机器学习

Machine learning

第九章 概率图模型

Probabilistic Graphical Model

授课人：周晓飞
zhouxiaofei@iie.ac.cn

2021-12-11

第九章 概率图模型

近似推断

推断方法

精确推断：

变量消去、信念传播。

计算复杂度随着极大团规模的增长呈指数级增长，适用范围有限。

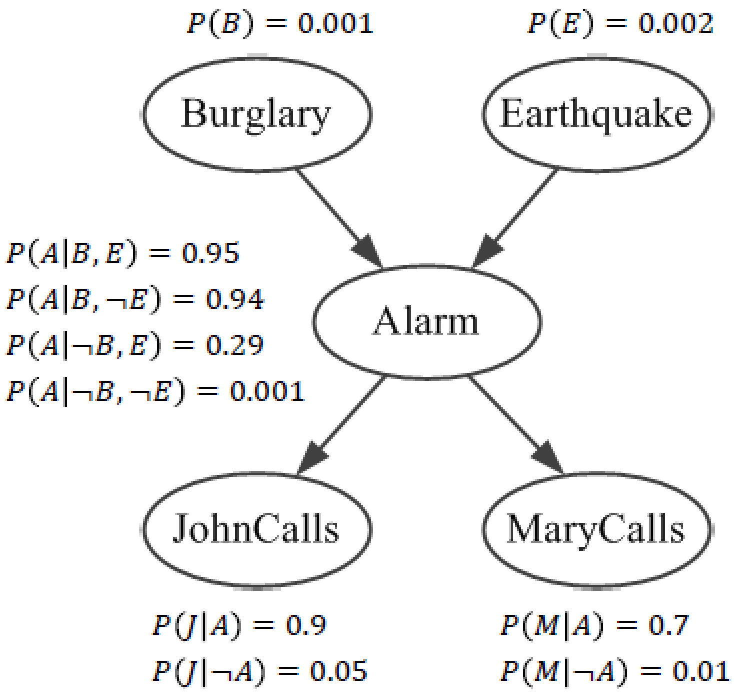
近似推断：

前向采样、吉布斯采样。

通过采样一组服从特定分布的样本，来近似原始分布，适用范围更广，操作性更强。

前向采样

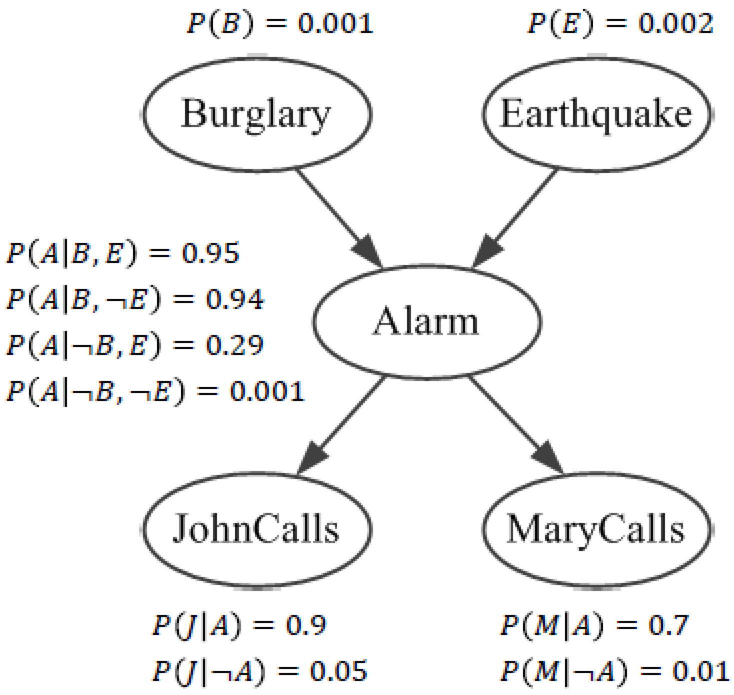
思路：依据贝叶斯网络的（条件）概率直接采样。



$$P(B=1 \mid E=0, J=1) = ?$$

前向采样

思路：依据贝叶斯网络的（条件）概率直接采样。



$$P(B=1|E=0, J=1)=?$$

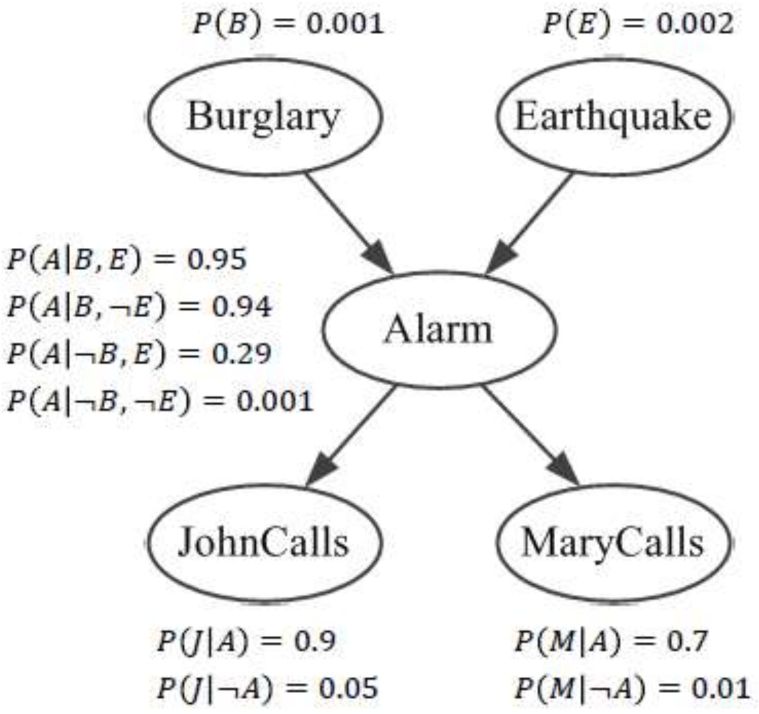
$$
\begin{aligned}
&P(B=1|E=0,J=1)\\
&=\frac{P(B=1,E=0,J=1)}{P(E=0,J=1)}\\
&=\frac{\sum_{A,M} P(B=1,E=0,A,J=1,M)}{\sum_{B,A,M} P(B,E=0,A,J=1,M)}\\
&=\cdots
\end{aligned}
$$

采样后，进行需要的概率统计！

近似推断

前向采样

思路：依据贝叶斯网络的（条件）概率直接采样。

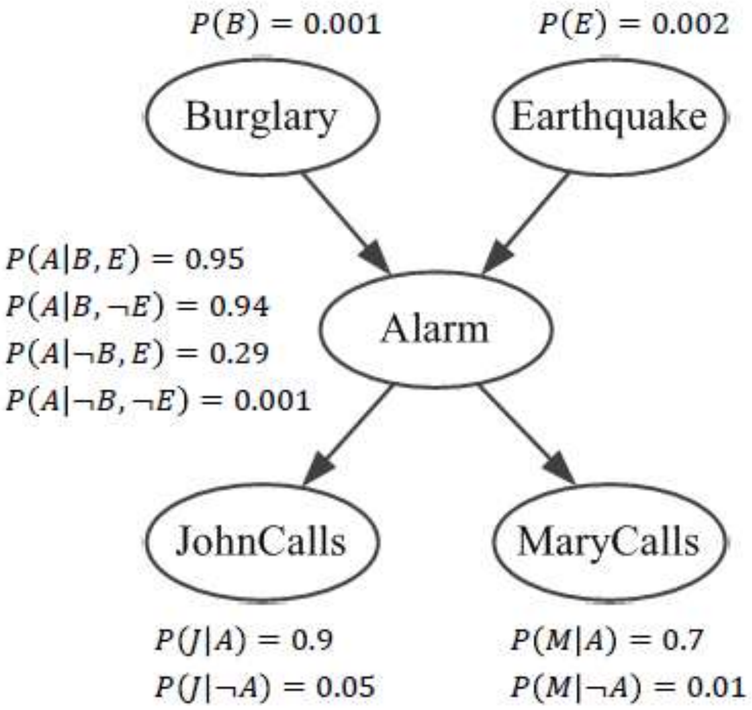


- 采样 $B \sim \mathrm{Ber}(0.001)$
  - 采样 $r \sim \mathrm{U}(0,1)$：若$r < 0.001$，$B = 1$；否则 $B = 0$

近似推断

前向采样

思路：依据贝叶斯网络的（条件）概率直接采样。

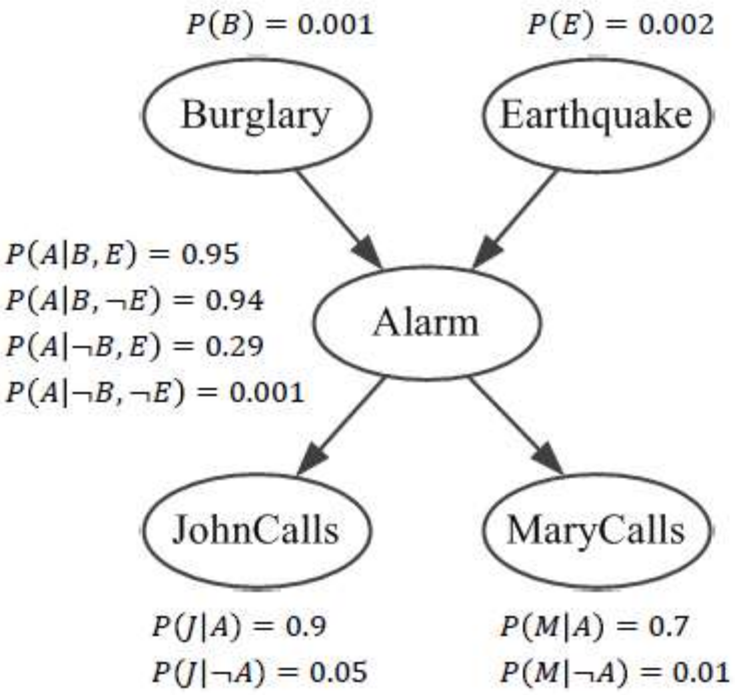


- 采样 $B \sim \mathrm{Ber}(0.001)$
- 采样 $E \sim \mathrm{Ber}(0.002)$

近似推断

前向采样

思路：依据贝叶斯网络的（条件）概率直接采样。



- 采样 $B \sim \mathrm{Ber}(0.001)$
- 采样 $E \sim \mathrm{Ber}(0.002)$
- 若 $B=1, E=1$，采样 $A \sim \mathrm{Ber}(0.95)$

  若 $B=1, E=0$，采样 $A \sim \mathrm{Ber}(0.94)$

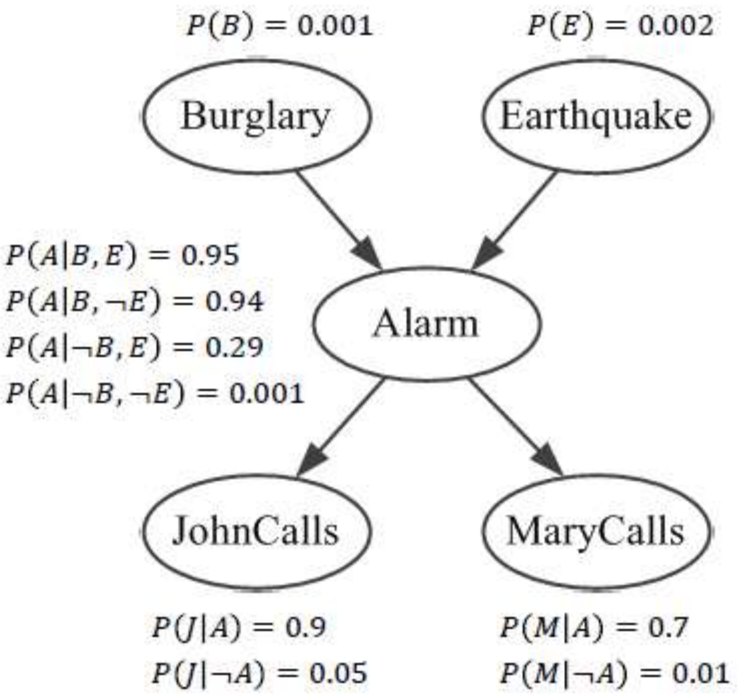
  若 $B=0, E=1$，采样 $A \sim \mathrm{Ber}(0.29)$

