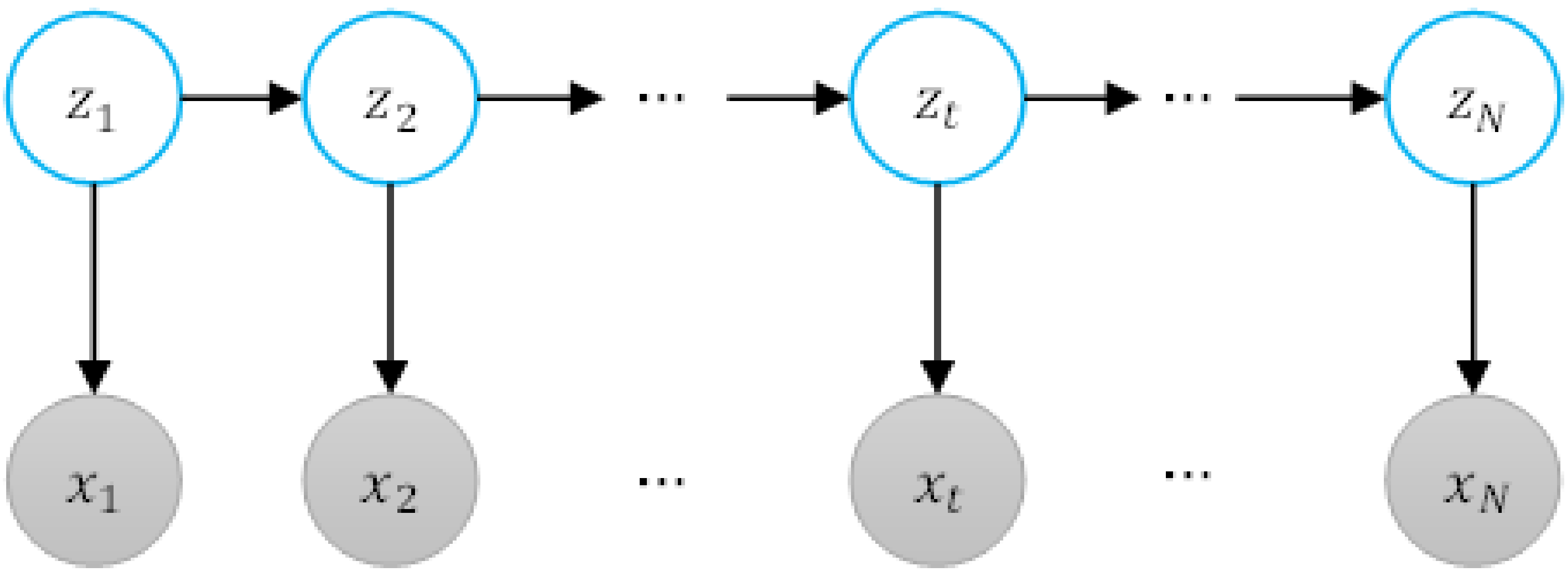
隐马尔可夫模型

Hidden Markov Model

- 隐马尔可夫模型(HMM)可用于标注问题，在语音识别、NLP、生物信息、模式识别等领域被证明是有效的算法
- HMM是关于时序的概率模型，描述由一个隐藏的马尔可夫链生成不可观测的状态随机序列，再由各个状态生成观测随机序列的过程。
- 隐马尔可夫模型随机生成的状态随机序列，成为状态序列
- 每个状态生成一个观测，由此产生的观测随机序列，称为观测序列，序列的每个位置可看做是一个时刻。



HMM由初始概率分布π、状态转移概率分布A及观测概率分布B确定

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

- 设Q是所有可能的状态集合

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

- V是所有可能的观测的集合

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

- I是长度为T的状态序列，O是对应的观测序列

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

- A是状态转移概率矩阵

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

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

- B是观测概率矩阵

$$B = [b_{ik}]_{N\times N}$$

其中，$b_{ik} = P(o_t = v_k | i_t = q_i)$

  - b_{ik}是在时刻t处于状态q_i的条件下生成观测v_k的概率。

- π是初始状态概率向量

$$\pi = (\pi_i)$$

其中，$\pi_i = P(i_1 = q_i)$

  - π_i是在时刻$t = 1$处于状态q_i的概率。

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

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

- 两个基本假设
  - 齐次假设, HMM在任意时刻t的状态只依赖于其前一时刻的状态，于其他时刻的状态及观测无关，也与时刻t无关

    $$P(i_t|i_{t-1}, o_{t-1}, i_{t-2}, o_{t-2}, \ldots, i_1, o_1) = P(i_t|i_{t-1}),\ t = 1, 2, \ldots, T$$
  - 观测独立性假设，即任意时刻的而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

$$P(o_t|i_T, o_T, \ldots, i_{t+1}, o_{t+1}, i_t, i_{t-1}, o_{t-1}, \ldots, i_1, o_1) = P(o_t|i_t)$$

