第10章　隐马尔可夫模型

隐马尔可夫模型（hidden Markov model, HMM）是可用于标注问题的统计学习模型，描述由隐藏的马尔可夫链随机生成观测序列的过程，属于生成模型．本章首先介绍隐马尔可夫模型的基本概念，然后分别叙述隐马尔可夫模型的概率计算算法、学习算法以及预测算法．隐马尔可夫模型在语音识别、自然语言处理、生物信息、模式识别等领域有着广泛的应用．

10.1　隐马尔可夫模型的基本概念

10.1.1　隐马尔可夫模型的定义

定义 10.1（隐马尔可夫模型）　隐马尔可夫模型是关于时序的概率模型，描述由一个隐藏的马尔可夫链随机生成不可观测的状态随机序列，再由各个状态生成一个观测而产生观测随机序列的过程．隐藏的马尔可夫链随机生成的状态的序列，称为状态序列（state sequence）；每个状态生成一个观测，而由此产生的观测的随机序列，称为观测序列（observation sequence）．序列的每一个位置又可以看作是一个时刻．

隐马尔可夫模型由初始概率分布、状态转移概率分布以及观测概率分布确定．隐马尔可夫模型的形式定义如下：

设 Q 是所有可能的状态的集合，V 是所有可能的观测的集合．

$$Q=\{q_1,q_2,\cdots,q_N\}\,,\quad V=\{v_1,v_2,\cdots,v_M\}$$

其中，N 是可能的状态数，M 是可能的观测数．

I 是长度为 T 的状态序列，O 是对应的观测序列．

$$I=(i_1,i_2,\cdots,i_T)\,,\quad O=(o_1,o_2,\cdots,o_T)$$

A 是状态转移概率矩阵：

$$A=\left[a_{ij}\right]_{N\times N} \tag{10.1}$$

其中，

$$a_{ij}=P(i_{t+1}=q_j\mid i_t=q_i)\,,\qquad i=1,2,\cdots,N;\ j=1,2,\cdots,N \tag{10.2}$$

是在时刻 t 处于状态 q_i 的条件下在时刻 $t+1$ 转移到状态 q_j 的概率．

B 是观测概率矩阵：

$$B=\left[b_j(k)\right]_{N\times M} \tag{10.3}$$

其中，

$$b_j(k)=P(o_t=v_k \mid i_t=q_j)\,,\quad k=1,2,\cdots,M;\ j=1,2,\cdots,N \tag{10.4}$$

是在时刻 t 处于状态 q_j 的条件下生成观测 v_k 的概率.

π 是初始状态概率向量：

$$\pi=(\pi_i) \tag{10.5}$$

其中，

$$\pi_i=P(i_1=q_i)\,,\quad i=1,2,\cdots,N \tag{10.6}$$

是时刻 $t=1$ 处于状态 q_i 的概率.

隐马尔可夫模型由初始状态概率向量 π、状态转移概率矩阵 A 和观测概率矩阵 B 决定. π 和 A 决定状态序列，B 决定观测序列. 因此，隐马尔可夫模型 λ 可以用三元符号表示，即

$$\lambda=(A,B,\pi) \tag{10.7}$$

A,B,π 称为隐马尔可夫模型的三要素.

状态转移概率矩阵 A 与初始状态概率向量 π 确定了隐藏的马尔可夫链，生成不可观测的状态序列. 观测概率矩阵 B 确定了如何从状态生成观测，与状态序列综合确定了如何产生观测序列.

从定义可知，隐马尔可夫模型作了两个基本假设：

（1）齐次马尔可夫性假设，即假设隐藏的马尔可夫链在任意时刻 t 的状态只依赖于其前一时刻的状态，与其他时刻的状态及观测无关，也与时刻 t 无关.

$$P(i_t \mid i_{t-1},o_{t-1},\cdots,i_1,o_1)=P(i_t \mid i_{t-1})\,,\quad t=1,2,\cdots,T \tag{10.8}$$

（2）观测独立性假设，即假设任意时刻的观测只依赖于该时刻的马尔可夫链的状态，与其他观测及状态无关.

$$P(o_t \mid i_T,o_T,i_{T-1},o_{T-1},\cdots,i_{t+1},o_{t+1},i_t,i_{t-1},o_{t-1},\cdots,i_1,o_1)=P(o_t \mid i_t) \tag{10.9}$$

隐马尔可夫模型可以用于标注，这时状态对应着标记. 标注问题是给定观测的序列预测其对应的标记序列. 可以假设标注问题的数据是由隐马尔可夫模型生成的. 这样我们可以利用隐马尔可夫模型的学习与预测算法进行标注.

下面看一个隐马尔可夫模型的例子.

例 10.1（盒子和球模型） 假设有 4 个盒子，每个盒子里都装有红白两种颜色的球，盒子里的红白球数由表 10.1 列出：

表 10.1 各盒子的红白球数

盒 子	1	2	3	4
红球数	5	3	6	8
白球数	5	7	4	2

按照下面的方法抽球，产生一个球的颜色的观测序列：开始，从 4 个盒子里以等概率随机选取 1 个盒子，从这个盒子里随机抽出 1 个球，记录其颜色后，放回；然后，从当前盒子随机转移到下一个盒子，规则是：如果当前盒子是盒子 1，那么下一盒子一定是盒子 2，如果当前是盒子 2 或 3，那么分别以概率 0.4 和 0.6 转移到左边或右边的盒子，如果当前是盒子 4，那么各以 0.5 的概率停留在盒子 4 或转移到盒子 3；确定转移的盒子后，再从这个盒子里随机抽出 1 个球，记录其颜色，放回；如此下去，重复进行 5 次，得到一个球的颜色的观测序列：

$$O=\{红, 红, 白, 白, 红\}$$

在这个过程中，观察者只能观测到球的颜色的序列，观测不到球是从哪个盒子取出的，即观测不到盒子的序列.

在这个例子中有两个随机序列，一个是盒子的序列（状态序列），一个是球的颜色的观测序列（观测序列）. 前者是隐藏的，只有后者是可观测的. 这是一个隐马尔可夫模型的例子，根据所给条件，可以明确状态集合、观测集合、序列长度以及模型的三要素.

