思考：给定文本标注词性

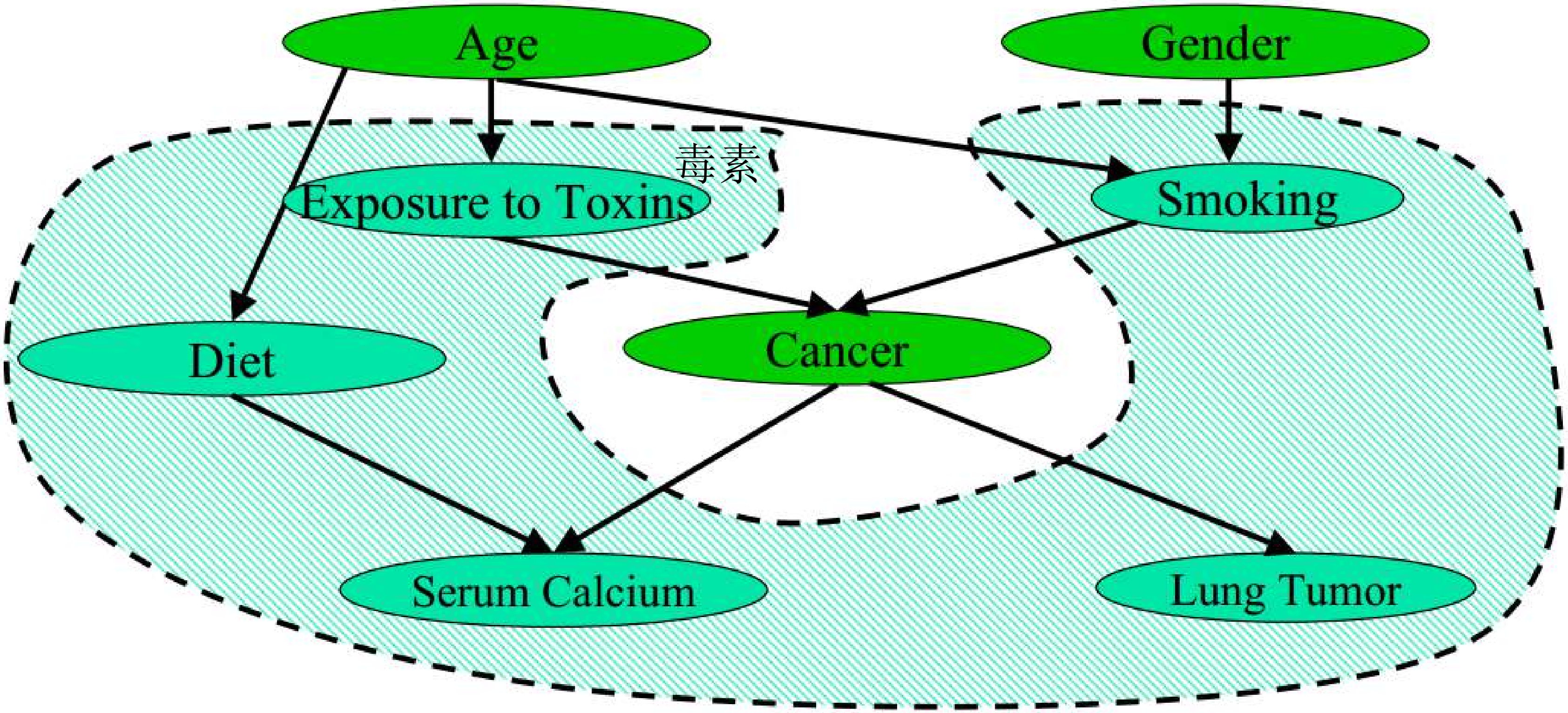
[PRP He] [VBZ reckons] [DT the] [JJ current] [NN account] [NN deficit] [MD will] [VB narrow] [TO to] [RB only] [# #] [CD 1.8] [CD billion] [IN in] [NNP September] [. .]