  若 $B=0, E=0$，采样 $A \sim \mathrm{Ber}(0.001)$

近似推断

前向采样

思路：依据贝叶斯网络的（条件）概率直接采样。

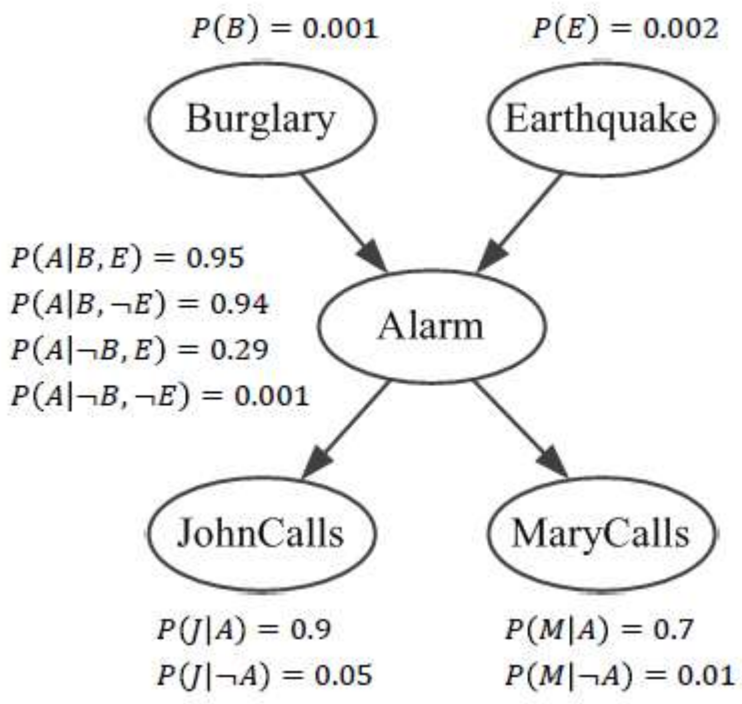


- 采样 $B \sim \mathrm{Ber}(0.001)$
- 采样 $E \sim \mathrm{Ber}(0.002)$
- 若 $B = 1, E = 1$，采样 $A \sim \mathrm{Ber}(0.95)$
  若 $B = 1, E = 0$，采样 $A \sim \mathrm{Ber}(0.94)$
  若 $B = 0, E = 1$，采样 $A \sim \mathrm{Ber}(0.29)$
  若 $B = 0, E = 0$，采样 $A \sim \mathrm{Ber}(0.001)$
- 若 $A = 1$，采样 $J \sim \mathrm{Ber}(0.9)$；否则采样 $J \sim \mathrm{Ber}(0.05)$

实例模型

前向采样

思路：依据贝叶斯网络的（条件）概率直接采样。



- 采样 $B \sim \mathrm{Ber}(0.001)$
- 采样 $E \sim \mathrm{Ber}(0.002)$
- 若 $B = 1, E = 1$，采样 $A \sim \mathrm{Ber}(0.95)$

  若 $B = 1, E = 0$，采样 $A \sim \mathrm{Ber}(0.94)$

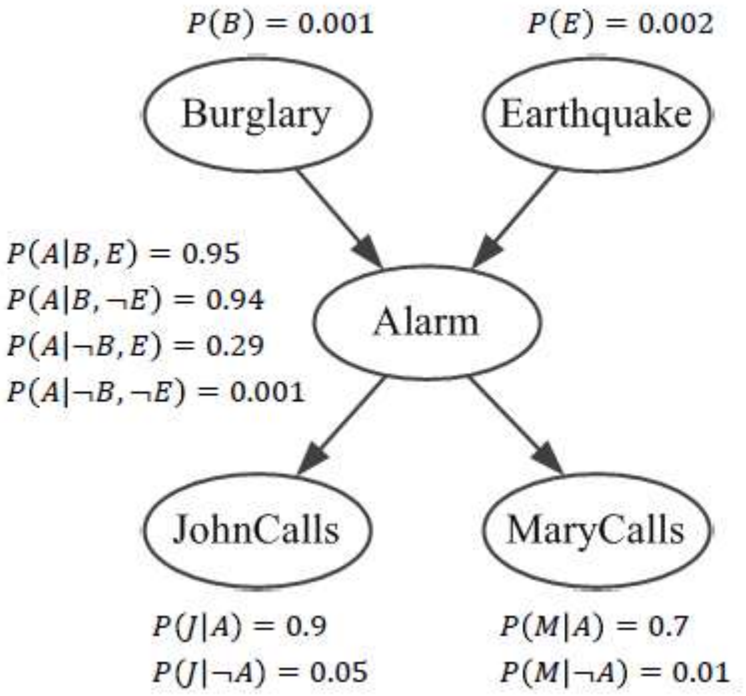
  若 $B = 0, E = 1$，采样 $A \sim \mathrm{Ber}(0.29)$

  若 $B = 0, E = 0$，采样 $A \sim \mathrm{Ber}(0.001)$
- 若 $A = 1$，采样 $J \sim \mathrm{Ber}(0.9)$；否则采样 $J \sim \mathrm{Ber}(0.05)$
- 若 $A = 1$，采样 $M \sim \mathrm{Ber}(0.7)$；否则采样 $M \sim \mathrm{Ber}(0.01)$

实例模型

前向采样

思路：依据贝叶斯网络的（条件）概率直接采样。

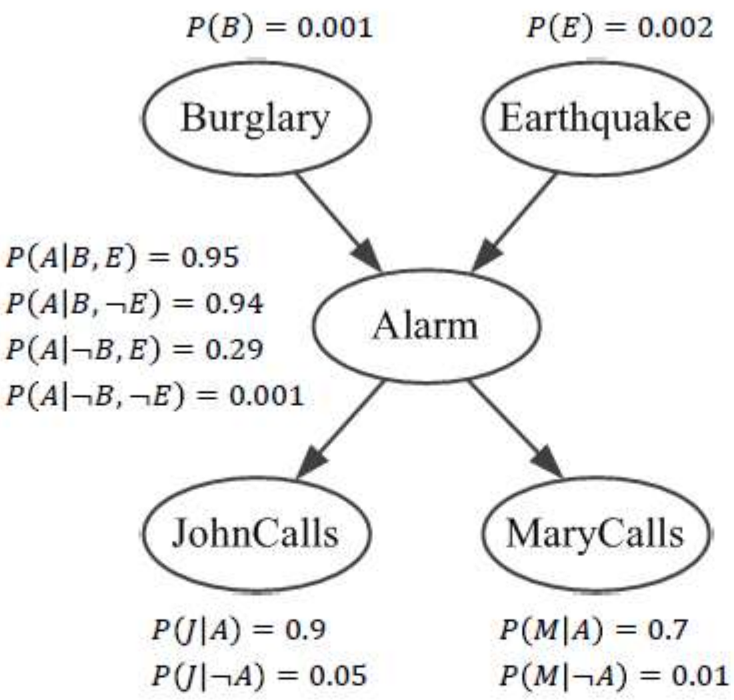


B	E	A	J	M
0	0	0	0	0
1	0	1	1	0
0	0	0	0	1
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
1	0	1	1	1
0	0	0	0	0
0	0	1	1	0
0	0	0	0	0

实例模型

前向采样

思路：依据贝叶斯网络的（条件）概率直接采样。



B	E	A	J	M
0	0	0	0	0
1	0	1	1	0
0	0	0	0	1
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
1	0	1	1	1
0	0	0	0	0
0	0	1	1	0
0	0	0	0	0

$$P(B=1|E=0,J=1) = {}^{2}/_{3}$$

实例模型

前向采样

思路：依据贝叶斯网络的（条件）概率直接采样。

缺点：I) 对于小概率事件采样困难，可能经过很多次采样也无法获得足够多的样本；

II) 仅适用于贝叶斯网络，不适用于马尔可夫随机场。

吉布斯采样

思路：直接依照条件概率 $P(\mathbf{x}_Q|\mathbf{x}_E)$采样。

Procedure Gibbs Sampling

Fix $\mathbf{x}_E$ and randomly initialize other variables $\mathbf{x}_O := \{x_1, \cdots, x_n\} \backslash \mathbf{x}_E$

repeat

  foreach $x_i \in \mathbf{x}_O$ do

    Sample $x \sim P(x_i|x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$ /* x_i 在其他所有变量当前取值下的条件概率 */