盒子对应状态，状态的集合是

$$Q=\{盒子1, 盒子2, 盒子3, 盒子4\}, \quad N=4$$

球的颜色对应观测. 观测的集合是

$$V=\{红, 白\}, \quad M=2$$

状态序列和观测序列长度 $T=5$.

初始概率分布为

$$\pi=(0.25, 0.25, 0.25, 0.25)^{\mathrm{T}}$$

状态转移概率分布为

$$A=\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 \\ 0 & 0.4 & 0 & 0.6 \\ 0 & 0 & 0.5 & 0.5 \end{bmatrix}$$

观测概率分布为

$$B=\begin{bmatrix}0.5 & 0.5\\0.3 & 0.7\\0.6 & 0.4\\0.8 & 0.2\end{bmatrix}$$

■

10.1.2 观测序列的生成过程

根据隐马尔可夫模型定义，可以将一个长度为 T 的观测序列 $O=(o_1,o_2,\cdots,o_T)$ 的生成过程描述如下：

算法 10.1（观测序列的生成）

输入：隐马尔可夫模型 $\lambda=(A,B,\pi)$，观测序列长度 T；

输出：观测序列 $O=(o_1,o_2,\cdots,o_T)$.

（1）按照初始状态分布 π 产生状态 i_1

（2）令 $t=1$

（3）按照状态 i_t 的观测概率分布 $b_{i_t}(k)$ 生成 o_t

（4）按照状态 i_t 的状态转移概率分布 $\{a_{i_t i_{t+1}}\}$ 产生状态 i_{t+1}， $i_{t+1}=1,2,\cdots,N$

（5）令 $t=t+1$；如果 $t<T$，转步 (3)；否则，终止 ■

10.1.3 隐马尔可夫模型的 3 个基本问题

隐马尔可夫模型有 3 个基本问题：

（1）概率计算问题．给定模型 $\lambda=(A,B,\pi)$ 和观测序列 $O=(o_1,o_2,\cdots,o_T)$，计算在模型 λ 下观测序列 O 出现的概率 $P(O|\lambda)$.

（2）学习问题．已知观测序列 $O=(o_1,o_2,\cdots,o_T)$，估计模型 $\lambda=(A,B,\pi)$ 参数，使得在该模型下观测序列概率 $P(O|\lambda)$ 最大．即用极大似然估计的方法估计参数．

（3）预测问题，也称为解码（decoding）问题．已知模型 $\lambda=(A,B,\pi)$ 和观测序列 $O=(o_1,o_2,\cdots,o_T)$，求对给定观测序列条件概率 $P(I|O)$ 最大的状态序列 $I=(i_1,i_2,\cdots,i_T)$．即给定观测序列，求最有可能的对应的状态序列．

下面各节将逐一介绍这些基本问题的解法．

10.2 概率计算算法

本节介绍计算观测序列概率 $P(O|\lambda)$ 的前向（forward）与后向（backward）算法．先介绍概念上可行但计算上不可行的直接计算法．

10.2.1 直接计算法

给定模型 $\lambda=(A,B,\pi)$ 和观测序列 $O=(o_1,o_2,\cdots,o_T)$，计算观测序列 O 出现的概率 $P(O|\lambda)$．最直接的方法是按概率公式直接计算．通过列举所有可能的长度为 T 的状态序列 $I=(i_1,i_2,\cdots,i_T)$，求各个状态序列 I 与观测序列 $O=(o_1,o_2,\cdots,o_T)$ 的联合概率 $P(O,I|\lambda)$，然后对所有可能的状态序列求和，得到 $P(O|\lambda)$．

状态序列 $I=(i_1,i_2,\cdots,i_T)$ 的概率是

$$P(I|\lambda)=\pi_{i_1}a_{i_1i_2}a_{i_2i_3}\cdots a_{i_{T-1}i_T} \tag{10.10}$$

对固定的状态序列 $I=(i_1,i_2,\cdots,i_T)$，观测序列 $O=(o_1,o_2,\cdots,o_T)$ 的概率是 $P(O|I,\lambda)$，

$$P(O|I,\lambda)=b_{i_1}(o_1)b_{i_2}(o_2)\cdots b_{i_T}(o_T) \tag{10.11}$$

O 和 I 同时出现的联合概率为

$$\begin{aligned}P(O,I|\lambda)&=P(O|I,\lambda)P(I|\lambda)\\&=\pi_{i_1}b_{i_1}(o_1)a_{i_1i_2}b_{i_2}(o_2)\cdots a_{i_{T-1}i_T}b_{i_T}(o_T)\end{aligned} \tag{10.12}$$

然后，对所有可能的状态序列 I 求和，得到观测序列 O 的概率 $P(O|\lambda)$，即

$$\begin{aligned}P(O|\lambda)&=\sum_I P(O|I,\lambda)P(I|\lambda)\\&=\sum_{i_1,i_2,\cdots,i_T}\pi_{i_1}b_{i_1}(o_1)a_{i_1i_2}b_{i_2}(o_2)\cdots a_{i_{T-1}i_T}b_{i_T}(o_T)\end{aligned} \tag{10.13}$$

但是，利用公式 (10.13) 计算量很大，是 $O(TN^T)$ 阶的，这种算法不可行．

下面介绍计算观测序列概率 $P(O|\lambda)$ 的有效算法：前向-后向算法（forward-backward algorithm）．

10.2.2 前向算法

首先定义前向概率．

定义 10.2（前向概率） 给定隐马尔可夫模型 λ，定义到时刻 t 部分观测序列为 $o_1,o_2,\cdots,o_t$ 且状态为 q_i 的概率为前向概率，记作

$$\alpha_t(i)=P(o_1,o_2,\cdots,o_t,i_t=q_i|\lambda) \tag{10.14}$$

可以递推地求得前向概率 $\alpha_t(i)$ 及观测序列概率 $P(O|\lambda)$．

算法 10.2（观测序列概率的前向算法）

输入：隐马尔可夫模型 λ，观测序列 O；

输出：观测序列概率 $P(O|\lambda)$．

（1）初值

$$\alpha_1(i)=\pi_i b_i(o_1)\,,\qquad i=1,2,\cdots,N \tag{10.15}$$

（2）递推　对 $t=1,2,\cdots,T-1$，

$$\alpha_{t+1}(i)=\left[\sum_{j=1}^{N}\alpha_t(j)a_{ji}\right]b_i(o_{t+1})\,,\qquad i=1,2,\cdots,N \tag{10.16}$$