- 他估算当前的赤字总额在9月份仅仅降低到18亿。
- NN、NNS、NNP、NNPS、PRP、DT、JJ分别代表普通名词单数形式、普通名词复数形式、专有名词单数形式、专有名词复数形式、代词、限定词、形容词

复习：Markov Blanket

- 一个结点的Markov Blanket是一个集合，在这个集合中的结点都给定的条件下，该结点条件独立于其他所有结点。
- 即：一个结点的Markov Blanket是它的parents,children以及spouses(孩子的其他parent)

Markov Blanket

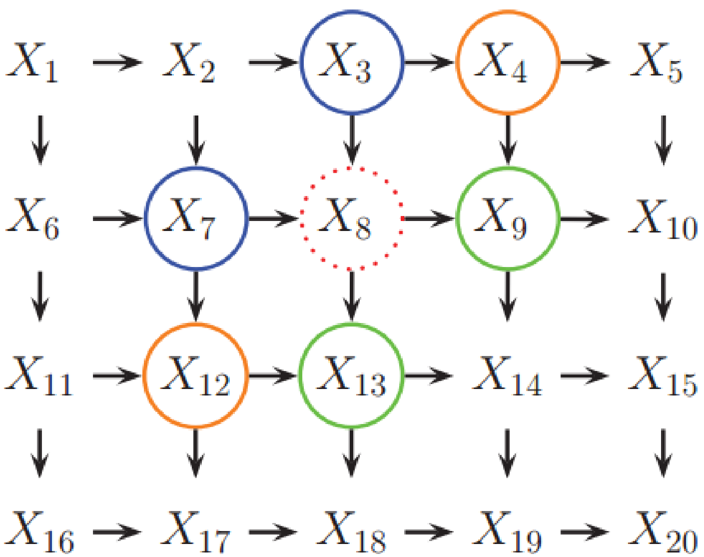


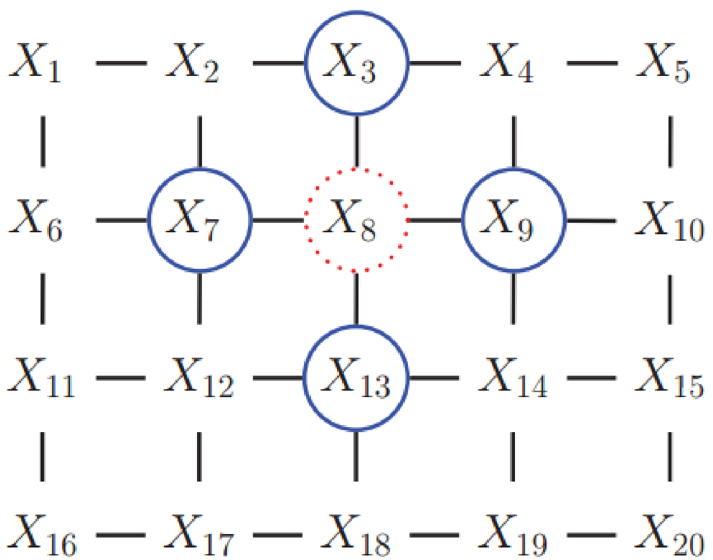
[Breese & Koller, 97]

补充知识：Serum Calcium(血清钙浓度)高于2.75mmol/L即为高钙血症。许多恶性肿瘤可并发高钙血症。以乳腺癌、骨肿瘤、肺癌、胃癌、卵巢癌、多发性骨髓瘤、急性淋巴细胞白血病等较为多见，其中乳腺癌约1/3 可发生高钙血症。

图像模型

- 考察X8的马尔科夫毯(Markov blanket)