HMM举例

- 假设有3个盒子，编号为1、2、3，每个盒子都装有红白两种颜色的小球，数目如下：
  - 盒子号: 1,2,3
  - 红球数: 5,4,7
  - 白球数: 5,6,3
- 按照下面的方法抽取小球，得到球颜色的观测序列:
  - 按照$\pi = (0.2, 0.4, 0.4)$的概率选择1个盒子，从盒子随机抽数1个球，记录颜色后放回盒子；
  - 按照某条件概率选择新的盒子，重复上述过程
  - 最终得到观测序列: "红红白白红"

- 状态集合：$Q = \{$盒子1，盒子2，盒子3$\}$
- 观测集合：$V = \{$红，白$\}$
- 状态序列和观测序列的长度$T = 5$
- 初始概率分布π、状态转移概率分布A、观测概率分布B

$$\pi = (0.2, 0.4, 0.4)^T$$

$$A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$$

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

观测序列的生成过程

- 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$O = (o_1, o_2, \ldots, o_T)$

(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$i_{t+1}, i_{t+1} = 1, 2, \ldots, N$

(5) 令$t = t + 1$; 如果$t < T$, 转步(3); 否则，终止。

HMM的三个基本问题

(1)概率计算问题(评估问题): 前向-后向算法--动态规划

- 给定模型$\lambda=(A,B,\pi)$和观测序列$O=(o_1,o_2,...,o_T)$, 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: Baum-Welch算法(状态未知)--EM算法

- 已知观测序列$O=(o_1,o_2,...,o_T)$, 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.

(3)预测问题(解码问题): Viterbi算法--动态规划

- 已知模型$\lambda=(A,B,\pi)$和观测序列$O=(o_1,o_2,...,o_T)$, 求对给定观测序列条件概率$P(I|O)$最大的状态序列$I=(i_1,i_2,...,i_T)$

- 概率计算问题
  - 直接算法
  - 前向算法
  - 后向算法
- 直接算法
  按照概率公式，列举所有可能的长度为T的状态序列$I=\{i_1,i_2,...,i_T\}$，求各个状态序列I与观测序列$O=\{o_1,o_2,...,o_T\}$的联合概率$P(O,I|\lambda)$，然后对所有可能的状态序列求和，从而得到$P(O|\lambda)$
- 状态序列$I=\{i_1,i_2,...,i_T\}$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:

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

- O和I同时出现的联合概率是:

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

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

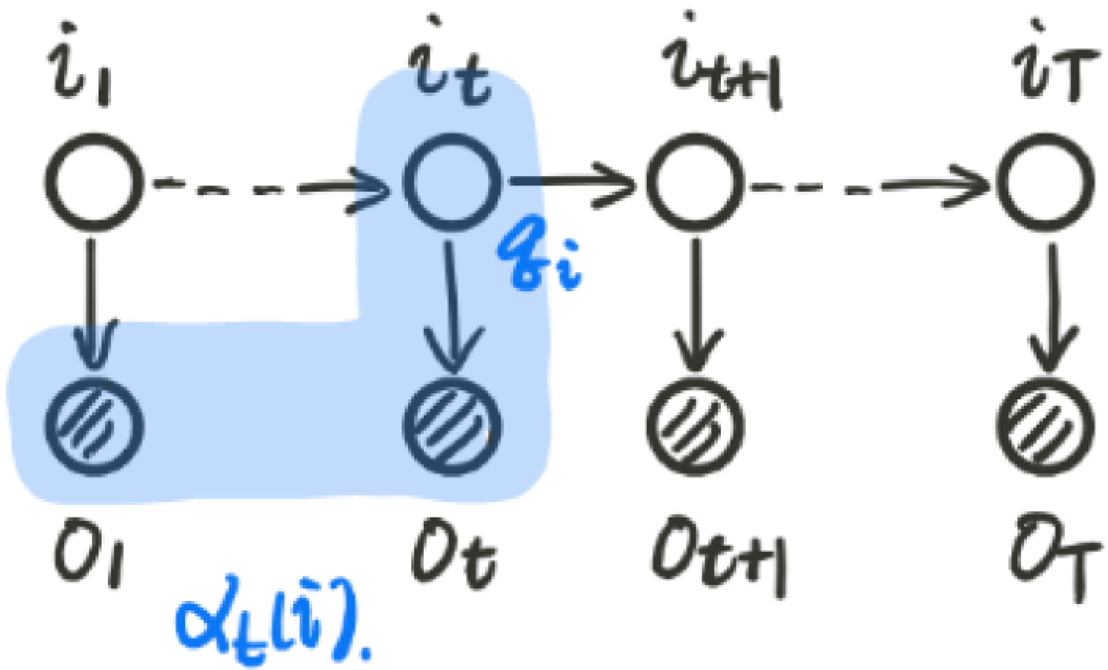
$$
\begin{aligned}
P(O|\lambda) &= \sum_I P(O, I|\lambda) \\
&= \sum_{i_1, i_2, \ldots, i_T} \pi_{i_1} b_{i_1 o_1} a_{i_1 i_2} b_{i_2 o_2} \cdots a_{i_{T-1} i_T} b_{i_T o_T}
\end{aligned}
$$

- 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$O(TN^T)$, 复杂度过高。

前向算法

- 前向概率: 给定λ, 定义到时刻t部分观测序列为$o_1, o_2, \ldots, o_t$且状态为q_i的概率称为前向概率

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



- 最终: $P(O|\lambda) = \sum_{i=1}^{N} P(O, i_T = q_i|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$
- 初值: $\alpha_1(i) = \pi_i b_{io_1}$
- 递推： $\alpha_{t+1}(i) = (\sum_{j=1}^{N} \alpha_t(j) a_{ji}) b_{io_{t+1}}$
- 考察盒子球模型，计算观测向量$O =$“红白红”的出现概率

$$\pi = (0.2, 0.4, 0.4)^T$$

$$A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$$

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

计算初值: $\alpha_1(1) = \pi_1 b_{1o_1} = 0.2 \times 0.5 = 0.1$

$\alpha_1(2) = \pi_2 b_{2o_1} = 0.4 \times 0.4 = 0.16$

$\alpha_1(3) = \pi_3 b_{3o_1} = 0.4 \times 0.7 = 0.28$

递推:

$$\begin{aligned}\alpha_2(1) &= (\textstyle\sum_{j=1}^{N} \alpha_1(j) a_{j1}) b_{1o_{t+1}} \\ &= (0.1 \times 0.5 + 0.16 \times 0.3 + 0.28 \times 0.2) \times 0.5 \\ &= 0.077\end{aligned}$$

$\alpha_2(2) = 0.1104, \alpha_2(3) = 0.0606,$
$\alpha_3(1) = 0.04187, \alpha_3(2) = 0.03551, \alpha_3(3) = 0.05284$