（3）终止

$$P(O|\lambda)=\sum_{i=1}^{N}\alpha_T(i) \tag{10.17}$$ ■

前向算法，步骤（1）初始化前向概率，是初始时刻的状态 $i_1=q_i$ 和观测 o_1 的联合概率．步骤（2）是前向概率的递推公式，计算到时刻 $t+1$ 部分观测序列为 $o_1,o_2,\cdots,o_t,o_{t+1}$ 且在时刻 $t+1$ 处于状态 q_i 的前向概率，如图 10.1 所示．在式 (10.16) 的方括弧里，既然 $\alpha_t(j)$ 是到时刻 t 观测到 $o_1,o_2,\cdots,o_t$ 并在时刻 t 处于状态 q_j 的前向概率，那么乘积 $\alpha_t(j)a_{ji}$ 就是到时刻 t 观测到 $o_1,o_2,\cdots,o_t$ 并在时刻 t 处于状态 q_j 而在时刻 $t+1$ 到达状态 q_i 的联合概率．对这个乘积在时刻 t 的所有可能的 N 个状态 q_j 求和，其结果就是到时刻 t 观测为 $o_1,o_2,\cdots,o_t$ 并在时刻 $t+1$ 处于状态 q_i 的联合概率．方括弧里的值与观测概率 $b_i(o_{t+1})$ 的乘积恰好是到时刻 $t+1$ 观测到 $o_1,o_2,\cdots,o_t,o_{t+1}$ 并在时刻 $t+1$ 处于状态 q_i 的前向概率 $\alpha_{t+1}(i)$．步骤（3）给出 $P(O|\lambda)$ 的计算公式．因为

$$\alpha_T(i)=P(o_1,o_2,\cdots,o_T,i_T=q_i|\lambda)$$

所以

$$P(O|\lambda)=\sum_{i=1}^{N}\alpha_T(i)$$

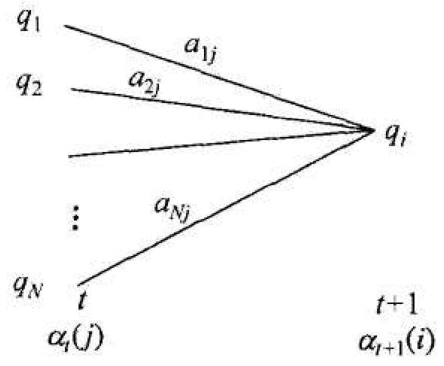

图 10.1　前向概率的递推公式

如图 10.2 所示，前向算法实际是基于“状态序列的路径结构”递推计算 $P(O|\lambda)$ 的算法．前向算法高效的关键是其局部计算前向概率，然后利用路径结构将前向概率“递推”到全局，得到 $P(O|\lambda)$．具体地，在时刻 $t=1$，计算 $\alpha_1(i)$ 的 N 个值

$(i=1,2,\cdots,N)$；在各个时刻 $t=1,2,\cdots,T-1$，计算 $\alpha_{t+1}(i)$ 的 N 个值$(i=1,2,\cdots,N)$，而且每个 $\alpha_{t+1}(i)$ 的计算利用前一时刻 N 个 $\alpha_t(j)$．减少计算量的原因在于每一次计算直接引用前一个时刻的计算结果，避免重复计算．这样，利用前向概率计算 $P(O|\lambda)$ 的计算量是 $O(N^2T)$ 阶的，而不是直接计算的 $O(TN^T)$ 阶．

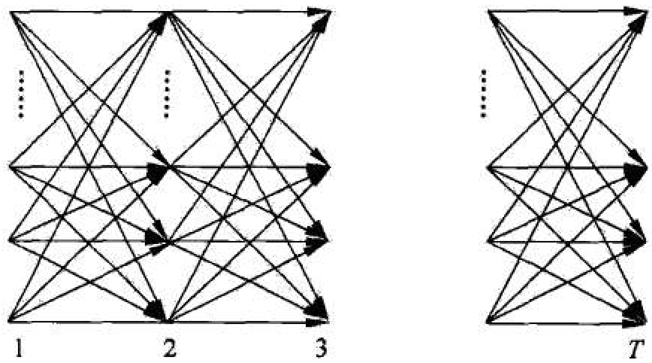

图 10.2 观测序列路径结构

例 10.2 考虑盒子和球模型 $\lambda=(A,B,\pi)$，状态集合 $Q=\{1,2,3\}$，观测集合 $V=\{红，白\}$，

$$A=\begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\end{bmatrix},\quad B=\begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\\ 0.7 & 0.3\end{bmatrix},\quad \pi=(0.2,0.4,0.4)^{\mathrm{T}}$$

设 $T=3$，$O=(红,白,红)$，试用前向算法计算 $P(O|\lambda)$．

解 按照算法 10.2

（1）计算初值

$$\alpha_1(1)=\pi_1 b_1(o_1)=0.10$$
$$\alpha_1(2)=\pi_2 b_2(o_1)=0.16$$
$$\alpha_1(3)=\pi_3 b_3(o_1)=0.28$$