无向图模型

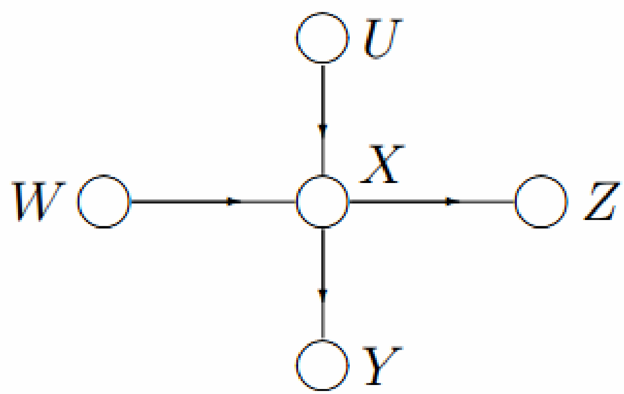
- 有向图模型，又称作贝叶斯网络(Directed Graphical Models, DGM, Bayesian Network)
- 在有些情况下，强制对某些结点之间的边增加方向是不合适的。
- 使用没有方向的无向边，形成了无向图模型(Undirected Graphical Model,UGM), 又被称为马尔科夫随机场或者马尔科夫网络(Markov Random Field, MRF or Markov network)

条件随机场

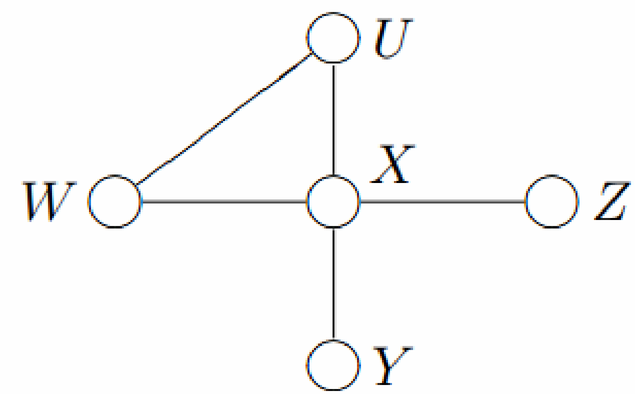
- 设X=(X1,X2…Xn)和Y=(Y1,Y2…Ym)都是联合随机变量，若随机变量Y构成一个无向图G=(V,E)表示的马尔科夫随机场(MRF)，则条件概率分布P(Y|X)称为条件随机场(Conditional Random Field, CRF)

- *注：大量文献将MRF和CRF混用，包括经典著作。后面将考察为何会有该混用。*

DGM转换成UGM

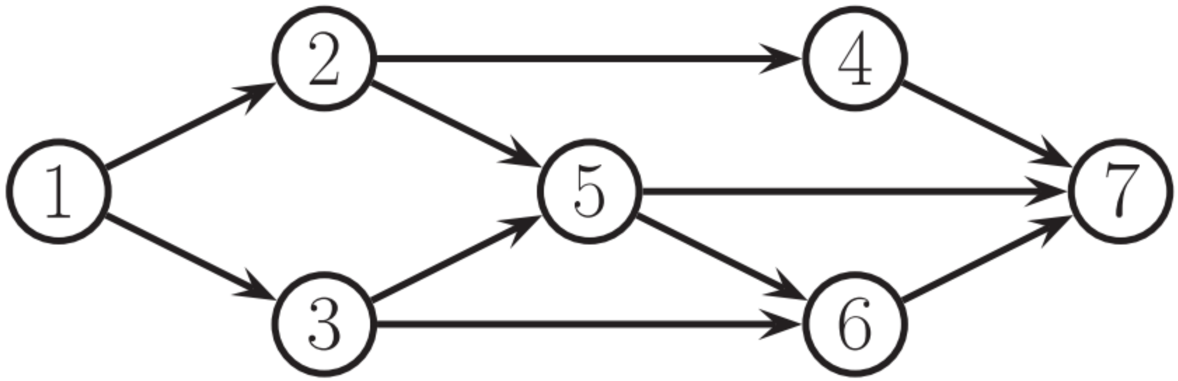


Bayesian network.