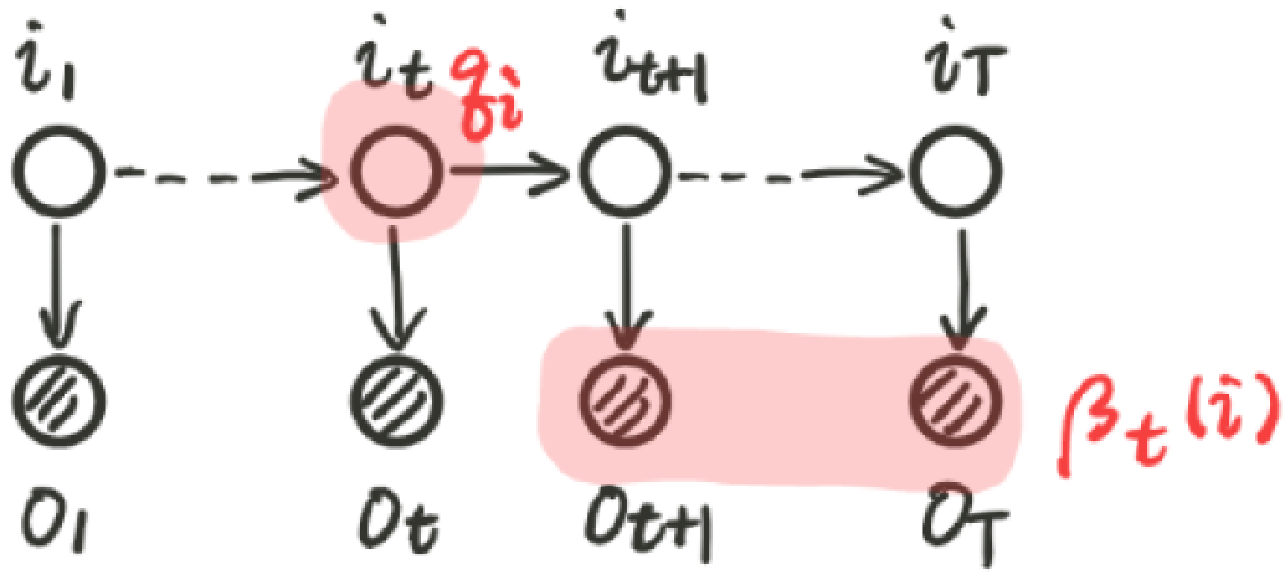
最终:

$$\begin{aligned}P(O|\lambda) &= \textstyle\sum_{i=1}^{3} \alpha_3(i) \\ &= 0.04187 + 0.03551 + 0.05284 \\ &= 0.13022\end{aligned}$$

后向算法

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

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



- 最终:

$$\begin{aligned} P(O|\lambda) &= \sum_{i=1}^{N} P(O, i_1 = q_i|\lambda) \\ &= \sum_{i=1}^{N} P(O|i_1 = q_i, \lambda)\pi_i \\ &= \sum_{i=1}^{N} \pi_i P(o_1|i_1 = q_i, \lambda)P(o_2, \ldots, o_T|i_1 = q_i, \lambda) \\ &= \sum_{i=1}^{N} \pi_i b_{io_1} \beta_1(i) \end{aligned}$$

- 初值: $\beta_T(i) = 1$
- 递推:

$$\begin{aligned} \beta_t(i) &= P(o_{t+1}, o_{t+2}, \ldots, o_T|i_t = q_i, \lambda) \\ &= \sum_{j=1}^{N} P(o_{t+1}, o_{t+2}, \ldots, o_T, i_{t+1} = q_j|i_t = q_i, \lambda) \\ &= \sum_{j=1}^{N} P(o_{t+1}, o_{t+2}, \ldots, o_T|i_t = q_i, i_{t+1} = q_j, \lambda)P(i_{t+1} = q_j|i_t = q_i) \\ &= \sum_{j=1}^{N} P(o_{t+1}, o_{t+2}, \ldots, o_T|i_{t+1} = q_j, \lambda)\alpha_{ij} \\ &= \sum_{j=1}^{N} P(o_{t+1}|i_{t+1} = q_j, \lambda)P(o_{t+2}, \ldots, o_T|i_{t+1} = q_j, \lambda)\alpha_{ij} \\ &= \sum_{j=1}^{N} b_{jo_{t+1}} \beta_{t+1}(j)\alpha_{ij} \end{aligned}$$

前向后向关系

$$\begin{aligned}\alpha_t(i)\beta_t(i) &= P(o_1, o_2, \ldots, o_t, i_t = q_i|\lambda)P(o_{t+1}, o_{t+2}, \ldots, o_T|i_t = q_i, \lambda) \\ &= P(o_1, o_2, \ldots, o_t|i_t = q_i, \lambda)P(i_t = q_i|\lambda)P(o_{t+1}, o_{t+2}, \ldots, o_T|i_t = q_i, \lambda) \\ &= P(O|i_t = q_i, \lambda)P(i_t = q_i|\lambda) \\ &= P(O, i_t = q_i|\lambda)\end{aligned}$$

一些概率与期望值的计算

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

- 给定模型λ和观测O，在时刻t处于状态q_i的概率.

$$r_t(i) = P(i_t = q_i|O, \lambda) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}$$