（2）递推计算

$$\alpha_2(1)=\left[\sum_{i=1}^{3}\alpha_1(i)a_{i1}\right]b_1(o_2)=0.154\times 0.5=0.077$$

$$\alpha_2(2)=\left[\sum_{i=1}^{3}\alpha_1(i)a_{i2}\right]b_2(o_2)=0.184\times 0.6=0.1104$$

$$\alpha_2(3)=\left[\sum_{i=1}^{3}\alpha_1(i)a_{i3}\right]b_3(o_2)=0.202\times 0.3=0.0606$$

$$\alpha_3(1)=\left[\sum_{i=1}^{3}\alpha_2(i)a_{i1}\right]b_1(o_3)=0.04187$$

$$\alpha_3(2)=\left[\sum_{i=1}^{3}\alpha_2(i)a_{i2}\right]b_2(o_3)=0.03551$$

$$\alpha_3(3)=\left[\sum_{i=1}^{3}\alpha_2(i)a_{i3}\right]b_3(o_3)=0.05284$$

（3）终止

$$P(O|\lambda)=\sum_{i=1}^{3}\alpha_3(i)=0.13022$$ ■

10.2.3 后向算法

定义 10.3（后向概率） 给定隐马尔可夫模型 λ，定义在时刻 t 状态为 q_i 的条件下，从 $t+1$ 到 T 的部分观测序列为 $o_{t+1},o_{t+2},\cdots,o_T$ 的概率为后向概率，记作

$$\beta_t(i)=P(o_{t+1},o_{t+2},\cdots,o_T|i_t=q_i,\lambda) \tag{10.18}$$

可以用递推的方法求得后向概率 $\beta_t(i)$ 及观测序列概率 $P(O|\lambda)$.

算法 10.3 (观测序列概率的后向算法)

输入：隐马尔可夫模型 λ，观测序列 O；

输出：观测序列概率 $P(O|\lambda)$.

（1）

$$\beta_T(i)=1,\quad i=1,2,\cdots,N \tag{10.19}$$

（2）对 $t=T-1,T-2,\cdots,1$

$$\beta_t(i)=\sum_{j=1}^{N}a_{ij}b_j(o_{t+1})\beta_{t+1}(j),\quad i=1,2,\cdots,N \tag{10.20}$$

（3）

$$P(O|\lambda)=\sum_{i=1}^{N}\pi_i b_i(o_1)\beta_1(i) \tag{10.21}$$ ■

步骤（1）初始化后向概率，对最终时刻的所有状态 q_i 规定 $\beta_T(i)=1$. 步骤（2）是后向概率的递推公式. 如图 10.3 所示，为了计算在时刻 t 状态为 q_i 条件下时刻 $t+1$ 之后的观测序列为 $o_{t+1},o_{t+2},\cdots,o_T$ 的后向概率 $\beta_t(i)$，只需考虑在时刻 $t+1$ 所有

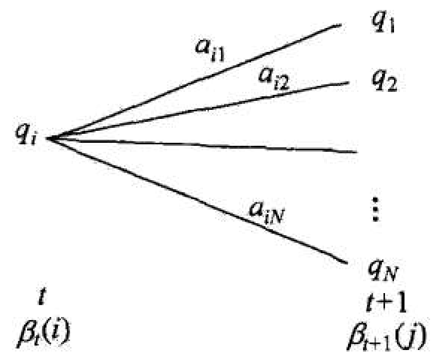

图 10.3 后向概率递推公式

可能的 N 个状态 q_j 的转移概率（即 a_{ij} 项），以及在此状态下的观测 o_{t+1} 的观测概率（即 $b_j(o_{t+1})$ 项），然后考虑状态 q_j 之后的观测序列的后向概率（即 $\beta_{t+1}(j)$ 项）．步骤（3）求 $P(O|\lambda)$ 的思路与步骤（2）一致，只是初始概率 π_i 代替转移概率．

利用前向概率和后向概率的定义可以将观测序列概率 $P(O|\lambda)$ 统一写成

$$P(O|\lambda)=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)\text{，}\quad t=1,2,\cdots,T-1 \tag{10.22}$$

此式当 $t=1$ 和 $t=T-1$ 时分别为式(10.17)和式(10.21)．

10.2.4 一些概率与期望值的计算

利用前向概率和后向概率，可以得到关于单个状态和两个状态概率的计算公式．

1．给定模型 λ 和观测 O，在时刻 t 处于状态 q_i 的概率．记

$$\gamma_t(i)=P(i_t=q_i\,|\,O,\lambda) \tag{10.23}$$

可以通过前向后向概率计算．事实上，

$$\gamma_t(i)=P(i_t=q_i\,|\,O,\lambda)=\frac{P(i_t=q_i,O\,|\,\lambda)}{P(O\,|\,\lambda)}$$

由前向概率 $\alpha_t(i)$ 和后向概率 $\beta_t(i)$ 定义可知：

$$\alpha_t(i)\beta_t(i)=P(i_t=q_i,O\,|\,\lambda)$$

于是得到：

$$\gamma_t(i)=\frac{\alpha_t(i)\beta_t(i)}{P(O\,|\,\lambda)}=\frac{\alpha_t(i)\beta_t(i)}{\sum\limits_{j=1}^{N}\alpha_t(j)\beta_t(j)} \tag{10.24}$$

2．给定模型 λ 和观测 O，在时刻 t 处于状态 q_i 且在时刻 $t+1$ 处于状态 q_j 的概率．记

$$\xi_t(i,j)=P(i_t=q_i,i_{t+1}=q_j\,|\,O,\lambda) \tag{10.25}$$

可以通过前向后向概率计算：

$$\xi_t(i,j)=\frac{P(i_t=q_i,i_{t+1}=q_j,O\,|\,\lambda)}{P(O\,|\,\lambda)}=\frac{P(i_t=q_i,i_{t+1}=q_j,O\,|\,\lambda)}{\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N}P(i_t=q_i,i_{t+1}=q_j,O\,|\,\lambda)}$$

而

$$P(i_t=q_i,i_{t+1}=q_j,O\,|\,\lambda)=\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)$$

所以

$$\xi_t(i,j)=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)} \tag{10.26}$$

3．将 $\gamma_t(i)$ 和 $\xi_t(i,j)$ 对各个时刻 t 求和，可以得到一些有用的期望值：

（1）在观测 O 下状态 i 出现的期望值

$$\sum_{t=1}^{T}\gamma_t(i) \tag{10.27}$$

（2）在观测 O 下由状态 i 转移的期望值

$$\sum_{t=1}^{T-1}\gamma_t(i) \tag{10.28}$$

（3）在观测 O 下由状态 i 转移到状态 j 的期望值

$$\sum_{t=1}^{T-1}\xi_t(i,j) \tag{10.29}$$

10.3 学 习 算 法

隐马尔可夫模型的学习，根据训练数据是包括观测序列和对应的状态序列还是只有观测序列，可以分别由监督学习与非监督学习实现．本节首先介绍监督学习算法，而后介绍非监督学习算法——Baum-Welch 算法（也就是 EM 算法）．