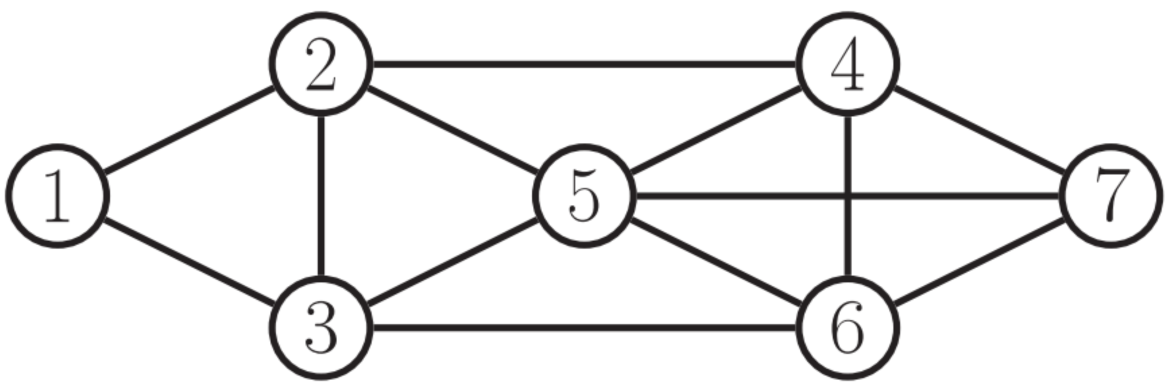


Markov random field.

