2} \times \max\left(\frac{1}{96} \times \frac{1}{3}, \frac{1}{192} \times \frac{1}{3}, \frac{1}{64} \times \frac{1}{3}\right) = \frac{1}{384}$$

$$V_4^2 = b_{2,2} \max_{\{y_2^i\}} a_{i,2} V_3^i = \frac{1}{4} \times \max\left(\frac{1}{96} \times \frac{1}{3}, \frac{1}{192} \times \frac{1}{3}, \frac{1}{64} \times \frac{1}{3}\right) = \frac{1}{768}$$

$$V_4^3 = b_{3,2} \max_{\{y_2^i\}} a_{i,3} V_3^i = \frac{3}{4} \times \max\left(\frac{1}{96} \times \frac{1}{3}, \frac{1}{192} \times \frac{1}{3}, \frac{1}{64} \times \frac{1}{3}\right) = \frac{1}{256}$$

当$t=5$时：已知$x_5 = orange$，所以

迭代

$$V_5^1 = b_{1,2} \max_{\{y_4^i\}} a_{i,1} V_4^i = \frac{1}{2} \times \max\left(\frac{1}{384} \times \frac{1}{3}, \frac{1}{768} \times \frac{1}{3}, \frac{1}{256} \times \frac{1}{3}\right) = \frac{1}{6*256} = \frac{1}{1536}$$

$$V_5^2 = b_{2,2} \max_{\{y_4^i\}} a_{i,2} V_4^i = \frac{1}{4} \times \max\left(\frac{1}{384} \times \frac{1}{3}, \frac{1}{768} \times \frac{1}{3}, \frac{1}{256} \times \frac{1}{3}\right) = \frac{1}{12*256} = \frac{1}{3072}$$

$$V_5^3 = b_{3,2} \max_{\{y_4^i\}} a_{i,3} V_4^i = \frac{3}{4} \times \max\left(\frac{1}{384} \times \frac{1}{3}, \frac{1}{768} \times \frac{1}{3}, \frac{1}{256} \times \frac{1}{3}\right) = \frac{1}{4*256} = \frac{1}{1024}$$

终止：T=5

$$P^* = \max_{1 \le i \le N} V_T^i = V_5^3 = \frac{1}{1024}$$

最优路径回溯：$\boldsymbol{y} = \{2,\ 2,\ 3,\ 3,\ 3\}$。

	t=1	t=2	t=3	t=4	t=5	
V_t^1	$\frac{1}{6}$	$\frac{1}{24}$	$\frac{1}{96}$	$\frac{1}{384}$	$\frac{1}{1536}$	
V_t^2	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{192}$	$\frac{1}{768}$	$\frac{1}{3072}$	
V_t^3	$\frac{1}{12}$	$\frac{1}{48}$	$\frac{1}{64}$	$\frac{1}{256}$	$\frac{1}{1024}$	

2. 学习: 参数估计

- HMM的参数θ 包括：初始概率$\boldsymbol{\pi}$，转移矩阵$\mathbf{A}$, 发射矩阵B
- 给定观测输出x，更新模型的参数$\boldsymbol{\theta}$= (π，A，B)，以最大化$p(\mathbf{x}|\boldsymbol{\theta})$---极大似然估计
  - 双重随机过程
    - 例子：不同骰子（y）密度不一样，投出具体点数（x）的概率（发射概率）就不一样。玩家改变骰子就是改变隐变量y_t状态，投一次骰子就是获取一次观测样本x_t
  - 监督学习：
    - 玩家投了10000次骰子，可以观察到玩家什么时候换哪个骰子，可以知道每次骰子投掷的结果
  - 无监督学习：
    - 只能观察到玩家投了10000次骰子，不知道玩家怎么换骰子，不知道每次是哪个骰子投掷的结果