    $x_i \leftarrow x$

until Convergence

return The later samples

实例模型

吉布斯采样

思路： 直接依照条件概率 $P(x_Q|x_E)$采样。

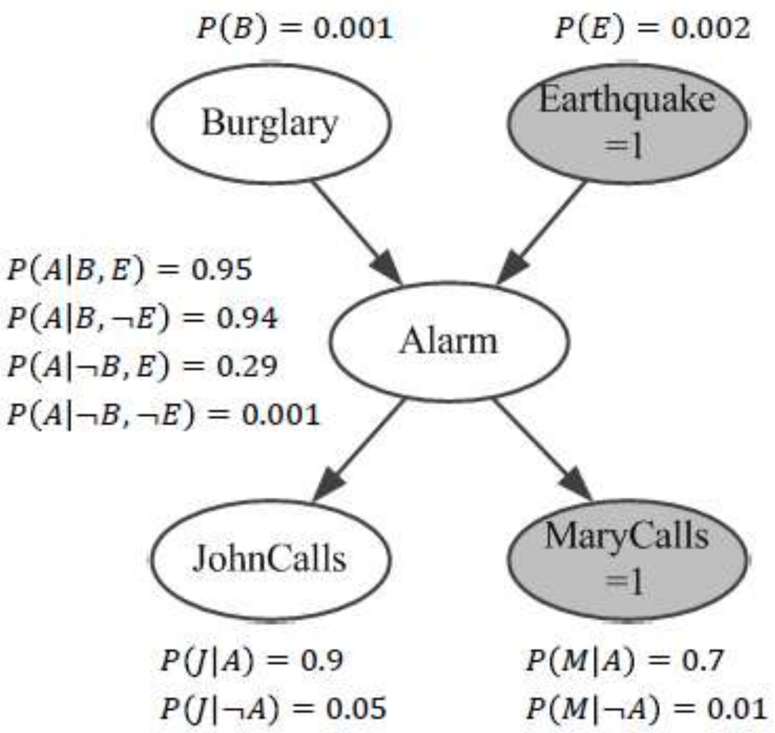
马尔可夫毯的性质：

$$P(x_i \mid x_1, \ldots x_{i-1}, x_{i-1}, \ldots, x_n) = P(x_i \mid MB(x_i))$$

实例模型

吉布斯采样

思路：直接依照条件概率 $P(x_Q|x_E)$采样。

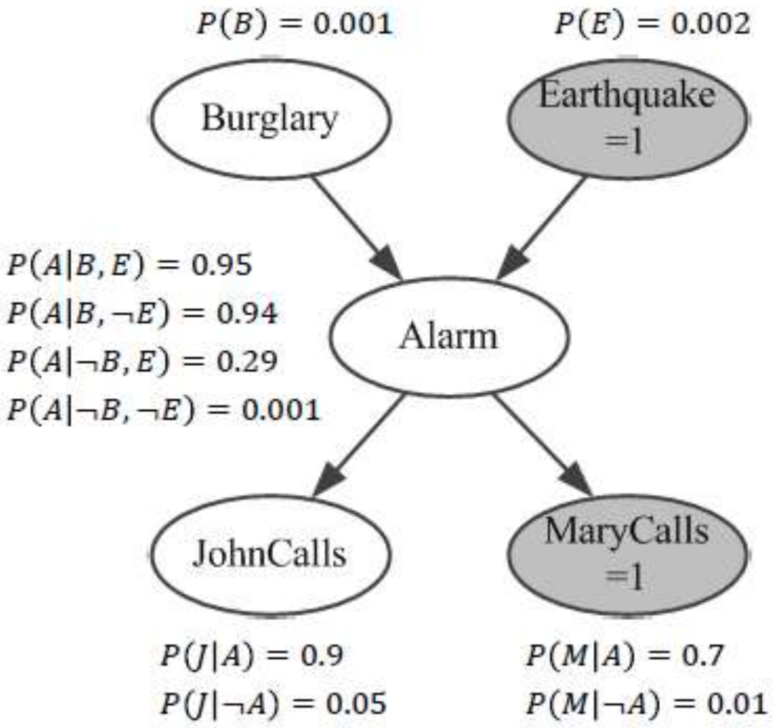


$$P(B=1|E=1,M=1)=?$$

实例模型

吉布斯采样

思路：直接依照条件概率 $P(x_Q|x_E)$采样。

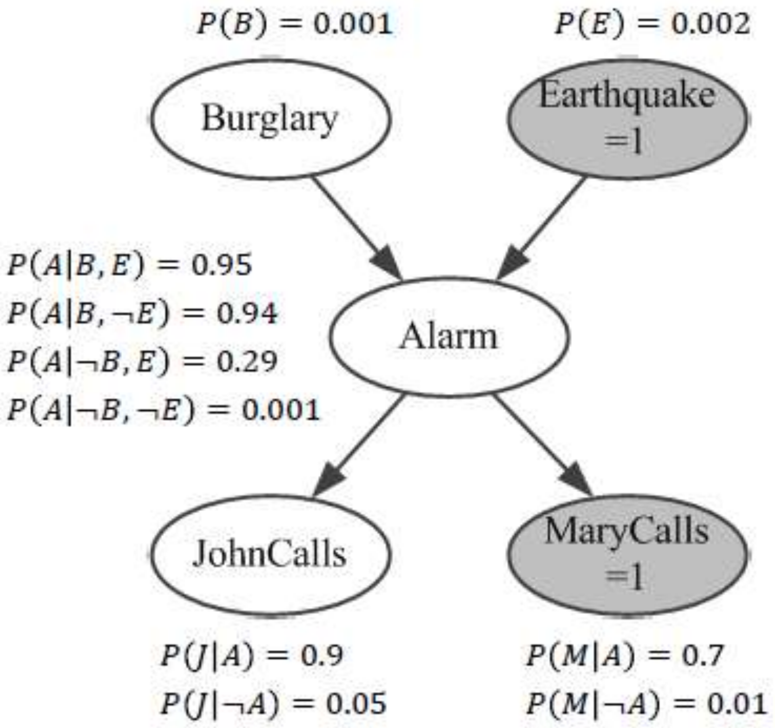


Iteration	B	E	A	J	M
$t=0$	0	1	0	0	1

实例模型

吉布斯采样

思路：直接依照条件概率 $P(x_Q|x_E)$采样。



Iteration	B	E	A	J	M
$t=0$	0	1	0	0	1
$t=1$					

- 采样 B，马尔可夫毯 $\{A, E\}$

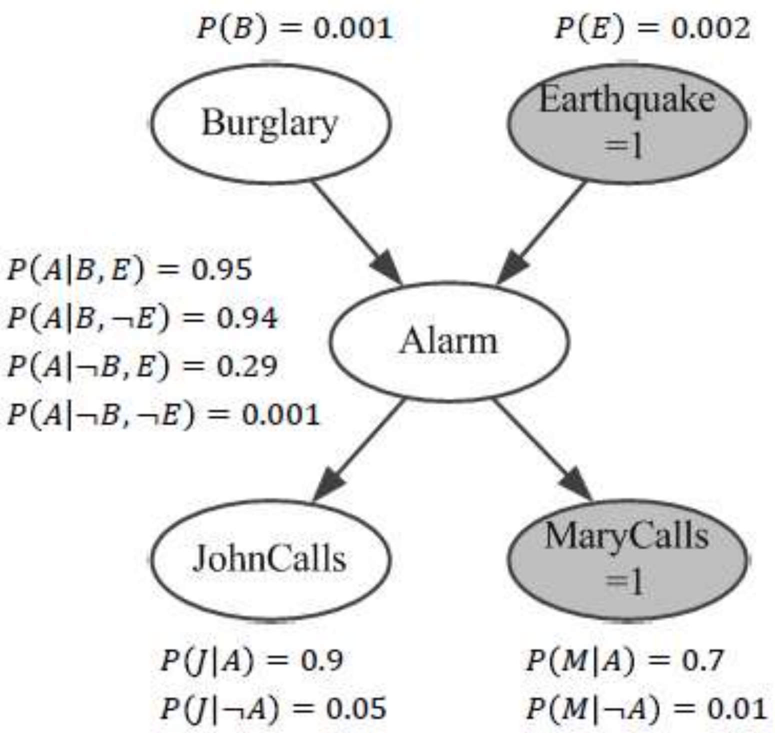
$$P(B|\neg A, E) \propto P(B)P(\neg A|B, E) = 0.001 \times 0.05$$

$$P(\neg B|\neg A, E) \propto P(\neg B)P(\neg A|\neg B, E) = 0.999 \times 0.71$$

实例模型

吉布斯采样

思路：直接依照条件概率 $P(x_Q|x_E)$采样。



Iteration	B	E	A	J	M
$t=0$	0	1	0	0	1
$t=1$	0				

- 采样 B，马尔可夫毯 $\{A,E\}$

  $P(B|\neg A,E) \propto P(B)P(\neg A|B,E) = 0.001 \times 0.05$

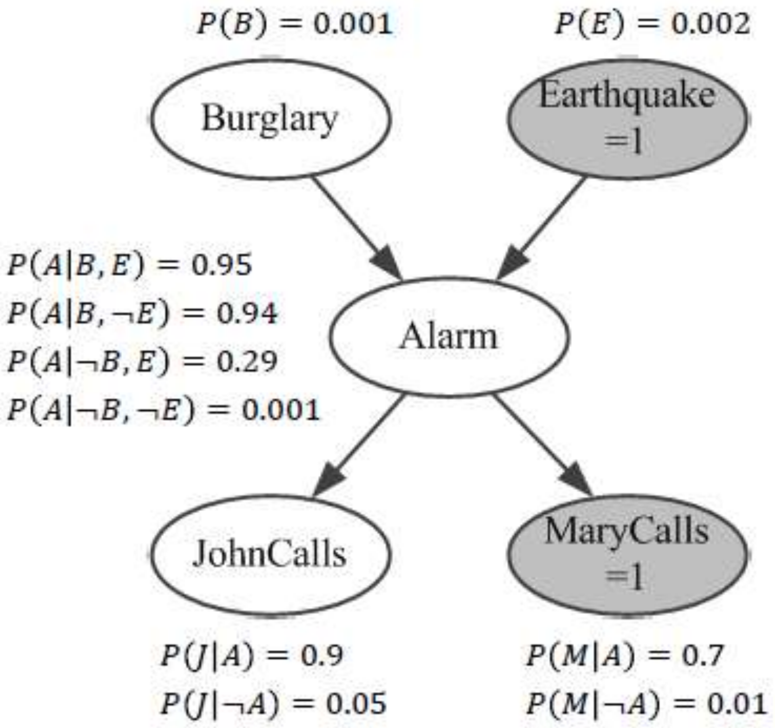
  $P(\neg B|\neg A,E) \propto P(\neg B)P(\neg A|\neg B,E) = 0.999 \times 0.71$

- 采得 $B=0$

实例模型

吉布斯采样

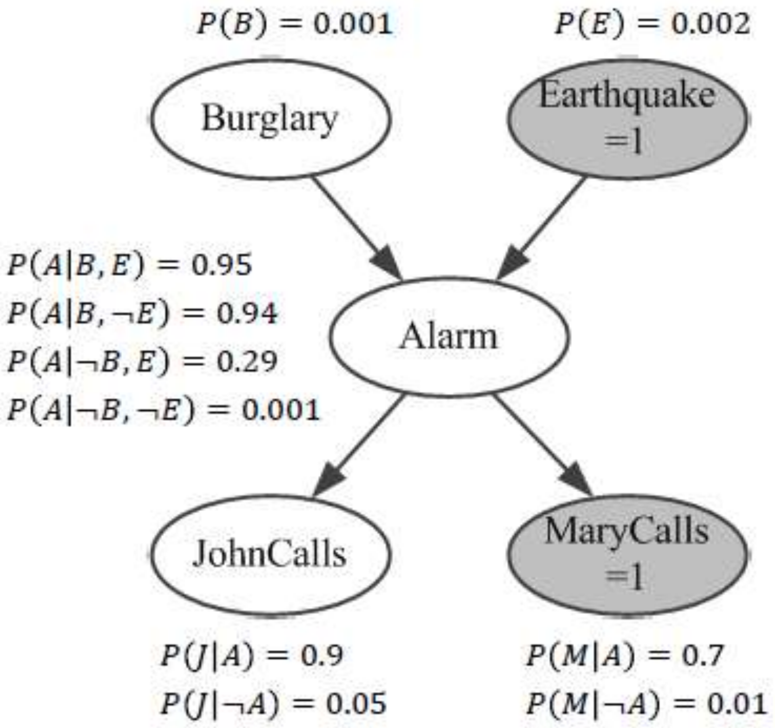
思路：直接依照条件概率 $P(x_Q|x_E)$采样。



Iteration	B	E	A	J	M
$t=0$	0	1	0	0	1
$t=1$	0	1			

吉布斯采样

思路：直接依照条件概率 $P(x_Q|x_E)$采样。



Iteration	B	E	A	J	M
$t=0$	0	1	0	0	1
$t=1$	0	1			

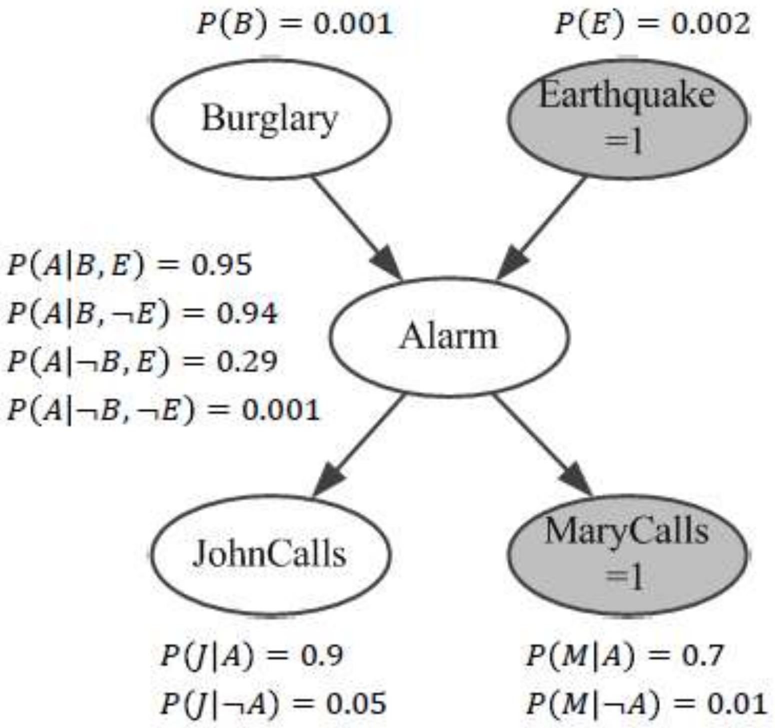
- 采样 A，马尔可夫毯 $\{B, E, J, M\}$

$$P(A|\neg B, E, \neg J, M) \propto P(A|\neg B, E)P(\neg J|A)P(M|A)$$
$$= 0.29 \times 0.1 \times 0.7$$
$$P(\neg A|\neg B, E, \neg J, M) \propto P(\neg A|\neg B, E)P(\neg J|\neg A)P(M|\neg A)$$
$$= 0.71 \times 0.95 \times 0.01$$

实例模型

吉布斯采样

思路：直接依照条件概率 $P(x_Q|x_E)$采样。



Iteration	B	E	A	J	M
$t=0$	0	1	0	0	1
$t=1$	0	1	1		

- 采样 A，马尔可夫毯 $\{B, E, J, M\}$

$$P(A|\neg B, E, \neg J, M) \propto P(A|\neg B, E)P(\neg J|A)P(M|A) = 0.29 \times 0.1 \times 0.7$$

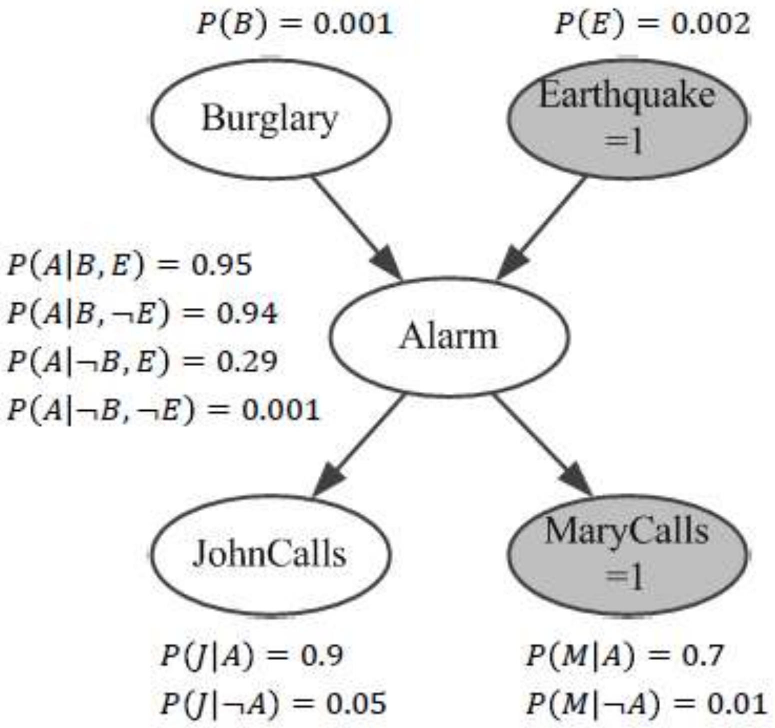
$$P(\neg A|\neg B, E, \neg J, M) \propto P(\neg A|\neg B, E)P(\neg J|\neg A)P(M|\neg A) = 0.71 \times 0.95 \times 0.01$$