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

$$\begin{aligned}
\xi_t(i,j) &= P(i_t = q_i, i_{t+1} = q_j | O, \lambda) \\
&= \frac{P(i_t = q_i, i_{t+1} = q_j, O | \lambda)}{P(O|\lambda)} \\
&= \frac{P(i_t = q_i, i_{t+1} = q_j, O | \lambda)}{\sum_{i=1}^{N} \sum_{j=1}^{N} P(i_t = q_i, i_{t+1} = q_j, O | \lambda)} \\
&= \frac{\alpha_t(i) a_{ij} b_{j o_{t+1}} \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_{j o_{t+1}} \beta_{t+1}(j)}
\end{aligned}$$

期望

- 在观测O下状态i出现的期望值

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

- 在观测O下由状态i转移的期望值

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

- 在观测O下由状态i转移到状态j的期望值

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

学习问题

- 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

假设已给定训练数据包含S个长度相同的观测序列和对应的状态序列 $\{(O_1, I_1), (O_2, I_2), \ldots, (O_s, I_s)\}$, 那么，可以直接利用Bernoulli大数定理的结论"频率的极限是概率"，给出HMM的参数估计。

- 初始概率: S个样本中初始状态为q_i的频率

$$\hat{\pi}_i = \frac{|q_i|}{\sum_i |q_i|}$$

- 转移概率: 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

$$\hat{a}_{ij} = \frac{|A_{ij}|}{\sum_i |A_{ij}|}$$

- 观测概率: 设样本状态为j并观测为k的频数是B_ij

$$\hat{B}_{ij} = \frac{|B_{ij}|}{\sum_i |B_{ij}|}$$

Baum-Welch算法

- 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。
- 所有观测数据写成$O=(o_1,o_2,\dots,o_T)$, 所有隐藏数据写成$I=(i_1,i_2,\dots,i_T)$
- 完全数据是$(O,I)=(o_1,o_2,\dots,o_T,i_1,i_2,\dots,i_T)$
- 完全数据的对数似然函数是$\ln P(O,I|\lambda)$
- 假设$\bar{\lambda}$是HMM参数的当前估计值，λ为待求的参数。

$$
\begin{aligned}
Q(\lambda,\bar{\lambda}) \quad &= \sum_I \ln P(O,I|\lambda)P(I|O,\bar{\lambda}) \\
&= \sum_I \ln P(O,I|\lambda)\tfrac{P(I,O|\bar{\lambda})}{P(O,\bar{\lambda})} \\
&\propto \sum_I \ln P(O,I|\lambda)P(I,O|\bar{\lambda})
\end{aligned}
$$

根据

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

函数可写成:

$$
\begin{aligned}
Q(\lambda,\bar{\lambda}) &= \sum_I \ln P(O,I|\lambda)P(I,O|\bar{\lambda}) \\
&= \sum_I \ln \pi_{i_1} P(O,I|\bar{\lambda}) \\
&\quad + \sum_I (\sum_{t=1}^{T-1} \ln a_{i_t,i_{t+1}}) P(O,I|\bar{\lambda}) \\
&\quad + \sum_I (\sum_{t=1}^{T} \ln b_{i_t,o_t}) P(O,I|\bar{\lambda})
\end{aligned}
$$

极大化Q, 求得参数A,B,π

由于该三个参数分别位于三个项中，可分别极大化

$$\sum_I \ln \pi_{i_1} P(O, I|\bar{\lambda}) = \sum_{i=1}^N \ln \pi_{i_1} P(O, i_1 = i|\bar{\lambda})$$

注意到π_{i_1}满足加和为1，利用拉格朗日乘子法，得到:

$$\sum_{i=1}^N \ln \pi_{i_1} P(O, i_1 = i|\bar{\lambda}) + \gamma(\sum_{i=1}^N \pi_i - 1)$$

上式相对于π_i求偏导，得到: $P(O, i_1 = i|\bar{\lambda}) + \gamma\pi_i = 0$

对i求和, 得到: $\gamma = -P(O|\bar{\lambda})$

从而得到初始状态概率: $\pi_i = \frac{P(O,i_1=i|\bar{\lambda})}{P(O|\bar{\lambda})} = \gamma_1(i)$

第二项可写成:

$$
\begin{aligned}
& \sum_I (\sum_{t=1}^{T-1} \ln a_{i_t, i_{t+1}}) P(O, I|\bar{\lambda}) \\
= \quad & \sum_{i=1}^N \sum_{j=1}^N \sum_{t=1}^{T-1} \ln a_{i,j} P(O, i_t = i, i_{t+1} = j|\bar{\lambda})
\end{aligned}
$$