2. 学习: 参数估计

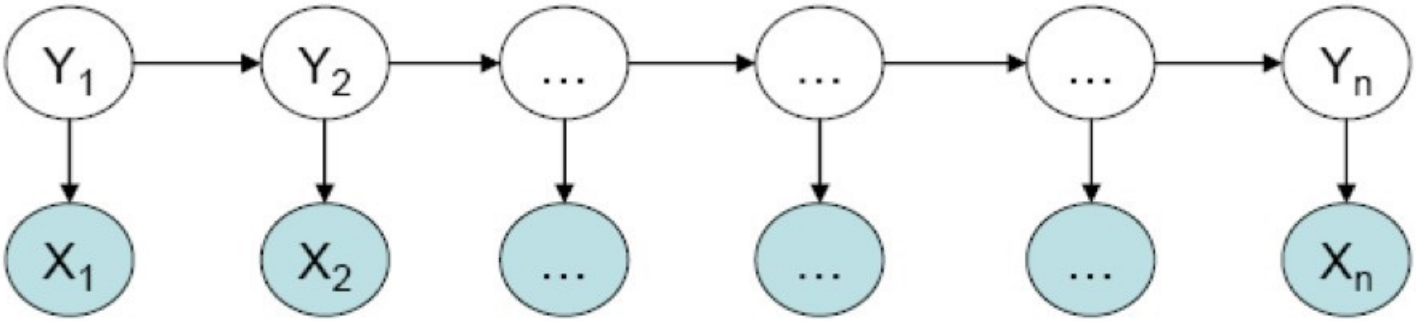
■ 对数似然：

$$p(\mathbf{x}|\boldsymbol{\theta}) = log \sum_{y_1} \sum_{y_2} \dots \sum_{y_T} \pi(y_1) \prod_{t=1}^{T-1} a_{y_t,y_{t+1}} \prod_{t=1}^{T} b_{y_t,x_t}$$

- 其中$b_{y_t,x_t} = p(x_t|y_t, \eta)$ 是输出分布
- 我们的目标是针对 θ 最大化表达

■ 将状态变量视作隐变量，采用EM算法估计HMM参数：Baum-Weich算法（课后阅读《统计学习方法》）

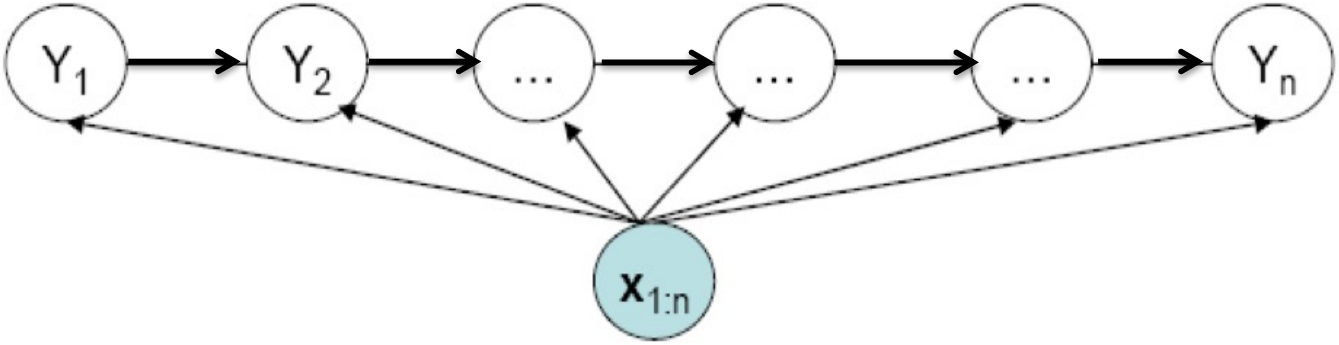
HMM缺点



- HMM模型仅捕捉了状态之间和状态及其对应输出之间的关系
  - NLP 例子：在一个句子分割的任务中，每一个分割状态可能不仅依赖单个词（ 和邻近的分割状态），还依赖于非局部特征，如整个长度和缩进，空格的数量等
- 学习目标和预测目标不匹配
  - HMM学习状态和观察的联合概率分布$P(Y,X)$
  - 但是在预测任务中，我们仅需要条件概率 $P(Y|X)$