- 采得 $A = 1$

实例模型

吉布斯采样

思路：直接依照条件概率 $P(x_Q|x_E)$采样。



Iteration	B	E	A	J	M
$t=0$	0	1	0	0	1
$t=1$	0	1	1		

- 采样 J，马尔可夫毯 $\{A\}$

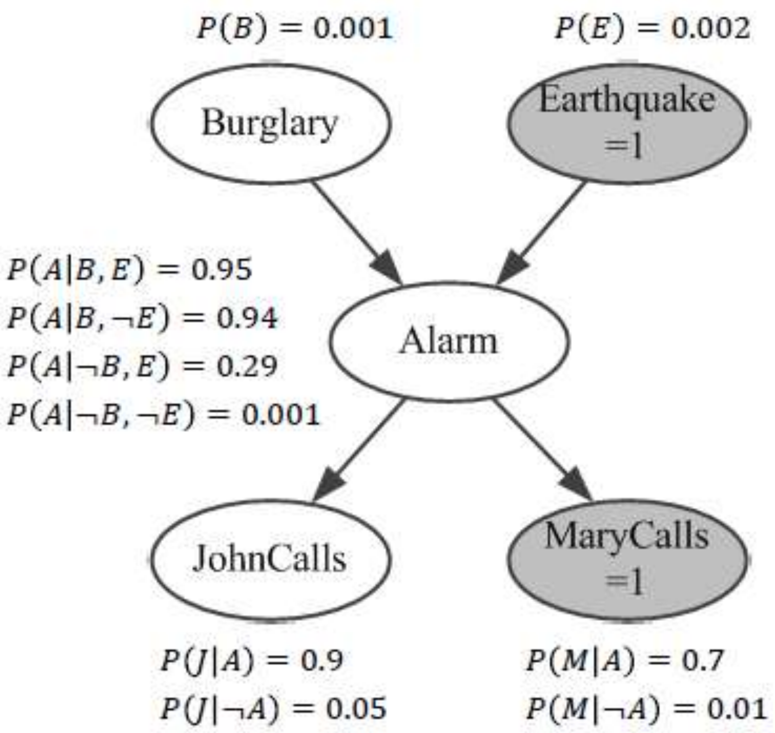
$P(J|A) = 0.9$

$P(\neg J|A) = 0.1$

实例模型

吉布斯采样

思路：直接依照条件概率 $P(x_Q|x_E)$采样。



Iteration	B	E	A	J	M
$t=0$	0	1	0	0	1
$t=1$	0	1	1	1	

- 采样 J，马尔可夫毯 $\{A\}$

  $P(J|A) = 0.9$

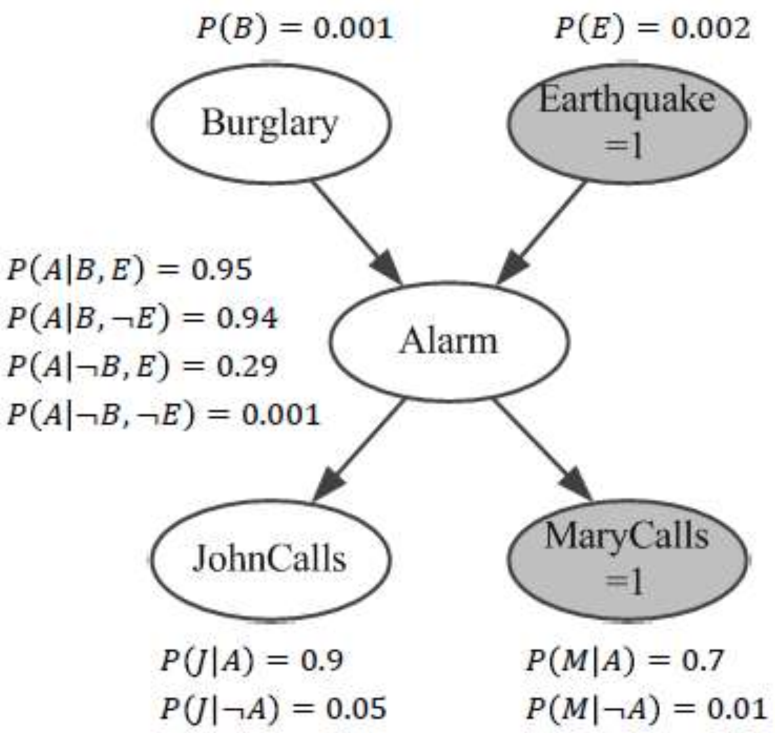
  $P(\neg J|A) = 0.1$

- 采得 $J = 1$

实例模型

吉布斯采样

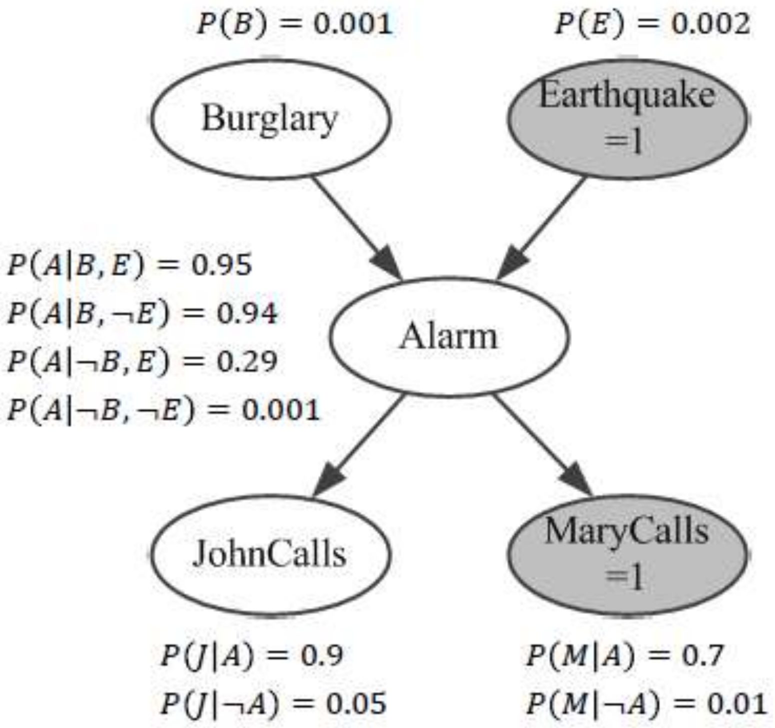
思路：直接依照条件概率 $P(x_Q|x_E)$采样。



Iteration	B	E	A	J	M
$t=0$	0	1	0	0	1
$t=1$	0	1	1	1	1

吉布斯采样

思路：直接依照条件概率 $P(x_Q|x_E)$采样。



Iteration	B	E	A	J	M
$t=0$	0	1	0	0	1
$t=1$	0	1	1	1	1
$t=2$	0	1	0	1	1
$t=3$	1	1	1	1	1

实例模型

吉布斯采样

思路：直接依照条件概率 $P(x_Q|x_E)$采样。

优点：I ）直接从 $P(x_Q|x_E)$采样，解决小概率事件采样难的问题；（Fixed X_E ）

II ）同时适用于贝叶斯网络和马尔可夫随机场；

III ）简单易推导，时间复杂度低。

第九章 概率图模型

实例模型

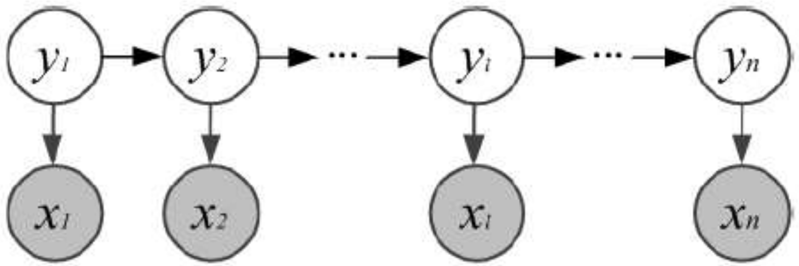
隐马尔可夫模型

条件随机场

隐马尔可夫模型

HMM 模型定义

隐马尔可夫模型是关于时序的概率模型，是最简单的动态贝叶斯网络模型。

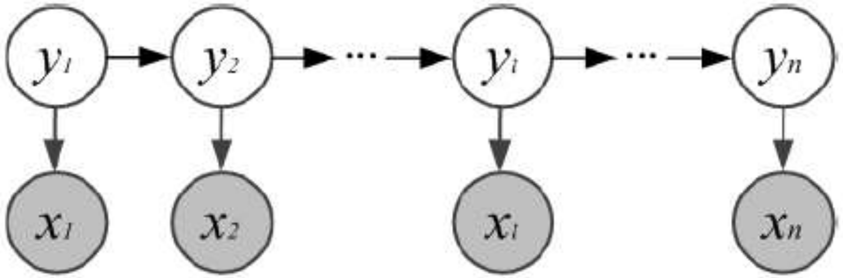


状态变量 $\{y_1, y_2, \ldots, y_n\}$ ， $y_t \in Y$ 表示第 t 时刻的系统状态，

隐马尔可夫模型

HMM 模型定义

隐马尔可夫模型是关于时序的概率模型，是最简单的动态贝叶斯网络模型。



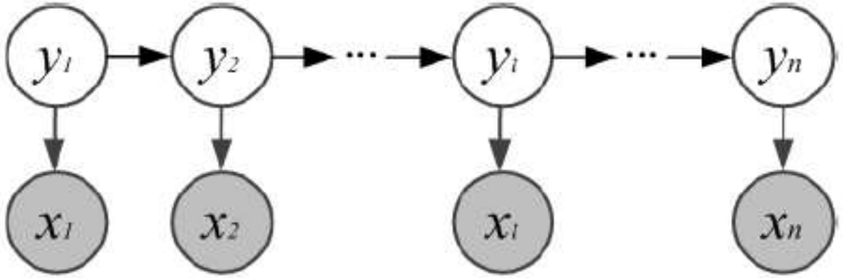
状态变量 $\{y_1, y_2, ..., y_n\}$ ，$y_t \in Y$ 表示第 t 时刻的系统状态，

观测变量 $\{x_1, x_2, ..., x_n\}$ ，$x_t \in X$ 表示第 t 时刻的观测值。

隐马尔可夫模型

HMM 模型定义

隐马尔可夫模型是关于时序的概率模型，是最简单的动态贝叶斯网络模型。



状态变量 $\{y_1, y_2, \ldots, y_n\}$ ， $y_t \in Y$ 表示第 t 时刻的系统状态，

观测变量 $\{x_1, x_2, \ldots, x_n\}$ ， $x_t \in X$ 表示第 t 时刻的观测值。

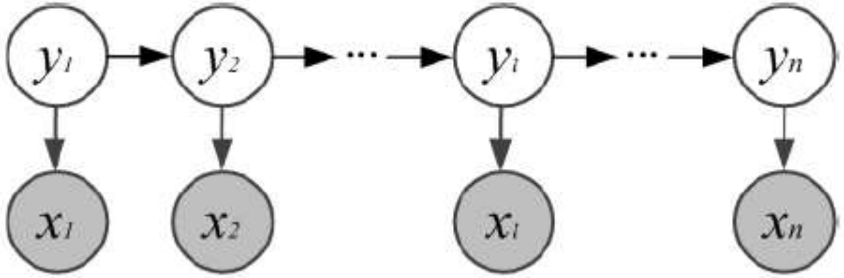
观测变量仅依赖于当前时刻的状态变量，当前状态仅依赖于前一时刻的状态。

状态集合 $Y = \{s_1, s_2, \ldots, s_N\}$ ，观测值集合 $X = \{o_1, o_2, \ldots, o_M\}$ 。

隐马尔可夫模型

HMM 模型定义

联合概率



$$P(x_1, y_1, \ldots, x_n, y_n) = P(y_1)P(x_1 \mid y_1)\prod_{t=2}^{n} P(x_t \mid y_t)P(y_t \mid y_{t-1})$$