10.3.1 监督学习方法

假设已给训练数据包含 S 个长度相同的观测序列和对应的状态序列 $\{(O_1,I_1),(O_2,I_2),\cdots,(O_S,I_S)\}$，那么可以利用极大似然估计法来估计隐马尔可夫模型的参数．具体方法如下．

1．转移概率 a_{ij} 的估计

设样本中时刻 t 处于状态 i 时刻 $t+1$ 转移到状态 j 的频数为 A_{ij}，那么状态转移概率 a_{ij} 的估计是

$$\hat{a}_{ij}=\frac{A_{ij}}{\sum_{j=1}^{N}A_{ij}}, \quad i=1,2,\cdots,N；\ j=1,2,\cdots,N \tag{10.30}$$

2．观测概率 $b_j(k)$ 的估计

设样本中状态为 j 并观测为 k 的频数是 B_{jk}，那么状态为 j 观测为 k 的概率

$b_j(k)$ 的估计是

$$\hat{b}_j(k)=\frac{B_{jk}}{\sum_{k=1}^{M}B_{jk}}, \quad j=1,2,\cdots,N;\ k=1,2,\cdots,M \tag{10.31}$$

3．初始状态概率 π_i 的估计 $\hat{\pi}_i$ 为 S 个样本中初始状态为 q_i 的频率

由于监督学习需要使用训练数据，而人工标注训练数据往往代价很高，有时就会利用非监督学习的方法.

10.3.2 Baum-Welch 算法

假设给定训练数据只包含 S 个长度为 T 的观测序列 $\{O_1,O_2,\cdots,O_S\}$ 而没有对应的状态序列，目标是学习隐马尔可夫模型 $\lambda=(A,B,\pi)$ 的参数. 我们将观测序列数据看作观测数据 O，状态序列数据看作不可观测的隐数据 I，那么隐马尔可夫模型事实上是一个含有隐变量的概率模型

$$P(O|\lambda)=\sum_I P(O|I,\lambda)P(I|\lambda) \tag{10.32}$$

它的参数学习可以由 EM 算法实现.

1．确定完全数据的对数似然函数

所有观测数据写成 $O=(o_1,o_2,\cdots,o_T)$，所有隐数据写成 $I=(i_1,i_2,\cdots,i_T)$，完全数据是 $(O,I)=(o_1,o_2,\cdots,o_T,i_1,i_2,\cdots,i_T)$. 完全数据的对数似然函数是 $\log P(O,I|\lambda)$.

2．EM 算法的 E 步：求 Q 函数 $Q(\lambda,\bar{\lambda})$①

$$Q(\lambda,\bar{\lambda})=\sum_I \log P(O,I|\lambda)P(O,I|\bar{\lambda}) \tag{10.33}$$

其中，$\bar{\lambda}$ 是隐马尔可夫模型参数的当前估计值，λ 是要极大化的隐马尔可夫模型参数.

$$P(O,I|\lambda)=\pi_{i_1}b_{i_1}(o_1)a_{i_1i_2}b_{i_2}(o_2)\cdots a_{i_{T-1}i_T}b_{i_T}(o_T)$$

于是函数 $Q(\lambda,\bar{\lambda})$ 可以写成：

$$\begin{aligned}Q(\lambda,\bar{\lambda})=&\sum_I \log\pi_{i_1}P(O,I|\bar{\lambda})\\&+\sum_I\left(\sum_{t=1}^{T-1}\log a_{i_ti_{t+1}}\right)P(O,I|\bar{\lambda})+\sum_I\left(\sum_{t=1}^{T}\log b_{i_t}(o_t)\right)P(O,I|\bar{\lambda})\end{aligned} \tag{10.34}$$

式中求和都是对所有训练数据的序列总长度 T 进行的.

① 按照 Q 函数的定义

$$Q(\lambda,\bar{\lambda})=E_I[\log P(O,I|\lambda)|O,\bar{\lambda}]$$

式 (10.33) 略去了对 λ 而言的常数因子 $1/P(O|\bar{\lambda})$.

3．EM 算法的 M 步：极大化 Q 函数 $Q(\lambda,\bar{\lambda})$ 求模型参数 A,B,π

由于要极大化的参数在式 (10.34) 中单独地出现在 3 个项中，所以只需对各项分别极大化.

（1）式 (10.34) 的第 1 项可以写成：

$$\sum_I \log \pi_{i_0} P(O,I\mid\bar{\lambda})=\sum_{i=1}^N \log \pi_i P(O,i_1=i\mid\bar{\lambda})$$

注意到 π_i 满足约束条件 $\sum_{i=1}^N \pi_i=1$，利用拉格朗日乘子法，写出拉格朗日函数：

$$\sum_{i=1}^N \log \pi_i P(O,i_1=i\mid\bar{\lambda})+\gamma\left(\sum_{i=1}^N \pi_i-1\right)$$

对其求偏导数并令结果为 0

$$\frac{\partial}{\partial \pi_i}\left[\sum_{i=1}^N \log \pi_i P(O,i_1=i\mid\bar{\lambda})+\gamma\left(\sum_{i=1}^N \pi_i-1\right)\right]=0 \tag{10.35}$$

得

$$P(O,i_1=i\mid\bar{\lambda})+\gamma\pi_i=0$$

对 i 求和得到 γ

$$\gamma=-P(O\mid\bar{\lambda})$$

代入式 (10.35) 即得

$$\pi_i=\frac{P(O,i_1=i\mid\bar{\lambda})}{P(O\mid\bar{\lambda})} \tag{10.36}$$

（2）式 (10.34) 的第 2 项可以写成

$$\sum_I\left(\sum_{t=1}^{T-1}\log a_{i_t i_{t+1}}\right)P(O,I\mid\bar{\lambda})=\sum_{i=1}^N\sum_{j=1}^N\sum_{t=1}^{T-1}\log a_{ij}P(O,i_t=i,i_{t+1}=j\mid\bar{\lambda})$$