DGM转换成UGM

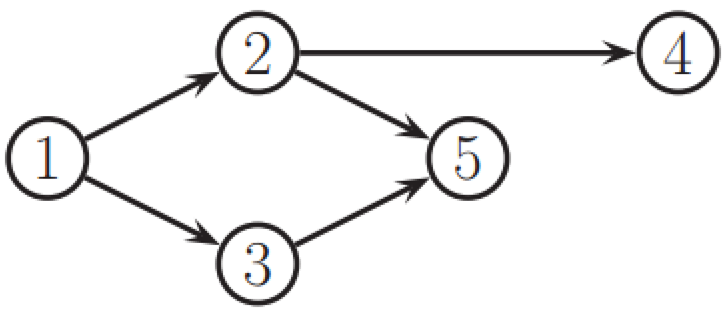


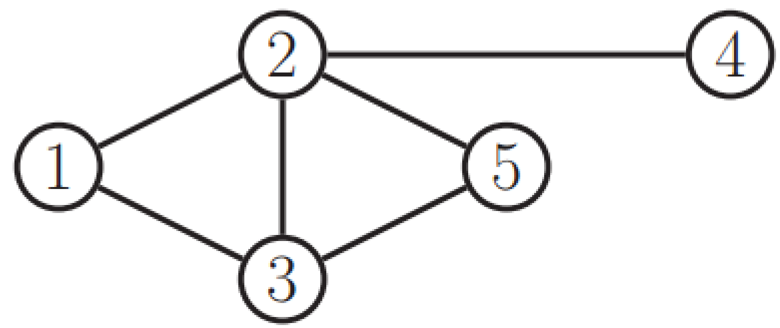
$$4 \perp 5 | 2$$



条件独立的破坏

- 靠考察是否有$A \perp B|C$，则计算U的祖先图(ancestral graph)：$U = A \cup B \cup C$





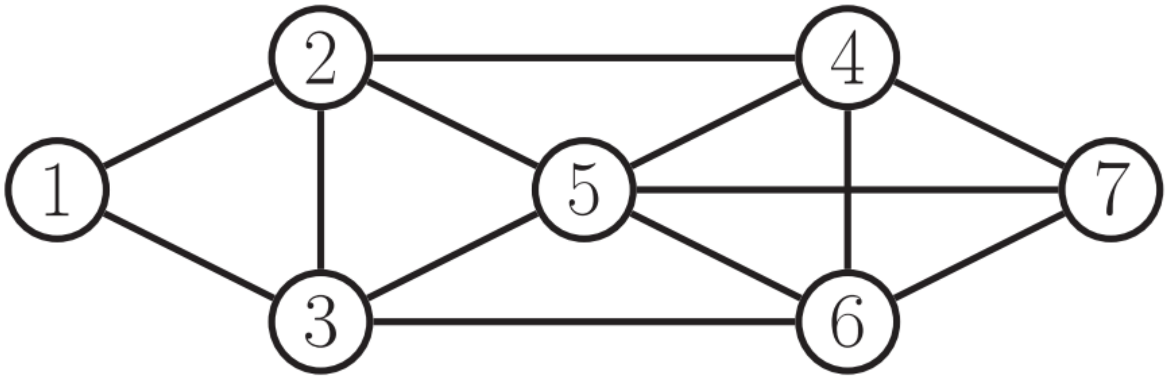
$$4 \perp 5|2$$

MRF的性质

- 成对马尔科夫性
  - parewise Markov property
- 局部马尔科夫性
  - local Markov property
- 全局马尔科夫性
  - global Markov property

- 表述说明：随机变量Y=(Y1,Y2…Ym)构成无向图G=(V,E)，结点v对应的随机变量是Yv。

考察结点间的独立性



Pairwise $1 \perp 7 | \text{rest}$

Local $1 \perp \text{rest} | 2, 3$

Global $1, 2 \perp 6, 7 | 3, 4, 5$

成对马尔科夫性

- 设u和v是无向图G中任意两个没有边直接连接的结点，G中其他结点的集合记做O；则在给定随机变量Yo的条件下，随机变量Yu和Yv条件独立。
- 即：$P(Y_u, Y_v|Y_o) = P(Y_u|Y_o) * P(Y_v|Y_o)$

局部马尔科夫性

- 设v是无向图G中任意一个结点，W是与v有边相连的所有结点，G中其他结点记做O；则在给定随机变量Yw的条件下，随机变量Yv和Yo条件独立。
- 即：P(Yv,Yo|Yw)= P(Yv|Yw)* P(Yo|Yw)

全局马尔科夫性

- 设结点集合A，B是在无向图G中被结点集合C分开的任意结点集合，则在给定随机变量Y_C的条件下，随机变量Y_A和Y_B条件独立。
- 即：$P(Y_A, Y_B | Y_C) = P(Y_A | Y_C) * P(Y_B | Y_C)$