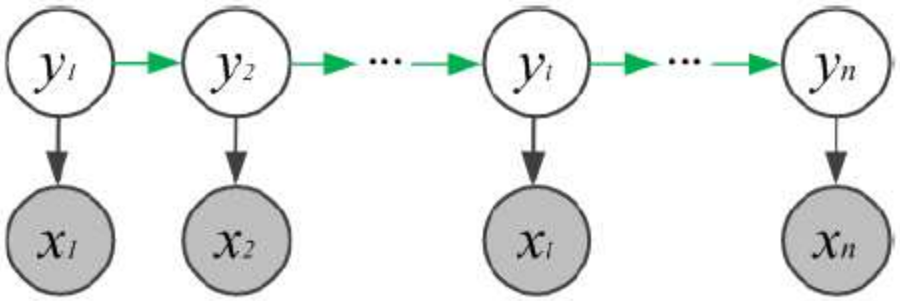
隐马尔可夫模型

HMM 模型参数

状态转移矩阵 A=[a_{ij}]$_{N\times N}$

$$a_{ij} = P(y_{t+1} = s_j \mid y_t = s_i) \qquad 1 \le i, j \le N$$

表示 t 时刻处于状态 s_i 的条件下，t+1 时刻转移到状态 s_j 的概率。



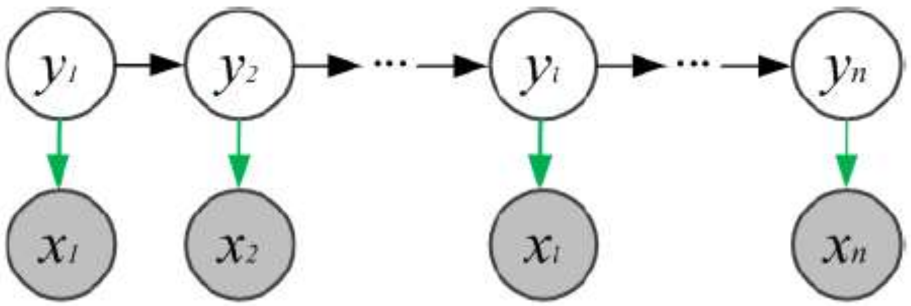
隐马尔可夫模型

HMM 模型参数

观测概率矩阵 $B=[b_{ij}]_{N\times M}$

$$b_{ij}=P(x_t=o_j|y_t=s_i) \qquad 1\le i\le N,\ 1\le j\le M$$

表示 t 时刻处于状态 s_i 的条件下观测到 o_j 的概率。



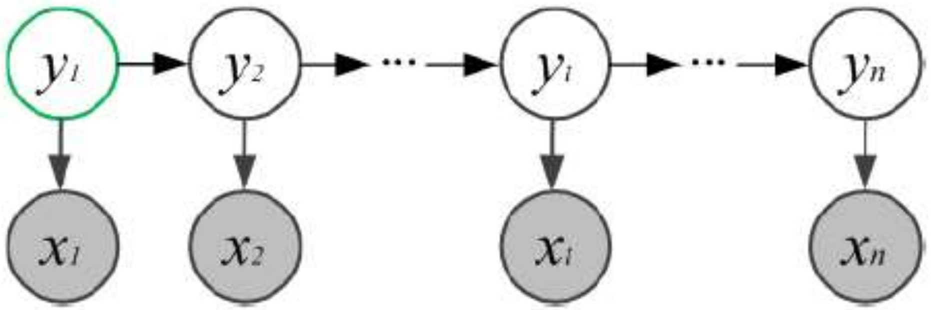
隐马尔可夫模型

HMM 模型参数

初始状态概率向量 $\boldsymbol{\pi}=(\pi_1,\pi_2,\ldots,\pi_N)$

$$\pi_i = P(y_1 = s_i) \qquad 1 \le i \le N$$

表示系统初始状态为 s_i 概率。



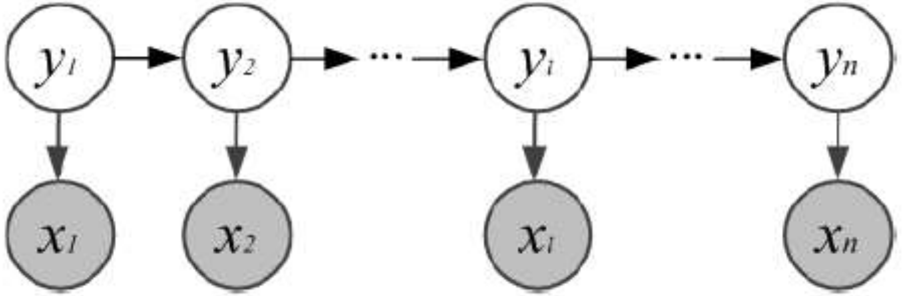
隐马尔可夫模型

HMM 模型参数

初始状态概率向量 $\pi=(\pi_1,\pi_2,\ldots,\pi_N)$

$$\pi_i = P(y_1 = s_i) \qquad 1 \le i \le N$$

表示系统初始状态为 s_i 概率。

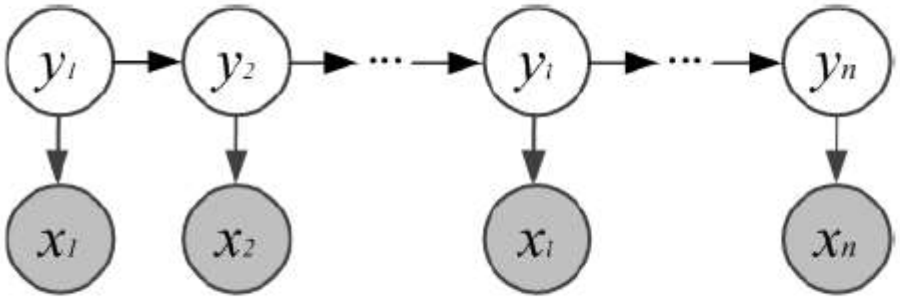


隐马尔可夫模型由 A，B，π 唯一确定，A，B，π 称为隐马尔可夫模型的三要素。

隐马尔可夫模型

生成过程

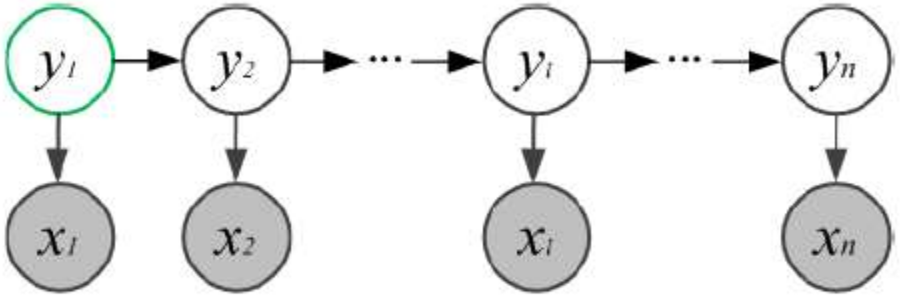
给定 A，B，π，生成观测序列$\{x_1, x_2, \ldots, x_n\}$。



隐马尔可夫模型

生成过程

给定 A，B，π，生成观测序列$\{x_1, x_2, \ldots, x_n\}$。

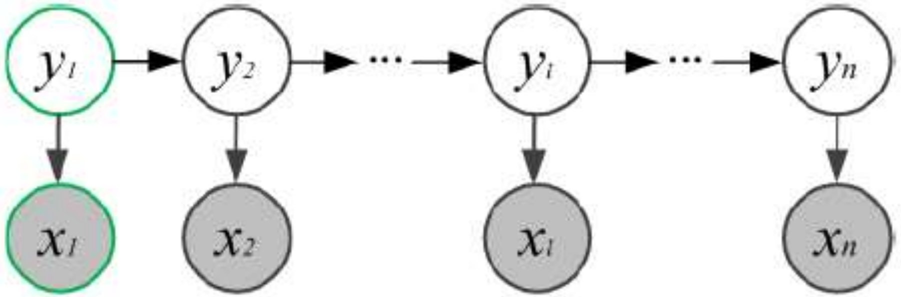


（1）设置 t=1，并根据初始状态概率 π 生成初始状态 y_1。

隐马尔可夫模型

生成过程

给定 A，B，π，生成观测序列$\{x_1, x_2, ..., x_n\}$。



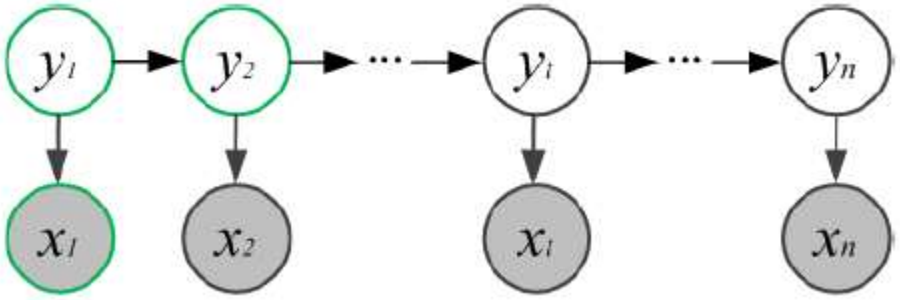
（1）设置 t=1，并根据初始状态概率 π 生成初始状态 y_1。

（2）根据 y_t 和观测概率矩阵 B 生成 x_t。

隐马尔可夫模型

生成过程

给定 A，B，π，生成观测序列$\{x_1, x_2, ..., x_n\}$。



（1）设置 t=1，并根据初始状态概率 π 生成初始状态 y_1。

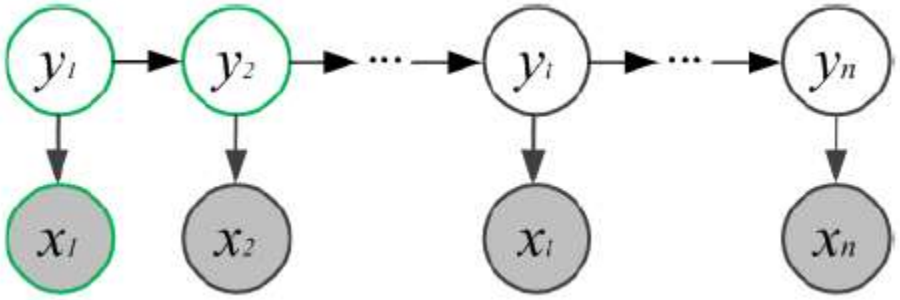
（2）根据 y_t 和观测概率矩阵 B 生成 x_t。

（3）根据 y_t 和状态转移矩阵 A 生成 y_{t+1}。

隐马尔可夫模型

生成过程

给定 A，B，π，生成观测序列$\{x_1, x_2, ..., x_n\}$。



（1）设置 t=1，并根据初始状态概率 π 生成初始状态 y_1。

（2）根据 y_t 和观测概率矩阵 B 生成 x_t。

（3）根据 y_t 和状态转移矩阵 A 生成 y_{t+1}。

（4）若 $t<n$，则设置 $t=t+1$，并转到第（2）步；否则，停止。

隐马尔可夫模型

三个基本问题

概率计算问题：给定模型 $\lambda=(A,B,\pi)$ 和观测序列 $\mathbf{x}=\{x_1,\dots,x_n\}$ 计算在模型 λ 下观测到 x 的概率 $P(\mathbf{x}\mid\lambda)$。（评估模型与观测序列之间的匹配程度）

隐马尔可夫模型

三个基本问题

概率计算问题：给定模型$\lambda=(A,B,\pi)$和观测序列$\mathbf{x}=\{x_1,...,x_n\}$计算在模型$\lambda$下观测到 $\mathbf{x}$ 的概率$P(\mathbf{x}\mid\lambda)$。（评估模型与观测序列之间的匹配程度）

预测问题：给定模型$\lambda=(A,B,\pi)$和观测序列$\mathbf{x}=\{x_1,...,x_n\}$，求使得条件概率$P(\mathbf{y}\mid\mathbf{x},\lambda)$最大的状态观测序列$\mathbf{y}=\{y_1,...,y_n\}$。（根据观测序列推测状态序列）

隐马尔可夫模型

三个基本问题

概率计算问题：给定模型 $\lambda=(A,B,\pi)$ 和观测序列 $\mathbf{x}=\{x_1,\ldots,x_n\}$ 计算在模型 λ 下观测到 $\mathbf{x}$ 的概率 $P(\mathbf{x}\,|\,\lambda)$。（评估模型与观测序列之间的匹配程度）

预测问题：给定模型 $\lambda=(A,B,\pi)$ 和观测序列 $\mathbf{x}=\{x_1,\ldots,x_n\}$ ，求使得条件概率 $P(\mathbf{y}\,|\,\mathbf{x},\lambda)$ 最大的状态观测序列 $\mathbf{y}=\{y_1,\ldots,y_n\}$。（根据观测序列推测状态序列）

学习问题：给定观测序列 $\mathbf{x}=\{x_1,\ldots,x_n\}$ ，调整模型 $\lambda=(A,B,\pi)$ 参数，使得该序列出现的概率 $P(\mathbf{x}\,|\,\lambda)$ 最大。（训练模型使其更好地描述观测序列）

隐马尔可夫模型

概率计算问题

给定模型 $\lambda = (A, B, \pi)$ 和观测序列 $\mathbf{x} = \{x_1, \dots, x_n\}$ 计算在模型 λ 下观测到 $\mathbf{x}$ 的概率 $P(\mathbf{x} \mid \lambda)$。

（评估模型与观测序列之间的匹配程度）

直接计算法

$$\Pr(\mathbf{x}|\lambda) = \sum_{\mathbf{y}} \Pr(\mathbf{x}, \mathbf{y}|\lambda) = \sum_{\mathbf{y}} \Pr(\mathbf{x}|\mathbf{y}, \lambda)\Pr(\mathbf{y}|\lambda)$$

前向后向算法

隐马尔可夫模型

直接计算法

（1）列举所有可能的长度为 n 的状态序列 $\mathbf{y}=\{y_1,\ldots,y_n\}$。

$$\Pr(\mathbf{y}|\lambda)=\pi_{y_1}a_{y_1y_2}a_{y_2y_3}\cdots a_{y_{n-1}y_n}$$

（2）求各状态序列 y 与观测序列 x 的联合概率 $P(\mathbf{x},\mathbf{y}\,|\,\lambda)$。

$$\Pr(\mathbf{x}|\mathbf{y},\lambda)=b_{y_1}(x_1)b_{y_2}(x_2)\cdots b_{y_n}(x_n)$$
$$\Pr(\mathbf{x},\mathbf{y}|\lambda)=\Pr(\mathbf{x}|\mathbf{y},\lambda)\Pr(\mathbf{y}|\lambda)=\pi_{y_1}b_{y_1}(x_1)a_{y_1y_2}b_{y_2}(x_2)\cdots a_{y_{n-1}y_n}b_{y_n}(x_n)$$

（3）对所有的状态序列 $\mathbf{y}=\{y_1,\ldots,y_n\}$ 求和，得到 $P(\mathbf{x}\,|\,\lambda)$。

$$\Pr(\mathbf{x}|\lambda)=\sum_{\mathbf{y}}\Pr(\mathbf{x},\mathbf{y}|\lambda)=\sum_{y_1,y_2,\cdots,y_n}\pi_{y_1}b_{y_1}(x_1)a_{y_1y_2}b_{y_2}(x_2)\cdots a_{y_{n-1}y_n}b_{y_n}(x_n)$$

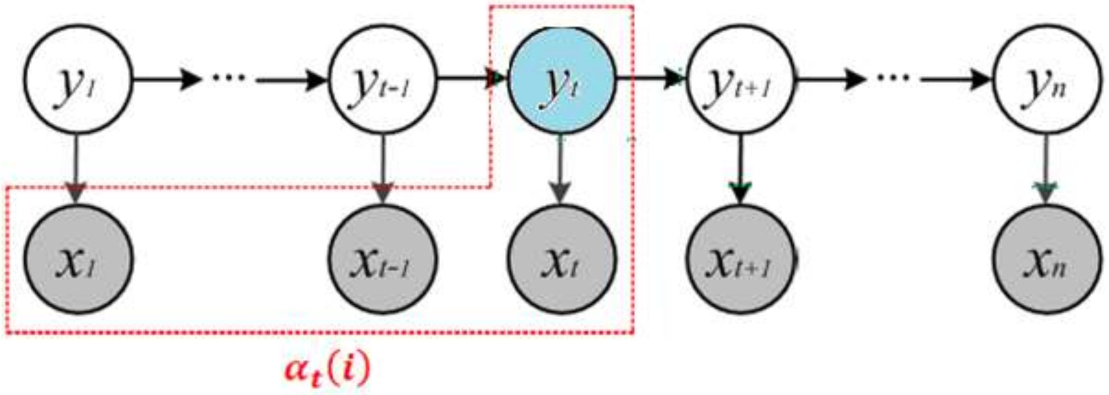
计算复杂度 $O(nN^n)$（每条路径有 2n 个乘法操作，共 N^n 个路径）