类似第 1 项，应用具有约束条件 $\sum_{j=1}^N a_{ij}=1$ 的拉格朗日乘子法可以求出

$$a_{ij}=\frac{\sum_{t=1}^{T-1}P(O,i_t=i,i_{t+1}=j\mid\bar{\lambda})}{\sum_{t=1}^{T-1}P(O,i_t=i\mid\bar{\lambda})} \tag{10.37}$$

（3）式 (10.34) 的第 3 项为

$$\sum_I\left(\sum_{t=1}^T\log b_{i_t}(o_t)\right)P(O,I\mid\bar{\lambda})=\sum_{j=1}^N\sum_{t=1}^T\log b_j(o_t)P(O,i_t=j\mid\bar{\lambda})$$

同样用拉格朗日乘子法，约束条件是$\sum_{k=1}^{M} b_j(k)=1$. 注意，只有在$o_t=v_k$时$b_j(o_t)$对$b_j(k)$的偏导数才不为 0，以$I(o_t=v_k)$表示. 求得

$$b_j(k)=\frac{\sum_{t=1}^{T} P(O,i_t=j\mid\bar{\lambda})I(o_t=v_k)}{\sum_{t=1}^{T} P(O,i_t=j\mid\bar{\lambda})} \tag{10.38}$$

10.3.3 Baum-Welch 模型参数估计公式

将式 (10.36) ～式 (10.38) 中的各概率分别用$\gamma_t(i)$，$\xi_t(i,j)$表示，则可将相应的公式写成：

$$a_{ij}=\frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{10.39}$$

$$b_j(k)=\frac{\sum_{t=1,o_t=v_k}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)} \tag{10.40}$$

$$\pi_i=\gamma_1(i) \tag{10.41}$$

其中，$\gamma_t(i)$，$\xi_t(i,j)$分别由式 (10.24) 及式 (10.26) 给出. 式 (10.39) ～式 (10.41) 就是 Baum-Welch 算法（Baum-Welch algorithm），它是 EM 算法在隐马尔可夫模型学习中的具体实现，由 Baum 和 Welch 提出.

算法 10.4（Baum-Welch 算法）

输入：观测数据$O=(o_1,o_2,\cdots,o_T)$；

输出：隐马尔可夫模型参数.

（1）初始化

对$n=0$，选取$a_{ij}^{(0)}$，$b_j(k)^{(0)}$，$\pi_i^{(0)}$，得到模型$\lambda^{(0)}=(A^{(0)},B^{(0)},\pi^{(0)})$.

（2）递推. 对$n=1,2,\cdots$，

$$a_{ij}^{(n+1)}=\frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}$$

$$b_j(k)^{(n+1)} = \frac{\sum_{t=1, o_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

$$\pi_i^{(n+1)} = \gamma_1(i)$$

右端各值按观测 $O = (o_1, o_2, \cdots, o_T)$ 和模型 $\lambda^{(n)} = (A^{(n)}, B^{(n)}, \pi^{(n)})$ 计算．式中 $\gamma_t(i)$，$\xi_t(i,j)$ 由式 (10.24) 和式 (10.26) 给出．

（3）终止．得到模型参数 $\lambda^{(n+1)} = (A^{(n+1)}, B^{(n+1)}, \pi^{(n+1)})$． ■

10.4 预测算法

下面介绍隐马尔可夫模型预测的两种算法：近似算法与维特比算法（Viterbi algorithm）．

10.4.1 近似算法

近似算法的想法是，在每个时刻 t 选择在该时刻最有可能出现的状态 i_t^*，从而得到一个状态序列 $I^* = (i_1^*, i_2^*, \cdots, i_T^*)$，将它作为预测的结果．

给定隐马尔可夫模型 λ 和观测序列 O，在时刻 t 处于状态 q_i 的概率 $\gamma_t(i)$ 是

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \tag{10.42}$$

在每一时刻 t 最有可能的状态 i_t^* 是

$$i_t^* = \arg\max_{1 \leqslant i \leqslant N} [\gamma_t(i)], \quad t = 1, 2, \cdots, T \tag{10.43}$$

从而得到状态序列 $I^* = (i_1^*, i_2^*, \cdots, i_T^*)$．

近似算法的优点是计算简单，其缺点是不能保证预测的状态序列整体是最有可能的状态序列，因为预测的状态序列可能有实际不发生的部分．事实上，上述方法得到的状态序列中有可能存在转移概率为 0 的相邻状态，即对某些 i, j，$a_{ij} = 0$ 时．尽管如此，近似算法仍然是有用的．

10.4.2 维特比算法

维特比算法实际是用动态规划解隐马尔可夫模型预测问题，即用动态规划（dynamic programming）求概率最大路径（最优路径）．这时一条路径对应着一个状态序列．

根据动态规划原理，最优路径具有这样的特性：如果最优路径在时刻 t 通过结点 i_t^*，那么这一路径从结点 i_t^* 到终点 i_T^* 的部分路径，对于从 i_t^* 到 i_T^* 的所有可能

的部分路径来说，必须是最优的．因为假如不是这样，那么从 i_t^* 到 i_T^* 就有另一条更好的部分路径存在，如果把它和从 i_1^* 到达 i_t^* 的部分路径连接起来，就会形成一条比原来的路径更优的路径，这是矛盾的．依据这一原理，我们只需从时刻 $t=1$ 开始，递推地计算在时刻 t 状态为 i 的各条部分路径的最大概率，直至得到时刻 $t=T$ 状态为 i 的各条路径的最大概率．时刻 $t=T$ 的最大概率即为最优路径的概率 P^*，最优路径的终结点 i_T^* 也同时得到．之后，为了找出最优路径的各个结点，从终结点 i_T^* 开始，由后向前逐步求得结点 $i_{T-1}^*,\cdots,i_1^*$，得到最优路径 $I^*=(i_1^*,i_2^*,\cdots,i_T^*)$．这就是维特比算法．