从HMM 到MEMM

■ Maximum Entropy Markov Model（MEMM）

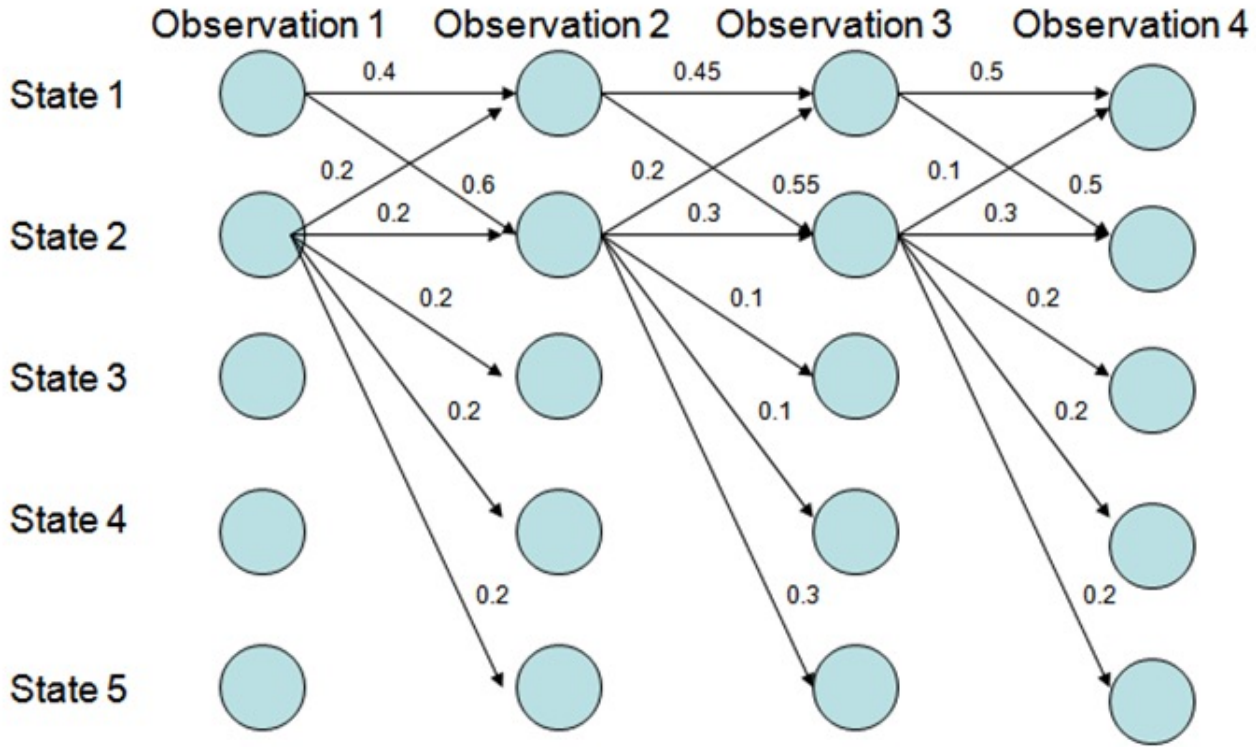


$$P(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}) = \prod_{i=1}^{n} P(y_i|y_{i-1}, x_{1:n})$$

■ MEMM是一个判别式模型（ discriminative model ）

- 建模了每个状态和整个观测序列的依赖

Label Bias 问题

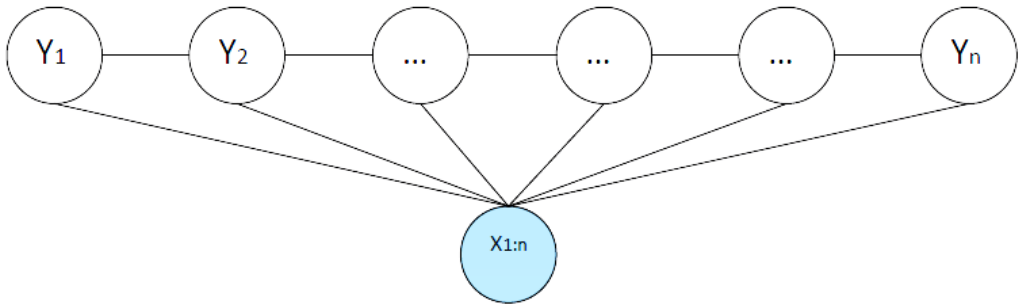


- 局部概率角度，从State1开始，倾向于一直留在State2？
- 1->2->2->2？ P=0.6*0.3*0.3 =0.054
- 1->1->1->1？ P=0.4*0.45*0.5=0.09
- 整体看，留在State1分支少，概率更大