隐马尔可夫模型

前向算法

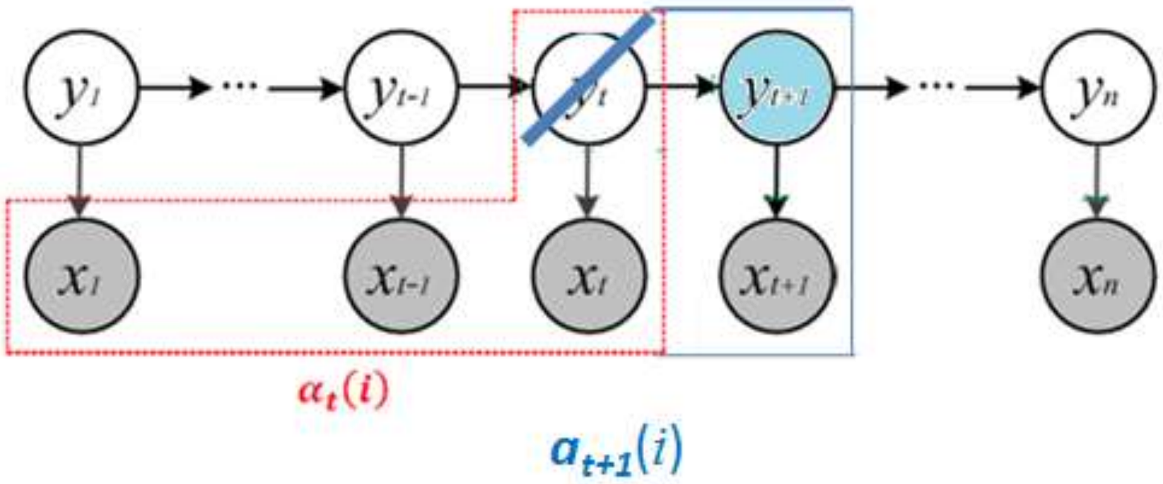
前向概率： 给定隐马尔可夫模型 $\lambda=(A,B,\pi)$ ，定义到时间 t 的部分观测序列为 $x_1, x_2,\ldots, x_t$ 并且状态为 s_i 的概率为前向概率，记作

$$\alpha_t(i)=\Pr(x_1,x_2,\ldots,x_t,y_t=s_i\mid\lambda)$$



前向算法

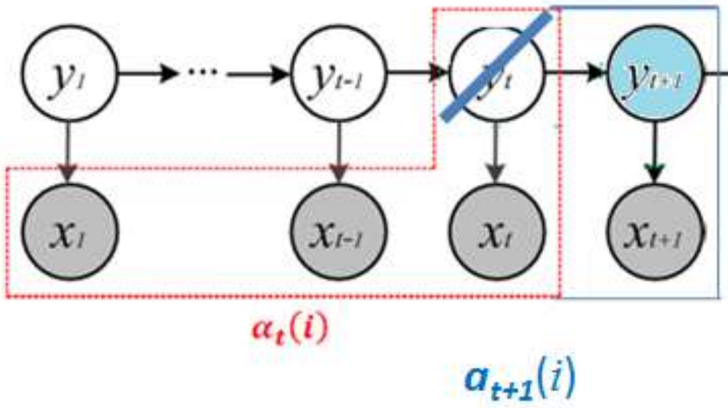
前向概率递推公式：

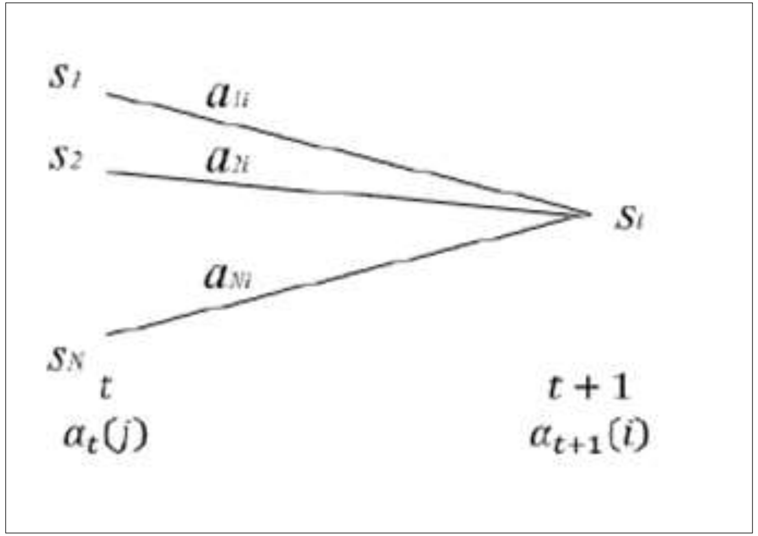


隐马尔可夫模型

前向算法

前向概率递推公式：





$$
\begin{aligned}
\alpha_{t+1}(i) &= b_i(x_{t+1})\sum_j \Pr(x_1, x_2, \cdots, x_t, y_t = s_j|\lambda)\Pr(y_{t+1} = s_i|y_t = s_j, \lambda) \\
&= b_i(x_{t+1})\sum_j \alpha_t(j)a_{ji}, \quad i = 1, 2, \cdots, N.
\end{aligned}
$$

隐马尔可夫模型

前向算法

Algorithm 1 观测序列概率的前向算法

Input: 隐马尔可夫模型 $\lambda = [\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}]$, 观测序列 $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$

1: 初始化前向概率 $\alpha_1(i) = \pi_i b_i(x_1), \quad i = 1, 2, \cdots, N$

2: **for** $t = 1 : (n-1)$ **do**

3: 递推计算前向概率 $\alpha_{t+1}(i) = b_i(x_{t+1}) \sum_{j=1}^{N} \alpha_t(j) a_{ji}, \quad i = 1, 2, \cdots, N$

4: **end for**

Output: 观测序列概率 $\Pr(\mathbf{x}|\lambda) = \sum_{i=1}^{N} \Pr(\mathbf{x}, y_n = s_i|\lambda) = \sum_{i=1}^{N} \alpha_n(i)$

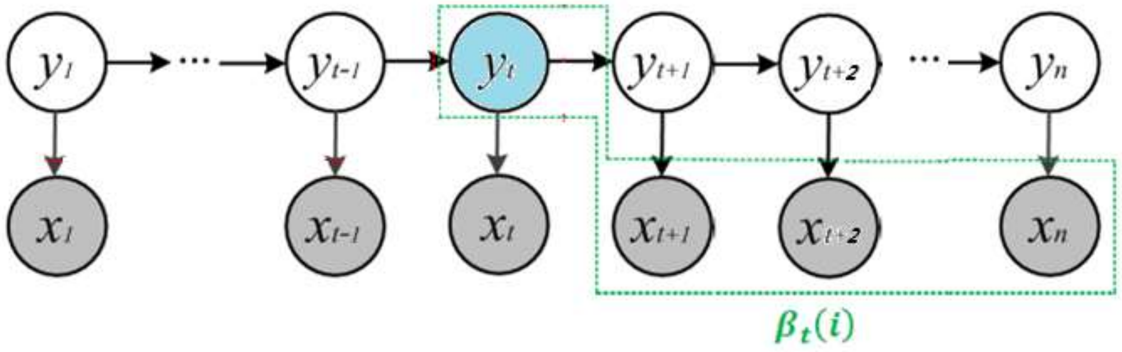
计算复杂度 O(nN²) （每层有 N^2 + N 个乘操作，共 n 层）

隐马尔可夫模型

后向算法

后向概率： 给定隐马尔可夫模型 $\lambda=(A,B,\pi)$ ，定义时间 t 状态为 s_i 的条件下，从 t+1 时刻到 n 时刻的部分观测序列为 $x_{t+1}, x_{t+2},\ldots,x_n$ 的概率为后向概率，记作

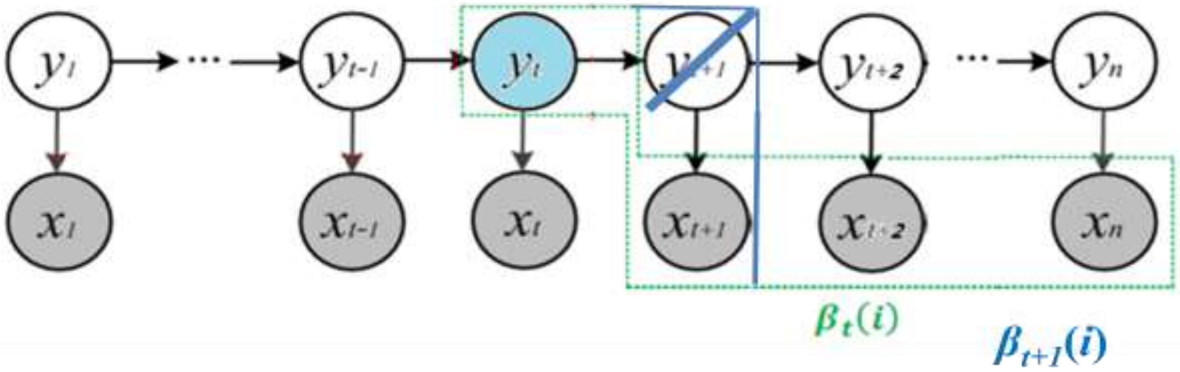
$$\beta_t(i)=\Pr(x_{t+1},x_{t+2},\ldots,x_n \mid y_t=s_i,\lambda)$$



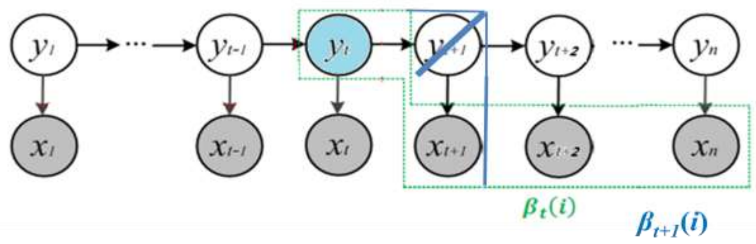
隐马尔可夫模型

后向算法

后向概率递推公式：

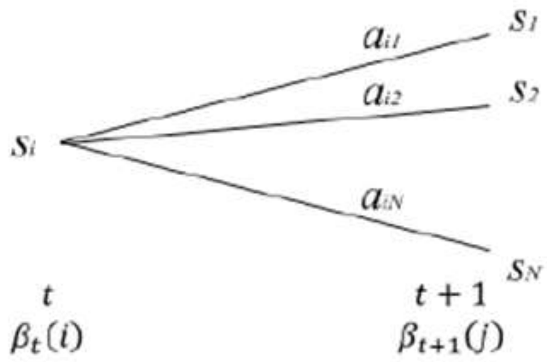


后向算法



$$\begin{aligned}
\beta_t(i) &= \Pr(x_{t+1}, x_{t+2}, \cdots, x_n | y_t = s_i, \lambda) \\
&= \sum_j \Pr(x_{t+1}, x_{t+2}, \cdots, x_n, y_{t+1} = s_j | y_t = s_i, \lambda) \\
&= \sum_j \Pr(y_{t+1} = s_j | y_t = s_i, \lambda) \Pr(x_{t+1}, x_{t+2}, \cdots, x_n | y_{t+1} = s_j, \boxed{y_t = s_i, \lambda)} \\
&= \sum_j \Pr(y_{t+1} = s_j | y_t = s_i, \lambda) \Pr(x_{t+1}, x_{t+2}, \cdots, x_n | y_{t+1} = s_j, \lambda) \\
&= \sum_j a_{ij} \Pr(x_{t+1} | y_{t+1} = s_j, \lambda) \Pr(x_{t+2}, \cdots, x_n | y_{t+1} = s_j, \lambda) \\
&= \sum_j a_{ij} b_j(x_{t+1}) \beta_{t+1}(j), \quad i = 1, 2, \cdots, N.
\end{aligned}$$

y_t 与 $x_{t+1},\ldots x_n$ 无关



后向算法

Algorithm 2 观测序列概率的后向算法

Input: 隐马尔可夫模型 $\lambda = [\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}]$, 观测序列 $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$

1: 初始化后向概率 $\beta_n(i) = 1, \quad i = 1, 2, \cdots, N$

2: **for** $t = (n-1) : 1$ **do**

3: 递推计算后向概率 $\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(x_{t+1}) \beta_{t+1}(j), \quad i = 1, 2, \cdots, N$

4: **end for**

Output: 观测序列概率 $\Pr(\mathbf{x}|\lambda) = \sum_{i=1}^{N} \pi_i b_i(x_1) \beta_1(i)$