首先导入两个变量 δ 和 ψ．定义在时刻 t 状态为 i 的所有单个路径 $(i_1,i_2,\cdots,i_t)$ 中概率最大值为

$$\delta_t(i)=\max_{i_1,i_2,\cdots,i_{t-1}} P(i_t=i,i_{t-1},\cdots,i_1,o_t,\cdots,o_1\mid\lambda),\quad i=1,2,\cdots,N \tag{10.44}$$

由定义可得变量 δ 的递推公式：

$$\begin{aligned}\delta_{t+1}(i)&=\max_{i_1,i_2,\cdots,i_t} P(i_{t+1}=i,i_t,\cdots,i_1,o_{t+1},\cdots,o_1\mid\lambda)\\&=\max_{1\leqslant j\leqslant N}[\delta_t(j)a_{ji}]b_i(o_{t+1}),\qquad i=1,2,\cdots,N;\ t=1,2,\cdots,T-1\end{aligned} \tag{10.45}$$

定义在时刻 t 状态为 i 的所有单个路径 $(i_1,i_2,\cdots,i_{t-1},i)$ 中概率最大的路径的第 $t-1$ 个结点为

$$\psi_t(i)=\arg\max_{1\leqslant j\leqslant N}[\delta_{t-1}(j)a_{ji}],\qquad i=1,2,\cdots,N \tag{10.46}$$

下面介绍维特比算法．

算法 10.5（**维特比算法**）

输入：模型 $\lambda=(A,B,\pi)$ 和观测 $O=(o_1,o_2,\cdots,o_T)$；

输出：最优路径 $I^*=(i_1^*,i_2^*,\cdots,i_T^*)$．

（1）初始化

$$\delta_1(i)=\pi_i b_i(o_1),\qquad i=1,2,\cdots,N$$
$$\psi_1(i)=0,\qquad i=1,2,\cdots,N$$

（2）递推．对 $t=2,3,\cdots,T$

$$\delta_t(i)=\max_{1\leqslant j\leqslant N}[\delta_{t-1}(j)a_{ji}]b_i(o_t),\qquad i=1,2,\cdots,N$$
$$\psi_t(i)=\arg\max_{1\leqslant j\leqslant N}[\delta_{t-1}(j)a_{ji}],\qquad i=1,2,\cdots,N$$

（3）终止

$$P^*=\max_{1\leqslant i\leqslant N}\delta_T(i)$$
$$i_T^*=\arg\max_{1\leqslant i\leqslant N}[\delta_T(i)]$$

（4）最优路径回溯．对 $t=T-1,T-2,\cdots,1$

$$i_t^* = \psi_{t+1}(i_{t+1}^*)$$

求得最优路径 $I^*=(i_1^*,i_2^*,\cdots,i_T^*)$． ■

下面通过一个例子来说明维特比算法．

例 10.3 例 10.2 的模型 $\lambda=(A,B,\pi)$，

$$A=\begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\end{bmatrix},\quad B=\begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\\ 0.7 & 0.3\end{bmatrix},\quad \pi=(0.2,0.4,0.4)^{\mathrm{T}}$$

已知观测序列 $O=(红,白,红)$，试求最优状态序列，即最优路径 $I^*=(i_1^*,i_2^*,i_3^*)$．

解 如图 10.4 所示，要在所有可能的路径中选择一条最优路径，按照以下步骤处理：

（1）初始化．在 $t=1$ 时，对每一个状态 i，$i=1,2,3$，求状态为 i 观测 o_1 为红的概率，记此概率为 $\delta_1(i)$，则

$$\delta_1(i)=\pi_i b_i(o_1)=\pi_i b_i(红),\quad i=1,2,3$$

代入实际数据

$$\delta_1(1)=0.10,\quad \delta_1(2)=0.16,\quad \delta_1(3)=0.28$$

记 $\psi_1(i)=0$，$i=1,2,3$．

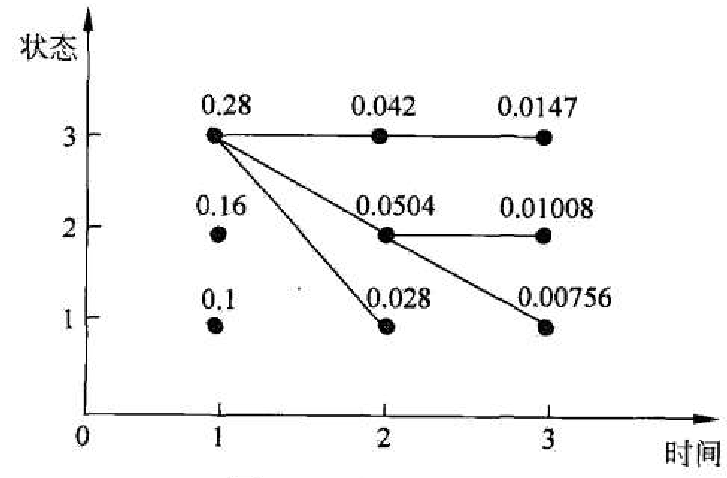

图 10.4 求最优路径

（2）在 $t=2$ 时，对每个状态 i，$i=1,2,3$，求在 $t=1$ 时状态为 j 观测为红并在 $t=2$ 时状态为 i 观测 o_2 为白的路径的最大概率，记此最大概率为 $\delta_2(i)$，则

$$\delta_2(i)=\max_{1\leqslant j\leqslant 3}[\delta_1(j)a_{ji}]b_i(o_2)$$

同时，对每个状态 i ， $i=1,2,3$ ，记录概率最大路径的前一个状态 j :

$$\psi_2(i)=\arg\max_{1\leqslant j\leqslant 3}[\delta_1(j)a_{ji}], \qquad i=1,2,3$$

计算:

$$\begin{aligned}\delta_2(1)&=\max_{1\leqslant j\leqslant 3}[\delta_1(j)a_{j1}]b_1(o_2)\\&=\max_j\{0.10\times 0.5, 0.16\times 0.3, 0.28\times 0.2\}\times 0.5\\&=0.028\\\psi_2(1)&=3\\\delta_2(2)&=0.0504,\quad \psi_2(2)=3\\\delta_2(3)&=0.042,\quad \psi_2(3)=3\end{aligned}$$