从HMM 到 CRF

■ 条件随机场（CRF）：无向图

- Local Probability → Local Potential
- Global Nomalizor $Z(X_{1:n})$
- 判别式



$$P(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}) = \frac{1}{Z(\mathbf{x}_{1:n})}\prod_{i=1}^{n} P(y_i, y_{i-1}, x_{1:N})$$

$$= \frac{1}{Z(\mathbf{x}_{1:n})}\prod_{i=1}^{n} \exp(\mathbf{w}^T \boldsymbol{f}(y_i, y_{i-1}, \mathbf{x}_{1:n}))$$

课后补充：

ICML 2011 Test of Time Award for the 10 Year Best Paper

Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data

John Lafferty[†*] LAFFERTY@CS.CMU.EDU
Andrew McCallum[*†] MCCALLUM@WHIZBANG.COM
Fernando Pereira[*‡] FPEREIRA@WHIZBANG.COM

[*]WhizBang! Labs–Research, 4616 Henry Street, Pittsburgh, PA 15213 USA
[†]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA
[‡]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104 USA

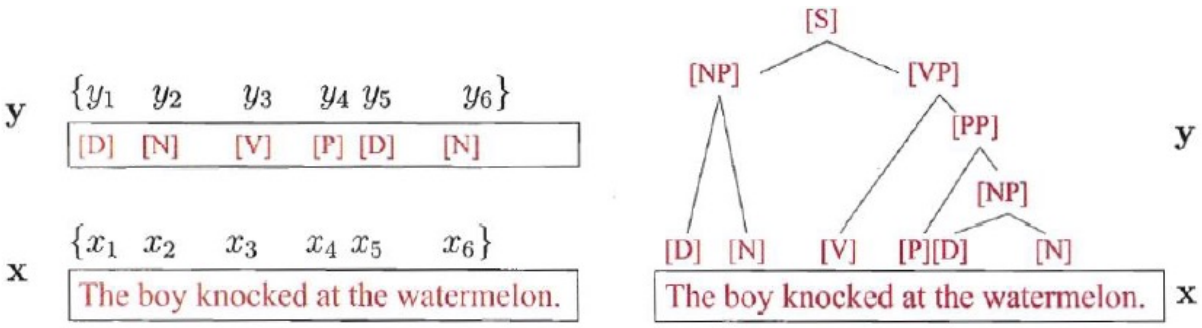
Abstract

We present *conditional random fields*, a framework for building probabilistic models to segment and label sequence data. Conditional random fields offer several advantages over hidden Markov models and stochastic grammars for such tasks, including the ability to relax strong independence assumptions made in those models. Conditional random fields also avoid a fundamental limitation of maximum entropy Markov models (MEMMs) and other discriminative Markov models based on directed graphical models, which can be biased towards states with few successor states. We present iterative mize the joint likelihood of training examples. To define a joint probability over observation and label sequences, a generative model needs to enumerate all possible observation sequences, typically requiring a representation in which observations are task-appropriate atomic entities, such as words or nucleotides. In particular, it is not practical to represent multiple interacting features or long-range dependencies of the observations, since the inference problem for such models is intractable.