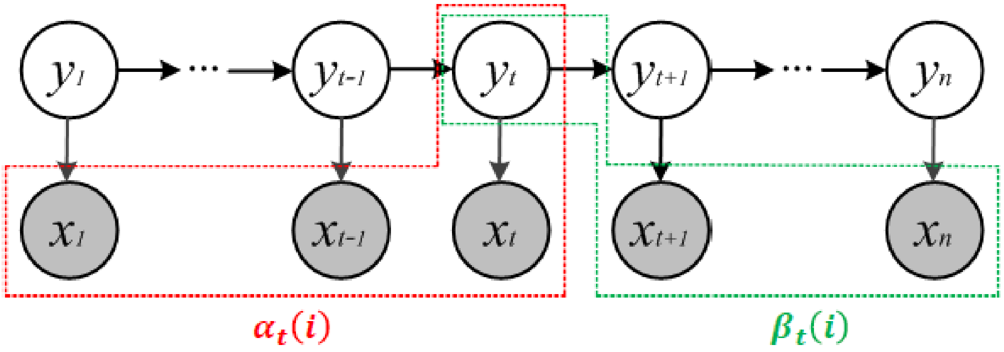
计算复杂度 O(nN²) （每层有 N^2 + N 个乘操作，共 n 层）

隐马尔可夫模型

前向-后向算法

利用前向和后向概率，可以将观测概率统一写成

$$\Pr(\mathbf{x} \mid \lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i), \quad t = 1, 2, \ldots, n$$



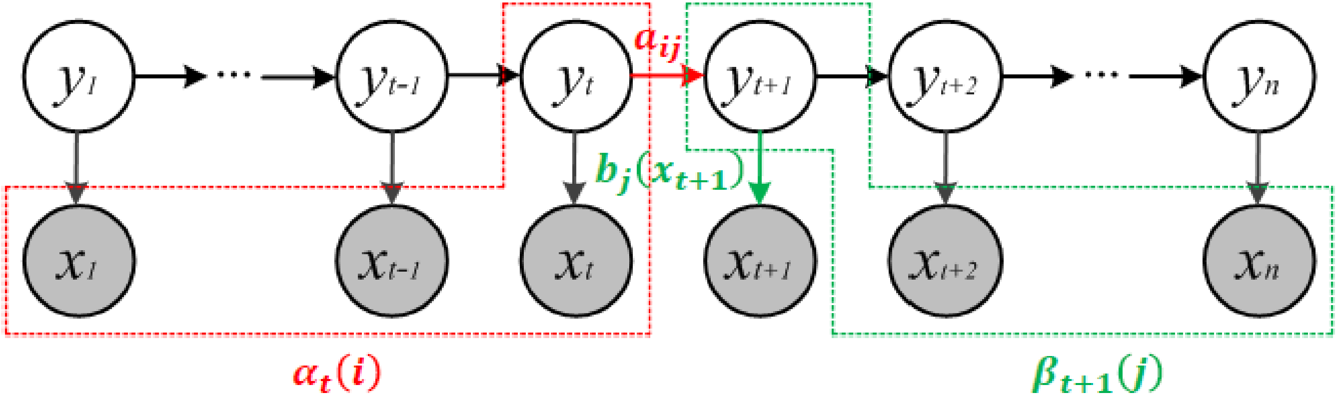
$$\begin{aligned}
\alpha_t(i)\,\beta_t(i) &= \Pr(x_1, x_2, \ldots, x_t, y_t = s_i \mid \lambda)\,\Pr(x_{t+1}, x_{t+2}, \ldots, x_n \mid y_t = s_i, \lambda) \\
&= \Pr(x_1, x_2, \ldots, x_t, y_t = s_i \mid \lambda)\,\Pr(x_{t+1}, x_{t+2}, \ldots, x_n \mid y_t = s_i, \boxed{x_1, x_2, \ldots, x_t}, \lambda) \\
&= \Pr(x_1, x_2, \ldots, x_t, x_{t+1}, x_{t+2}, \ldots, x_n, y_t = s_i \mid \lambda)
\end{aligned}$$

与 $x_{t+1},\ldots,x_n$ 无关

前向-后向算法

利用前向和后向概率，也可以将观测概率统一写成

$$\Pr(\mathbf{x}\,|\,\lambda)=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(x_{t+1})\beta_{t+1}(i),\quad t=1,2,\ldots,n-1$$

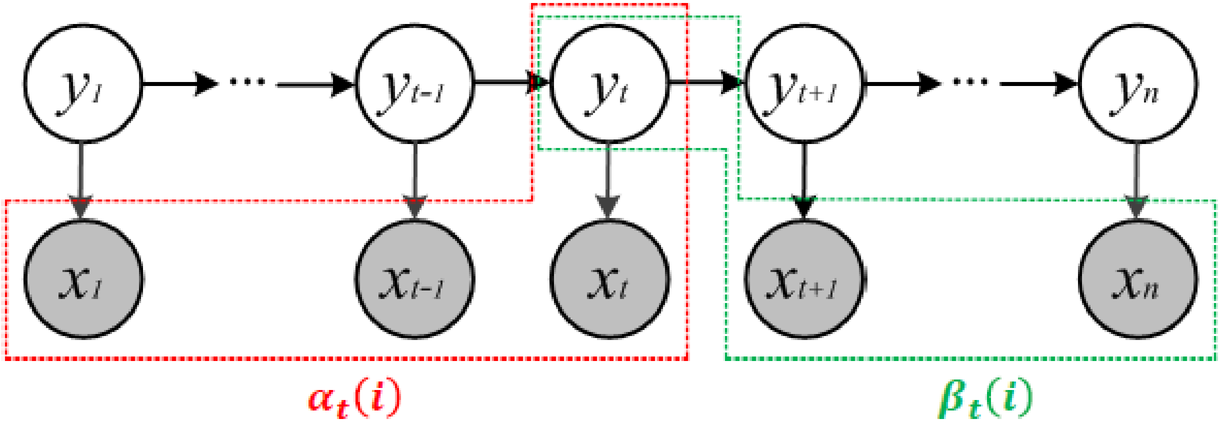


隐马尔可夫模型

相关概率和期望的计算

给定隐马尔可夫模型 $\lambda=(A,B,\pi)$ 和观测序列 **x**，在时刻 t 处于状态 s_i 的概率：

$$\gamma_t(i)=\Pr(y_t=s_i|\mathrm{x},\lambda)=\frac{\Pr(y_t=s_i,\mathrm{x}|\lambda)}{\Pr(\mathrm{x}|\lambda)}=\frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)}$$



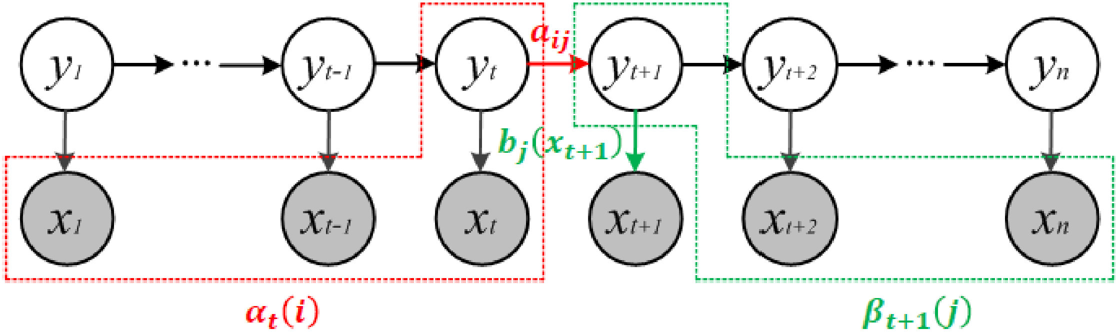
隐马尔可夫模型

相关概率和期望的计算

给定隐马尔可夫模型$\lambda=(A,B,\pi)$和观测序列 **x**，在时刻 t 处于状态 s_i 并且在时刻 t+1 处于状态 s_j 的概率：

$$\xi_t(i,j)=\Pr(y_t=s_i,y_{t+1}=s_j|\mathbf{x},\lambda)=\frac{\Pr(y_t=s_i,y_{t+1}=s_j,\mathbf{x}|\lambda)}{\Pr(\mathbf{x}|\lambda)}$$

$$=\frac{\alpha_t(i)a_{ij}b_j(x_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^N\sum_{j=1}^N\alpha_t(i)a_{ij}b_j(x_{t+1})\beta_{t+1}(j)}.$$



隐马尔可夫模型

相关概率和期望的计算

在观测 **x** 下状态 s_i 出现的期望值（所有时刻 i 出现上的期望和）：

$$\sum_{t=1}^{n} \gamma_t(i)$$

在观测 **x** 下由状态 s_i 转移的期望值（所有时刻 i 移出的期望和）：

$$\sum_{t=1}^{n-1} \gamma_t(i)$$

在观测 **x** 下由状态 s_i 转移到状态 s_j 的期望值（所有时刻 i 转移 j 的期望和）：

$$\sum_{t=1}^{n-1} \xi_t(i,j)$$

隐马尔可夫模型

预测问题

给定模型$\lambda=(A,B,\pi)$和观测序列$\mathbf{x}=\{x_1,\ldots,x_n\}$，求使得条件概率$P(\mathbf{y}\mid\mathbf{x},\lambda)$最大的状态观测序列$\mathbf{y}=\{y_1,\ldots,y_n\}$，即根据观测序列推测状态序列。

贪婪算法（精确算法）：找最大值路径

比较每条 y 路径的 $\Pr(\mathbf{x},\mathbf{y}|\lambda)$

近似算法

维特比算法（精确算法）

近似算法

思路： 在每个时刻 t 选择最有可能出现的状态 y_t^* ，得到一个状态序列 $\mathbf{y}^* = \{y_1^*, y_2^*, ..., y_n^*\}$ ，将这作为预测结果。

$$
\begin{aligned}
y_t^* &= \arg\max_{1 \le i \le N} \Pr(y_t = s_i | \mathrm{x}, \lambda) \\
&= \arg\max_{1 \le i \le N} \gamma_t(i) = \arg\max_{1 \le i \le N} \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}, \quad t = 1, 2, \cdots, n
\end{aligned}
$$

优点：计算简单，但不能保证预测结果是合法的（ $a_{ij}=0$ ）。

隐马尔可夫模型

维特比算法

思路：利用动态规划求解概率最大路径，这里一条路径对应着一个状态序列。

如果最优路径$\mathbf{y}^*$在 t 时刻通过结点y_t^*，那么这条路径从起始结点到结点y_t^*的路径中,局部路径$\mathbf{y}_{1:t}^*$ $(y_1^* \sim y_t^*)$一定是最优的。(每个结点y_t^*对应一个最优路径)

假定从起始时刻到 t 时刻上各个状态的最优路径已经找到，那么在计算从起始时刻到 t+1 时刻上的某个状态 s_j 的最优路径时，只需要考虑从起始时刻到上一时刻所有 N 个状态 s_i 的最优路径，以及从 s_i 到 s_j 的"距离"。

隐马尔可夫模型

维特比算法

维特比变量：在时刻 t，隐马尔可夫模型沿着一条路径到达状态 s_i，并输出观测序列 $\mathbf{x}=\{x_1,\ldots,x_n\}$ 的最大概率。

$$\delta_t(i)=\max_{y_1,y_2,\cdots,y_{t-1}}\Pr(y_t=s_i,y_{t-1},\cdots,y_1,x_t,x_{t-1},\cdots,x_1|\lambda)$$

$$\delta_1(i)=\pi_i b_i(x_1)\qquad \delta_t(i)=\max_{1\le j\le N}\delta_{t-1}(j)a_{ji}b_i(x_t),\ t=2,3,\cdots,n$$

路径变量：记录该路径上状态 s_i 的前一个状态。

$$\phi_t(i)=\arg\max_{1\le j\le N}\delta_{t-1}(j)a_{ji}b_i(x_t),\ t=2,3,\cdots,n$$

维特比算法

Algorithm 3 维特比算法

Input: 隐马尔可夫模型 $\lambda = [\mathbf{A}, \mathbf{B}, \pi]$, 观测序列 $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$

1: 初始化: $\delta_1(i) = \pi_i b_i(x_1)$, $\phi_1(i) = 0$, $\quad i = 1, 2, \cdots, N$

2: **for** $t = 2 : n$ **do**

3: 递推计算: $\delta_t(i) = \max\limits_{1 \le j \le N} \delta_{t-1}(j) a_{ji} b_i(x_t)$, $\quad i = 1, 2, \cdots, N$

4: 记忆路径: $\phi_t(i) = \arg \max\limits_{1 \le j \le N} \delta_{t-1}(j) a_{ji} b_i(x_t)$, $\quad i = 1, 2, \cdots, N$

5: **end for**

6: 终结: $P^* = \max\limits_{1 \le i \le N} \delta_n(i)$, $y_n^* = \arg \max\limits_{1 \le i \le N} \delta_n(i)$

7: **for** $t = (n-1) : 1$ **do**

8: 回溯最优路径: $y_t^* = \phi_{t+1}(y_{t+1}^*)$

9: **end for**

Output: 最优状态序列 $\mathbf{y}^* = \{y_1^*, y_2^*, \cdots, y_n^*\}$

计算复杂度 O(nN^2)（每层都是 N^2+N 个乘法，共 n 层）

隐马尔可夫模型

学习问题

给定观测序列 $\mathbf{x}=\{x_1,\ldots,x_n\}$ ，调整模型 $\lambda=(A,B,\pi)$ 参数，使得该序列出现的概率 $P(\mathbf{x}\mid\lambda)$ 最大，即训练模型使其更好地描述观测序列。