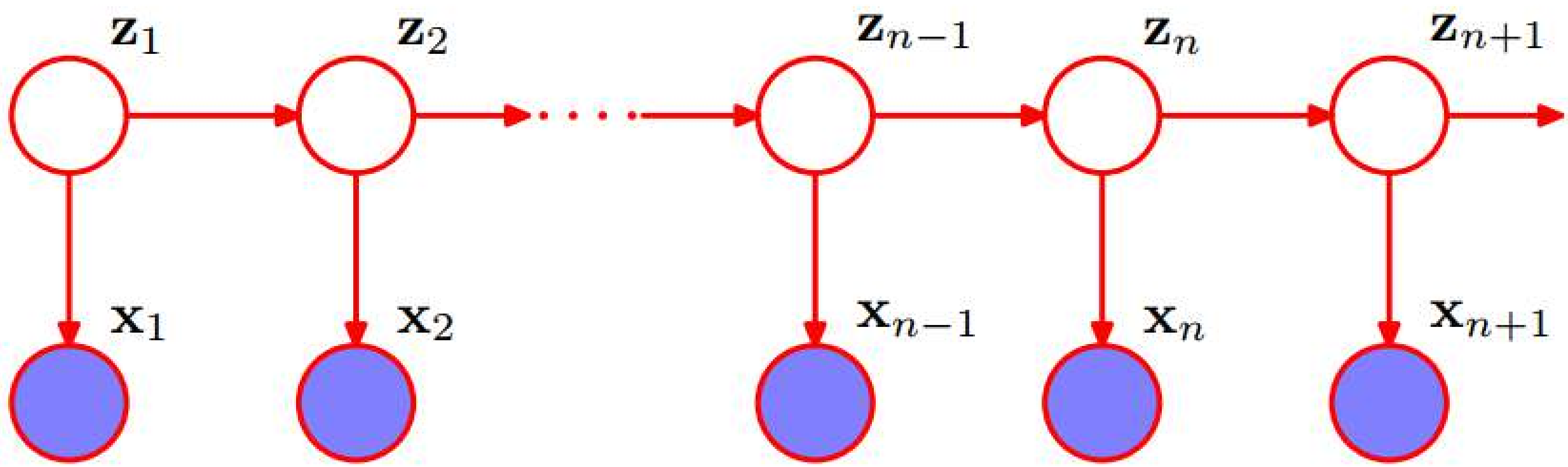
三个性质的等价性

$$
\begin{tikzcd}
G \arrow[r, Rightarrow] & L \arrow[r, Rightarrow] & P \arrow[ll, Rightarrow, bend left, "p(x) > 0"]
\end{tikzcd}
$$

- 根据全局马尔科夫性，能够得到局部马尔科夫性；
- 根据局部马尔科夫性，能够得到成对马尔科夫性；
- 根据成对马尔科夫性，能够得到全局马尔科夫性；

- 可以反向思考：满足这三个性质(或其一)的无向图，称为概率无向图模型。

复习：隐马尔科夫模型


$\mathbf{z}_1$
$\mathbf{z}_2$
$\mathbf{z}_{n-1}$
$\mathbf{z}_n$
$\mathbf{z}_{n+1}$
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_{n-1}$
$\mathbf{x}_n$
$\mathbf{x}_{n+1}$

HMM的确定

- HMM由初始概率分布π、状态转移概率分布A以及观测概率分布B确定。

$$\lambda = (A, B, \pi)$$

HMM的参数

- Q是所有可能的状态的集合
  - N是可能的状态数
- V是所有可能的观测的集合
  - M是可能的观测数

$$Q = \{q_1, q_2, \ldots q_N\}$$

$$V = \{v_1, v_2, \ldots v_M\}$$

HMM的参数

- I是长度为T的状态序列，O是对应的观测序列

$$I = \{i_1, i_2, \ldots i_T\} \qquad O = \{o_1, o_2, \ldots o_T\}$$

- A是状态转移概率矩阵

$$A = [a_{ij}]_{N\times N}$$

- 其中 $a_{ij} = P(i_{t+1} = q_j | i_t = q_i)$
- aij是在时刻t处于状态qi的条件下时刻t+1转移到状态qj的概率。

HMM的参数

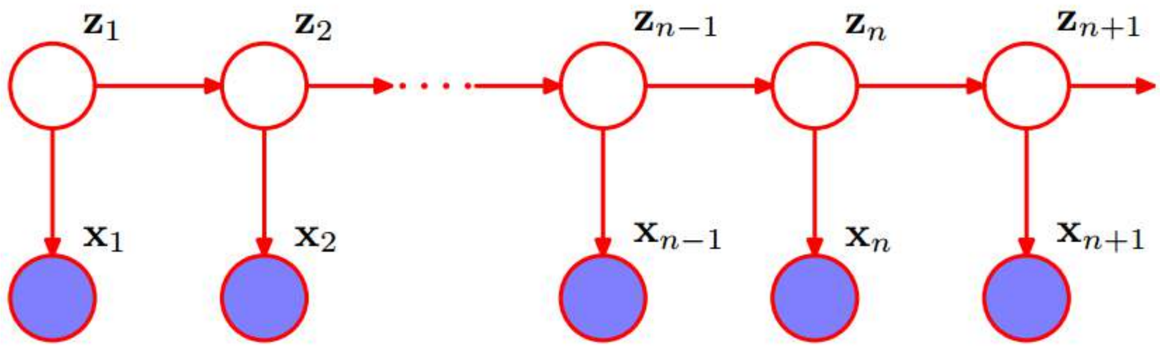
- B是观测概率矩阵 $B = \left[b_{ik}\right]_{N\times M}$
- 其中，$b_{ik} = P\left(o_t = v_k \middle| i_t = q_i\right)$
  - bik是在时刻t处于状态qi的条件下生成观测vk的概率。
- π是初始状态概率向量： $\pi = \left(\pi_i\right)$
- 其中，$\pi_i = P\left(i_1 = q_i\right)$
  - πi是时刻t=1处于状态qi的概率。