This difficulty is one of the main motivations for looking at conditional models as an alternative. A conditional model specifies the probabilities of possible label sequences given an observation sequence. Therefore, it does not expend modeling effort on the observations, which at test time

例：词性标注

- **一些经验：**
- 使用相同的特征集合：HMM ≈ CRF > MEMM
- 使用额外的重叠特征：CRF+ > MEMM+ >>HMM



(a) 词性标注　　(b) 语法分析

自然语言处理中的词性标注和语法分析任务

model	*error*	*oov error*
HMM	5.69%	45.99%
MEMM	6.37%	54.61%
CRF	5.55%	48.05%
MEMM+	4.81%	26.99%
CRF+	4.27%	23.76%

+Using spelling features

小结

- 概率图模型是自然/完美的工具，针对：
  - 表示(数据结构)和
  - 推断(算法)
- 两种类型的概率图模型
  - 贝叶斯网络
  - 马尔科夫随机场
- 典型的概率图模型
  - 隐马尔科夫是有向的生成模型
  - 条件随机场是无向的判别模型

参考文献

- Koller D, Friedman N. Probabilistic graphical models: principles and techniques. MIT press, 2009
- Michael I. Jordan 《An Introduction to Graphical Models》
- 周志华，《机器学习》
- Christopher Bishop 《Pattern Recognition and Machine Learning》
- J. Lafferty, A. McCallum and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. ICML 2001 （10 year best paper）
- Carlos Guestrin, Ben Taskar and Daphne Koller. Max-margin Markov Networks. NIPS 2003.
- Eric P. Xing Graphical models Course

作业

- **1、假设我们要采用HMM实现一个英文的词性标注系统,系统中共有20种词性,则状态转移矩阵B的大小为( )**

  A、 20

  B、 40

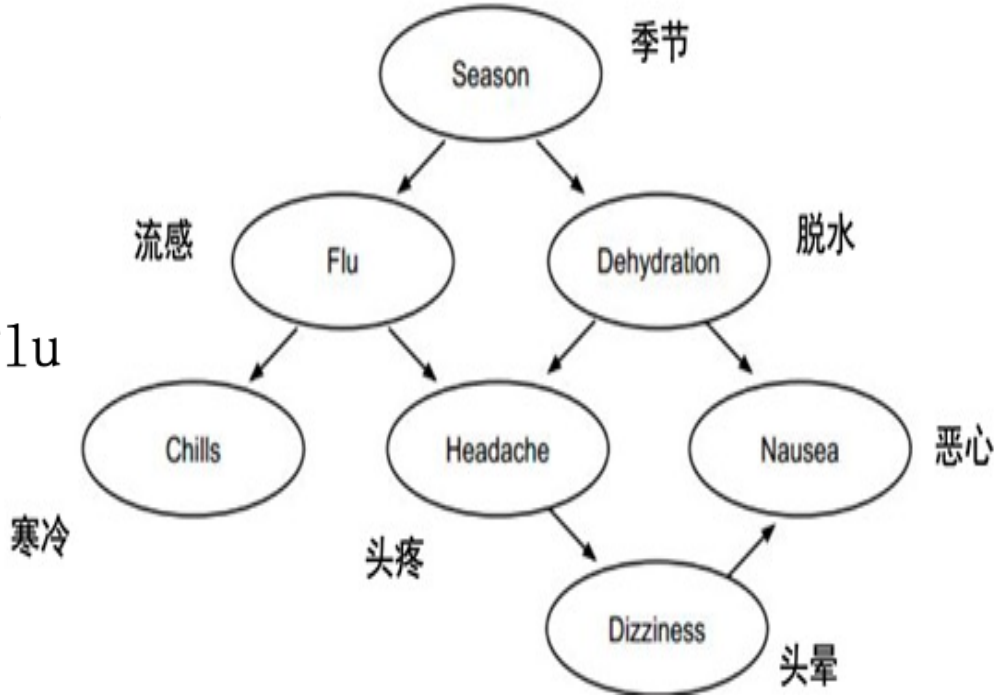
  C、 400

- **2. 已知以下贝叶斯网络,包含 7 个变量,即 Season (季节), Flu (流感), Dehydration (脱水), Chills (发 冷), Headache (头疼), Nausea (恶心), Dizziness (头晕),则下列(条件)独立成立的是( )**