同样，在 $t=3$ 时，

$$\begin{aligned}\delta_3(i)&=\max_{1\leqslant j\leqslant 3}[\delta_2(j)a_{ji}]b_i(o_3)\\\psi_3(i)&=\arg\max_{1\leqslant j\leqslant 3}[\delta_2(j)a_{ji}]\\\delta_3(1)&=0.00756,\quad \psi_3(1)=2\\\delta_3(2)&=0.01008,\quad \psi_3(2)=2\\\delta_3(3)&=0.0147,\quad \psi_3(3)=3\end{aligned}$$

（3）以 P^* 表示最优路径的概率，则

$$P^*=\max_{1\leqslant i\leqslant 3}\delta_3(i)=0.0147$$

最优路径的终点是 i_3^* :

$$i_3^*=\arg\max_i[\delta_3(i)]=3$$

（4）由最优路径的终点 i_3^* ，逆向找到 i_2^*,i_1^* :

$$\text{在 } t=2 \text{ 时，}\quad i_2^*=\psi_3(i_3^*)=\psi_3(3)=3$$

$$\text{在 } t=1 \text{ 时，}\quad i_1^*=\psi_2(i_2^*)=\psi_2(3)=3$$

于是求得最优路径，即最优状态序列 $I^*=(i_1^*,i_2^*,i_3^*)=(3,3,3)$. ■

本章概要

1．隐马尔可夫模型是关于时序的概率模型，描述由一个隐藏的马尔可夫链随机生成不可观测的状态的序列，再由各个状态随机生成一个观测而产生观测的序列的过程.

隐马尔可夫模型由初始状态概率向量π、状态转移概率矩阵A和观测概率矩阵B决定．因此，隐马尔可夫模型可以写成$\lambda=(A,B,\pi)$．

隐马尔可夫模型是一个生成模型，表示状态序列和观测序列的联合分布，但是状态序列是隐藏的，不可观测的．

隐马尔可夫模型可以用于标注，这时状态对应着标记．标注问题是给定观测序列预测其对应的标记序列．

2．概率计算问题．给定模型$\lambda=(A,B,\pi)$和观测序列$O=(o_1,o_2,\cdots,o_T)$，计算在模型λ下观测序列O出现的概率$P(O|\lambda)$．前向-后向算法是通过递推地计算前向-后向概率可以高效地进行隐马尔可夫模型的概率计算．

3．学习问题．已知观测序列$O=(o_1,o_2,\cdots,o_T)$，估计模型$\lambda=(A,B,\pi)$参数，使得在该模型下观测序列概率$P(O|\lambda)$最大．即用极大似然估计的方法估计参数．Baum-Welch 算法，也就是 EM 算法可以高效地对隐马尔可夫模型进行训练．它是一种非监督学习算法．

4．预测问题．已知模型$\lambda=(A,B,\pi)$和观测序列$O=(o_1,o_2,\cdots,o_T)$，求对给定观测序列条件概率$P(I|O)$最大的状态序列$I=(i_1,i_2,\cdots,i_T)$．维特比算法应用动态规划高效地求解最优路径，即概率最大的状态序列．

继续阅读

隐马尔可夫模型的介绍可见文献[1, 2]，特别地，文献[1]是经典的介绍性论文．关于 Baum-Welch 算法可见文献[3, 4]．可以认为概率上下文无关文法（probabilistic context-free grammar）是隐马尔可夫模型的一种推广，隐马尔可夫模型的不可观测数据是状态序列，而概率上下文无关文法的不可观测数据是上下文无关文法树[5]．动态贝叶斯网络（dynamic Bayesian network）是定义在时序数据上的贝叶斯网络，它包含隐马尔可夫模型，是一种特例[6]．

习　题

10.1 给定盒子和球组成的隐马尔可夫模型$\lambda=(A,B,\pi)$，其中，

$$A=\begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\end{bmatrix},\quad B=\begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\\ 0.7 & 0.3\end{bmatrix},\quad \pi=(0.2,0.4,0.4)^{\mathrm{T}}$$

设$T=4$，$O=(红,白,红,白)$，试用后向算法计算$P(O|\lambda)$．

10.2 考虑盒子和球组成的隐马尔可夫模型 $\lambda=(A,B,\pi)$，其中，

$$A=\begin{bmatrix}0.5 & 0.1 & 0.4\\0.3 & 0.5 & 0.2\\0.2 & 0.2 & 0.6\end{bmatrix},\quad B=\begin{bmatrix}0.5 & 0.5\\0.4 & 0.6\\0.7 & 0.3\end{bmatrix},\quad \pi=(0.2,0.3,0.5)^{\mathrm{T}}$$

设 $T=8$，$O=$(红,白,红,红,白,红,白,白)，用前向后向概率计算 $P(i_4=q_3|O,\lambda)$.

10.3 在习题 10.1 中，试用维特比算法求最优路径 $I^*=(i_1^*,i_2^*,i_3^*,i_4^*)$.

10.4 试用前向概率和后向概率推导

$$P(O|\lambda)=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j),\quad t=1,2,\cdots,T-1$$

10.5 比较维特比算法中变量 δ 的计算和前向算法中变量 α 的计算的主要区别.

参考文献

[1] Rabiner L, Juang B. An introduction to hidden markov Models. IEEE ASSP Magazine, January 1986

[2] Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of IEEE, 1989

[3] Baum L, et al. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. Annals of Mathematical Statistics, 1970, 41: 164–171

[4] Bilmes JA. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. http://ssli.ee.washington.edu/~bilmes/mypubs/bilmes1997-em.pdf

[5] Lari K, Young SJ. Applications of stochastic context-free grammars using the Inside-Outside algorithm, Computer Speech & Language, 1991, 5(3): 237–257

[6] Ghahramani Z. Learning Dynamic Bayesian Networks. Lecture Notes in Computer Science, Vol. 1387, 1997, 168–197