HMM的参数总结

- HMM由初始概率分布π、状态转移概率分布A以及观测概率分布B确定。π和A决定状态序列，B决定观测序列。因此，HMM可以用三元符号表示，称为HMM的三要素：

$$\lambda = (A, B, \pi)$$

HMM的两个基本性质



- 齐次假设：

$$P\left(i_t \middle| i_{t-1}, o_{t-1}, i_{t-2}, o_{t-2} \cdots i_1, o_1\right) = P\left(i_t \middle| i_{t-1}\right)$$

- 观测独立性假设：

$$P\left(o_t \middle| i_T, o_T, i_{T-1}, o_{T-1} \cdots i_1, o_1\right) = P\left(o_t \middle| i_t\right)$$

HMM的3个基本问题

- 概率计算问题
  - 给定模型 $\lambda=(A,B,\pi)$ 和观测序列$O=\{o_1,o_2,\ldots o_T\}$，计算模型λ下观测序列O出现的概率P(O| λ)
- 学习问题
  - 已知观测序列$O=\{o_1,o_2,\ldots o_T\}$，估计模型 $\lambda=(A,B,\pi)$的参数，使得在该模型下观测序列P(O| λ)最大
- 预测问题
  - 即解码问题：已知模型$\lambda=(A,B,\pi)$和观测序列 $O=\{o_1,o_2,\ldots o_T\}$ 求对给定观测序列条件概率P(I|O)最大的状态序列I

概率计算问题

- 直接算法
  - 暴力算法
- 前向算法
- 后向算法
  - 这二者是理解HMM的算法重点

直接计算法

- 按照概率公式，列举所有可能的长度为T的状态序列 $I=\{i_1,i_2,\ldots i_T\}$ ，求各个状态序列I与观测序列 $O=\{o_1,o_2,\ldots o_T\}$ 的联合概率P(O,I|λ)，然后对所有可能的状态序列求和，从而得到P(O|λ)

直接计算法

- 状态序列 $I=\{i_1,i_2,\ldots i_T\}$ 的概率是：

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

- 对固定的状态序列I，观测序列O的概率是：

$$P(O|I,\lambda)=b_{i_1o_1}b_{i_2o_2}\cdots b_{i_To_T}$$

直接计算法

- O和I同时出现的联合概率是：

$$P(O,I|\lambda)=P(O|I,\lambda)P(I|\lambda)$$

$$=\pi_{i_1}b_{i_1o_1}a_{i_1i_2}b_{i_2o_2}\cdots a_{i_{T-1}i_T}b_{i_To_T}$$

- 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|λ)

$$P(O|\lambda)=\sum_I P(O,I|\lambda)=\sum_I P(O|I,\lambda)P(I|\lambda)$$

$$=\sum_{i_1,i_2,\cdots i_T}\pi_{i_1}b_{i_1o_1}a_{i_1i_2}b_{i_2o_2}\cdots a_{i_{T-1}i_T}b_{i_To_T}$$