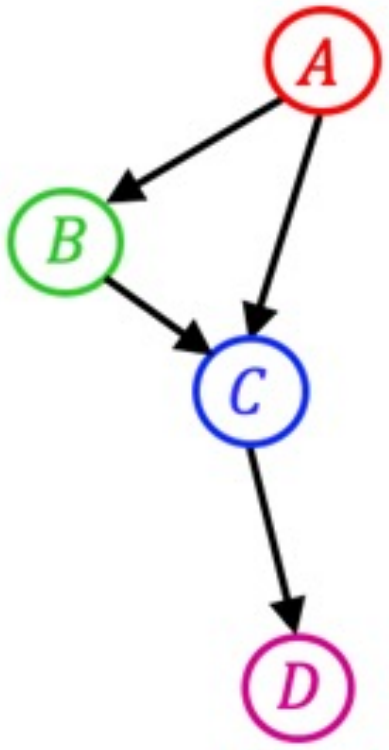
A、 Season ⊥ Chills | Flu

B、 Season ⊥ Chills

C、 Season ⊥ Headache | Flu



- **3. 已知以下贝叶斯网络,包含 4个二值变量,则该网络一共有( )个参数**



A、 4

B、 8

C、 9

D、 16

4. 假设你有三个盒子，每个盒子里都有一定数量的苹果和桔子。每次随机选择一个盒子，然后从盒子里选一个水果，并记录你的发现（a代表苹果，o代表橘子）。不幸的是，你忘了写下你所选的盒子，只是简单地记下了苹果和桔子。假设每个盒子中水果数量如下：

- 盒子一：2 个苹果，2 个桔子
- 盒子二 ：3 个苹果，1 个桔子
- 盒子三：1 个苹果，3 个桔子。

（1） 请给出 HMM 模型；

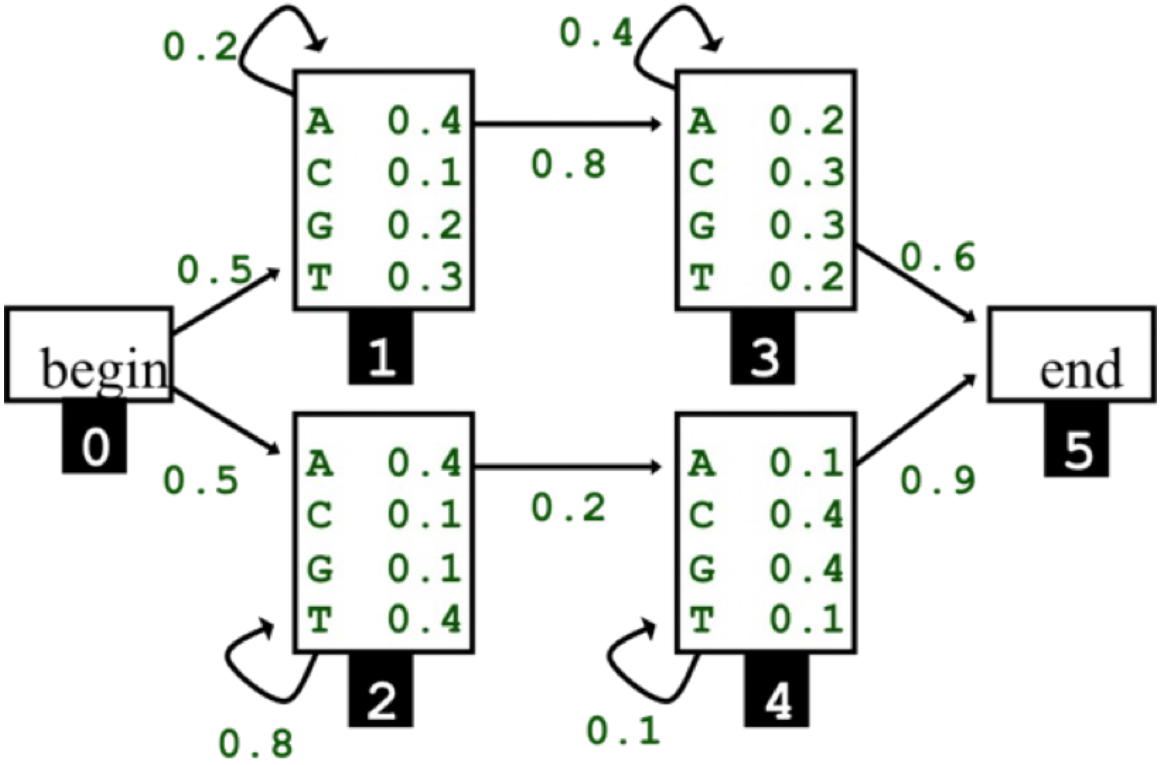
（2） 请给出水果序列 $\boldsymbol{x} = (a, a, o, o, o)$对应的最佳盒子序列。