监督学习方法

非监督学习方法（Baum-Welch 算法）

隐马尔可夫模型

监督学习方法

训练数据包含S个长度相同的观测序列和对应的状态序列$\{(x_1, y_1), (x_2, y_2), \ldots, (x_s, y_s)\}$，利用极大似然法估计隐马尔可夫模型的参数。 （P（X,Y）最大似然估计）

状态转移概率： $\hat{a}_{ij} = \dfrac{A_{ij}}{\sum_{j=1}^{N} A_{ij}}$ ，A_{ij}是由状态s_i转移到s_j的频数。

输出观测概率： $\hat{b}_{ik} = \dfrac{B_{ik}}{\sum_{k=1}^{M} B_{ik}}$ ，B_{ik}是由状态s_i观测到o_k的频数。

初始状态概率： $\hat{\pi}_i$ 为所有S个样本中初始状态为s_i的频率。

隐马尔可夫模型

非监督学习方法

训练数据只包含S个长度相同的观测序列$\{x_1, x_2, \ldots, x_S\}$而没有对应的状态序列。

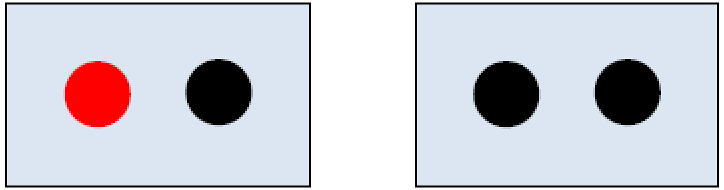
将观测序列看作观测数据$\mathcal{X}$，状态序列看作不可观测的隐数据$\mathcal{Y}$，隐马尔可夫模型等价于含有隐变量的概率模型。

$$P(\mathcal{X}|\lambda) = \sum_{\mathcal{Y}} P(\mathcal{X}|\mathcal{Y}, \lambda) P(\mathcal{Y}|\lambda)$$

相应的参数学习可以由EM算法实现。（《统计学习方法》李航，P 181页)

隐马尔可夫模型

示例：信封问题



（1）　开始等概率选择一个信封，并从中随机抽取出一个球，记录其颜色然后放回。

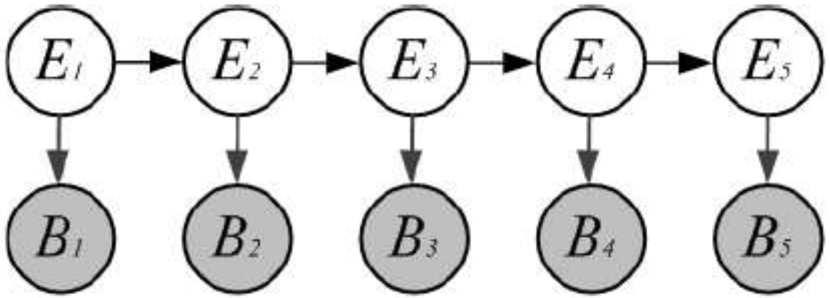
（2）　再次选择一个信封，其规则是如果上次选择的是第一个信封，则本次选第二个信封；否则等概率随机选择。

（3）　确定信封后，从中随机抽取出一个球，记录其颜色然后放回。

（4）　重复上述过程 5 次得到观测序列{红，黑，黑，黑，红}。

隐马尔可夫模型

示例：信封问题



状态变量：信封序列{E_1，E_2，E_3，E_4，E_5}（不可观测）

观测变量：球的颜色序列{B_1=红，B_2=黑，B_3=黑，B_4=黑，B_5=红}

状态集合{0, 1}，观测集合{红，黑}

$$\boldsymbol{\pi} = (0.5, 0.5) \qquad \mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0.5 & 0.5 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 0.5 & 0.5 \\ 0 & 1 \end{bmatrix}$$

实例模型

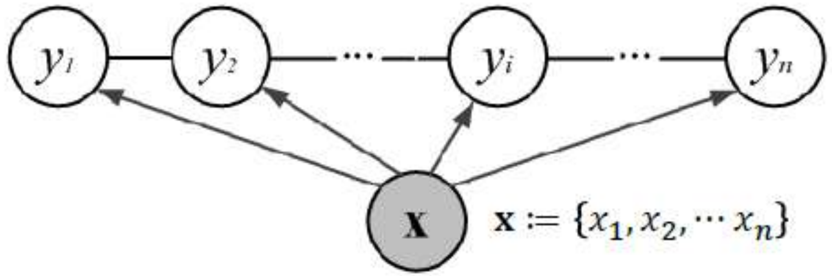
隐马尔可夫模型

条件随机场

条件随机场

CRF 模型定义

条件随机场(Conditional Random Field) 是给定随机变量x的条件下，随机变量y的马尔可夫随机场。$G(V,E)$是y中的随机变量构成的无向图，图中每个变量在给定x的条件下都满足马尔可夫性：



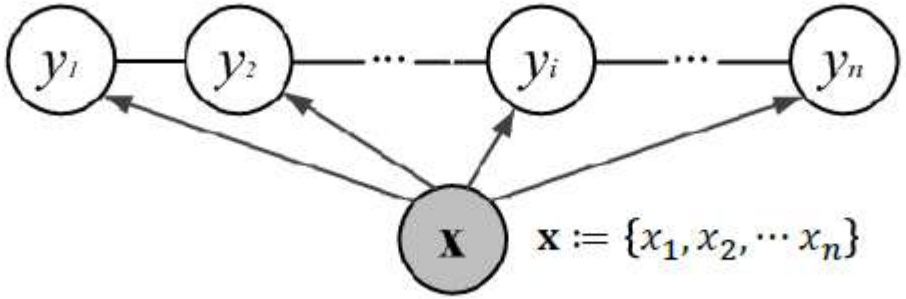
$$P\left(y_v \middle| \mathbf{x}, \mathbf{y}_{V\setminus\{v\}}\right) = P\left(y_v \middle| \mathbf{x}, \mathbf{y}_{MB(v)}\right)$$

$\mathbf{y}_{\mathrm{MB(v)}}$是$y_{\mathrm{v}}$的邻接变量.

条件随机场

线性链条件随机场

线性链条件随机场(linear-chain CRF) 是随机变量y为线性链时的条件随机场。



$\mathbf{x}=\{x_1, x_2,\cdots, x_n\}$ 是观测序列。$\mathbf{y}=\{y_1, y_2,\cdots, y_n\}$是标记序列（或称状态序列）。

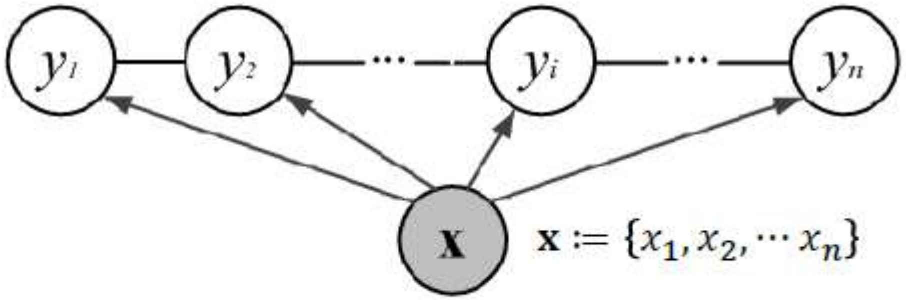
在给定x的条件下，y的条件分布$P(\mathbf{y}|\mathbf{x})$构成条件随机场。

$$P(y_i|\mathbf{x}, y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n) = P(y_i|\mathbf{x}, y_{i-1}, y_{i+1})$$

条件随机场

线性链条件随机场

线性链条件随机场(linear-chain CRF) 是随机变量y为线性链时的条件随机场。



$$P(\mathbf{y}\,|\,\mathbf{x}) = \frac{1}{Z}\exp\left(\sum_{i,j}\lambda_j t_j(y_{i+1}, y_i, \mathbf{x}, i) + \sum_{i,k}\mu_k s_k(y_i, \mathbf{x}, i)\right)$$

转移特征函数：$t_j(y_i, y_{i+1}, \mathbf{x}, i)$　团$\{y_i, y_{i+1}\}$上的势函数：$\Sigma_j \lambda_j t_j(y_i, y_{i+1}, \mathbf{x}, i)$

状态特征函数：$s_k(y_i, \mathbf{x}, i)$　团$\{y_i\}$上的势函数：$\Sigma_k \mu_k s_k(y_i, \mathbf{x}, i)$

条件随机场

三个基本问题

概率计算问题：给定条件随机场、观测序列$\mathbf{x}=x_1,\cdots,x_n$和状态序列$\mathbf{y}=y_1,\cdots,y_n$，计算条件概率$P(y_i|\mathbf{x})$，$P(y_i,y_{i+1}|\mathbf{x})$以及相应的数学期望。

预测问题：给定条件随机场和观测序列$\mathbf{x}=x_1,\cdots,x_n$，求使得条件概率$P(\mathbf{y}|\mathbf{x})$最大的状态序列$\mathbf{y}=y_1,\cdots,y_n$，即根据观测序列推断出隐藏的模型状态（词性标注、语音识别）。

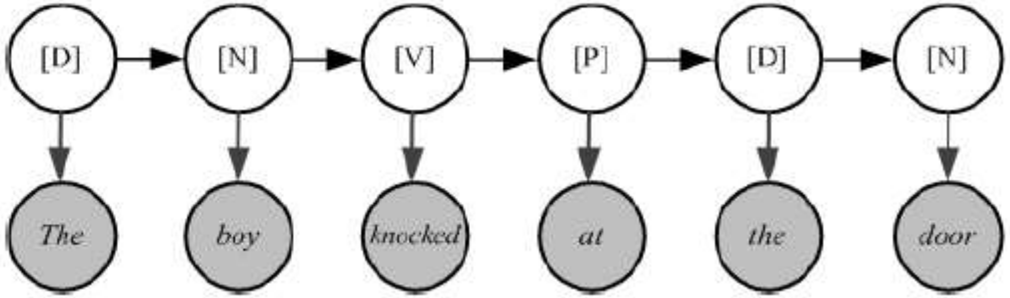
学习问题：给定训练数据$\{(\mathbf{x}_1,\mathbf{y}_1),(\mathbf{x}_2,\mathbf{y}_2),\cdots,(\mathbf{x}_S,\mathbf{y}_S)\}$，估计条件随机场的模型参数使得条件概率$P(\mathcal{Y}|\mathcal{X})$最大（训练模型使其能最好地描述训练数据）

自学《统计学习方法》李航，P207

条件随机场

示例：词性标注

利用隐马尔可夫模型进行词性标注：求解argmax $P(\mathbf{y}\,|\mathbf{x})$



每个词的词性仅和自身有关，和文本中其他所有词均无关。建模联合概率$P(\mathbf{x},\mathbf{y})$。

事实上，观测到“at”对判断“knocked”的词性有帮助。

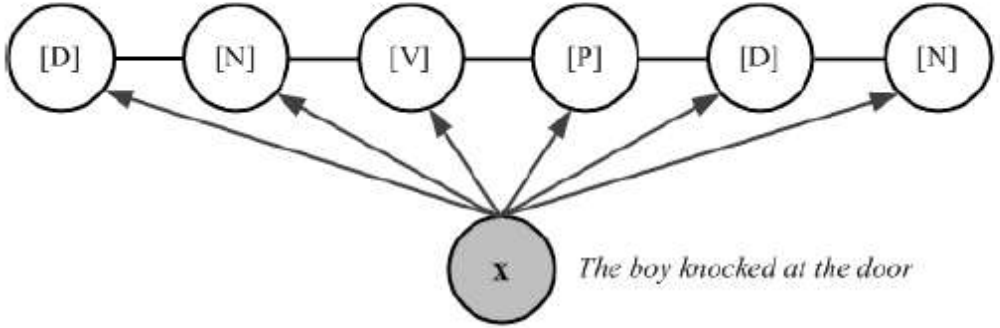
条件随机场

示例：词性标注

利用线性链条件随机场进行词性标注：

[D] [N] [V] [P] [D] [N]

The boy knocked at the door.



直接建模条件概率$P(\mathbf{y}|\mathbf{x})$ ，通过定义适当的特征函数，当前词的词性可以和其他词相关。当下一个观测词是“at”时，当前词对应的词性标记很可能是[V] 。

$$s_k(y_i, \mathbf{x}, i) = \begin{cases} 1, & \text{if } y_i = [V] \text{ and } x_{i+1} = \text{“at”} \\ 0, & \text{otherwise} \end{cases}$$

小结

1. **概率图模型**

   两类：贝叶斯网络（有向图）、马尔可夫随机场（无向图）。

   掌握：概率分布分解形式、条件独立性。

2. **学习与推断**

   学习：结构学习、参数学习（最大似然估计、EM 算法）。

   推断：精确推断（了解变量消去、信念传播）。

   近似推断（了解前向采样、吉布斯采样）。

小结

3. **隐马尔可夫模型**

   三要素：A、B、π。

   概率计算问题：掌握前向、后向算法。

   预测问题：掌握维特比算法。

   学习问题：了解最大似然估计和 EM 算法。

本讲参考文献

1. 《统计机器学习--第九章：概率图模型》课件，王泉，国科大网络安全学院，2017。

致谢王泉副研究员！感谢王泉提供了《概率图模型》课件供本章教学参考！