直接计算法

- 对于最终式

$$P(O|\lambda)=\sum_{I}P(O,I|\lambda)=\sum_{I}P(O|I,\lambda)P(I|\lambda)$$

$$=\sum_{i_1,i_2,\cdots i_T}\pi_{i_1}b_{i_1o_1}a_{i_1i_2}b_{i_2o_2}\cdots a_{i_{T-1}i_T}b_{i_To_T}$$

- 分析：加和符号中有2T个因子，I的遍历个数为N^T，因此，时间复杂度为$O(T\,N^T)$，过高。

前向算法

- 定义：给定λ，定义到时刻t部分观测序列为o1,o2…ot且状态为qi的概率为前向概率，记做：

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

- 可以递推的求得前向概率$\alpha_t(i)$及观测序列概率$P(O|\lambda)$

前向算法

- 初值：$\alpha_1(i) = \pi_i b_{io_1}$
- 递推：对于t=1,2…T-1

$$\alpha_{t+1}(i) = \left(\sum_{j=1}^{N} \alpha_t(j) a_{ji} \right) b_{io_{t+1}}$$

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

后向算法

- 定义：给定λ，定义到时刻t状态为qi的前提下，从t+1到T的部分观测序列为$o_{t+1}, o_{t+2} \ldots o_T$的概率为后向概率，记做：

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

- 可以递推的求得后向概率$\beta_t(i)$及观测序列概率$P(O|\lambda)$

后向算法

- 初值： $\beta_T(i)=1$
- 递推：对于t=T-1,T-2…,1

$$\beta_t(i)=\left(\sum_{j=1}^{N}a_{ij}b_{jo_{t+1}}\beta_{t+1}(j)\right)$$

- 最终： $P(O|\lambda)=\sum_{i=1}^{N}\pi_i b_{io_1}\beta_1(i)$

后向算法的说明

- 为了计算在时刻t状态为qi条件下时刻t+1之后的观测序列为$o_{t+1},o_{t+2}\ldots o_T$的后向概率$\beta_t(i)$，只需要考虑在时刻t+1所有可能的N个状态qj的转移概率(aij项)，以及在此状态下的观测o_{t+1}的观测概率(b_{jot+1})项，然后考虑状态qj之后的观测序列的后向概率$\beta_{t+1}(j)$

前向后向概率的关系

- 根据定义，证明下列等式

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

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

单个状态的概率

- 求给定模型λ和观测O，在时刻t处于状态qi的概率。
- 记：$\gamma_t(i)=P(i_t=q_i|O,\lambda)$

单个状态的概率

- 根据前向后向概率的定义，

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

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

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

γ的意义

- 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。
- 给定模型和观测序列，时刻t处于状态qi的概率为：

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

两个状态的联合概率

- 求给定模型λ和观测O，在时刻t处于状态qi并且时刻t+1处于状态qj的概率。

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

两个状态的联合概率

- 根据前向后向概率的定义，

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

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

期望

- 在观测O下状态i出现的期望：

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

- 在观测O下状态i转移到状态j的期望：

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

学习算法

- 若训练数据包括观测序列和状态序列，则HMM的学习非常简单，是监督性学习；
- 若训练数据只有观测序列，则HMM的学习需要使用EM算法，是非监督学习。

再次分析二项分布的参数估计

- 极大似然估计
- 简单的例子
  - 10次抛硬币的结果是：正正反正正正反反正正
- 假设p是每次抛硬币结果为正的概率。则：
- 得到这样的实验结果的概率是：

$$
\begin{aligned}
P &= pp(1-p)ppp(1-p)(1-p)pp \\
&= p^7(1-p)^3
\end{aligned}
$$

极大似然估计MLE

- 目标函数：$\max P = \max\limits_{0 \le p \le 1} p^7 (1-p)^3$
- 最优解是：p=0.7
  - 即：使用样本的均值可以作为全体的均值估计
- 一般形式：

$$L_{\bar{p}} = \prod_x p(x)^{\bar{p}(x)}$$

$p(x)$模型是估计的概率分布

$\bar{p}(x)$是实验结果的分布