提示：

盒子视为隐变量，拿出来的水果视为观测变量；

观测到了序列$x_1 \ldots x_t$，推断最可能的隐状态序列$y_1 \ldots y_t$，用Viterbi解码：
前向递推最佳状态的概率，回溯获得最佳路径（状态序列）

- 5. 给定如图所示HMM



(1) 采用前向算法计算序列 AGTT 出现的概率。

(2) 计算观测 TATA 最可能的状态序列。

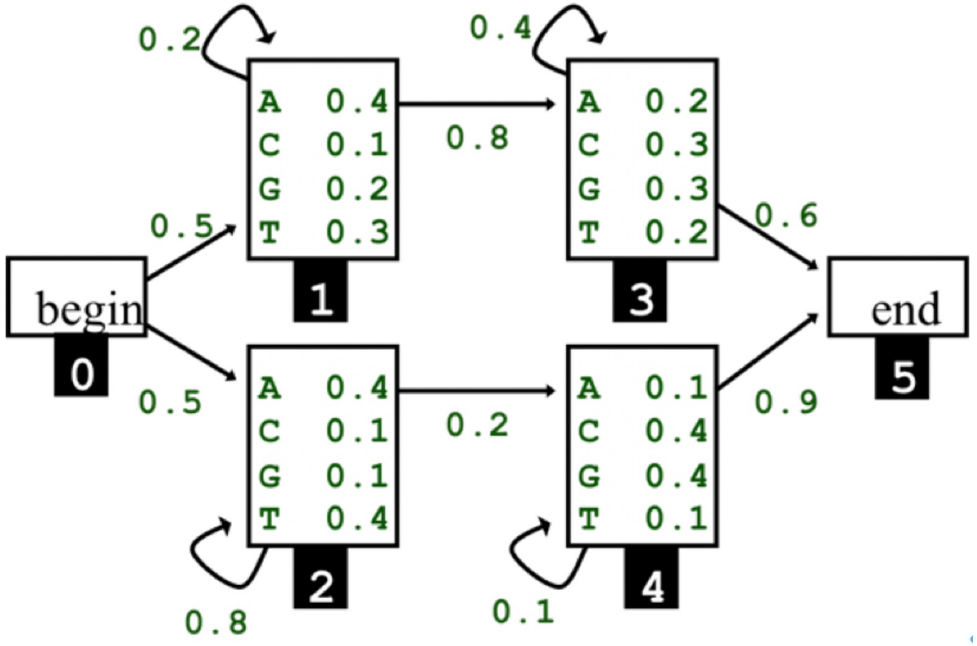
提示：

初始概率$\boldsymbol{\pi} = (1,0,0,0,0,0)$，

转移概率矩阵$\mathbf{A} = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0.2 & 06 \\ 0 & 0 & 0 & 0.4 & 0 & 0.6 \\ 0 & 0 & 0 & 0 & 0.1 & 0.9 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$，

（去掉状态 0/begin 和状态 5/end ）

发放概率矩阵$\mathbf{B} = \begin{bmatrix} 0.4 & 0.1 & 0.2 & 0.3 \\ 0.4 & 0.1 & 0.1 & 0.4 \\ 0.2 & 0.3 & 0.3 & 0.2 \\ 0.1 & 0.4 & 0.4 & 0.4 \end{bmatrix}$



- END