仍然使用拉格朗日乘子法，得到:

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

同理，得到:

$$
b_{ik} = \frac{\sum_{t=1}^{T} P(O, i_t = i|\bar{\lambda}) I(o_t = v_k)}{\sum_{t=1}^{T} P(O, i_t = i|\bar{\lambda})} = \frac{\sum_{t=1, o_t = v_k}^{T-1} \gamma_t(i)}{\sum_{t=1}^{T-1} \gamma_t(i)}
$$

Baum-Welch算法

输入: 观测数据$O = (o_1, o_2, \ldots, o_T)$
输出: 隐马尔可夫模型参数
(1) 初始化
对$n = 0$, 选取$a_{ij}^{(0)}, b_{jk}^{(0)}, \pi_i^{(0)}$, 得到模型$\lambda^{(0)} = (A^{(0)}, B^{(0)}, \pi^{(0)})$
(2) 递推. 对$n = 1, 2, \cdots,$

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

$$b_{ik}^{n+1} = \frac{\sum_{t=1, o_t=v_k}^{T-1} \gamma_t(i)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\pi_i = \gamma_1(i)$$

右端各值按观测$O = (o_1, o_2, ..., o_T)$和模型$\lambda^{(n)} = (A^{(n)}, B^{(n)}, \pi^{(n)})$计算.
(3)终止. 得到模型参数$\lambda^{(n+1)} = (A^{(n+1)}, B^{(n+1)}, \pi^{(n+1)})$

预测问题

- 近似算法
- Viterbi算法

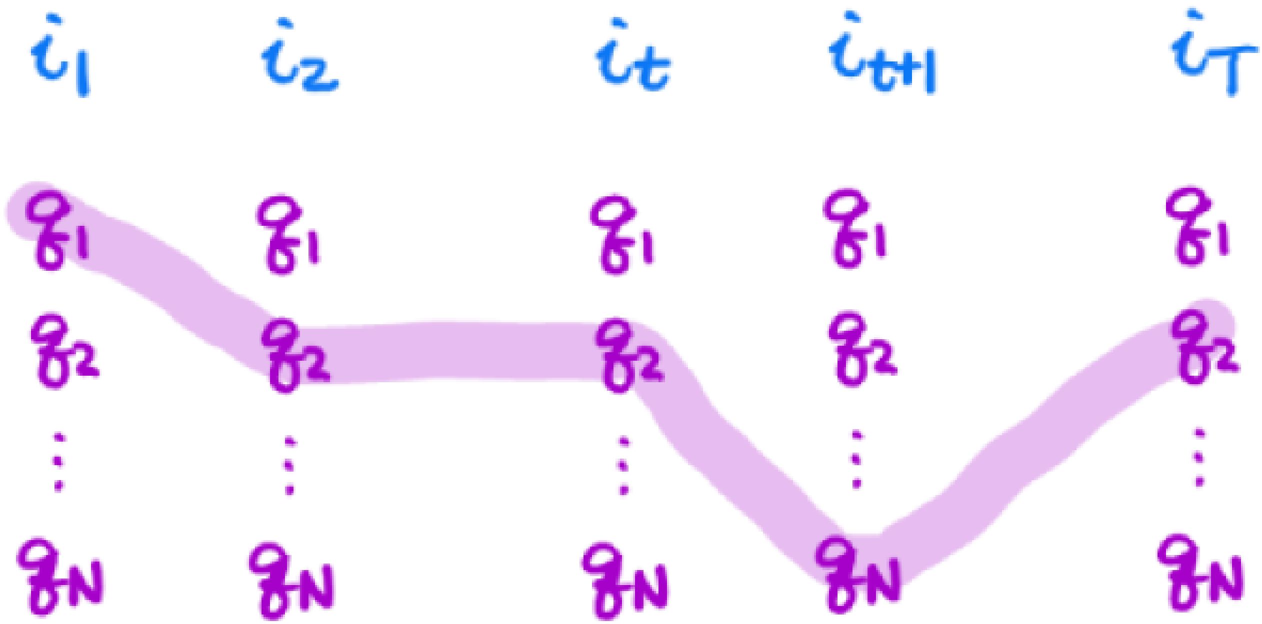
预测的近似算法

- 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$I^* = \{i_1^*, i_2^*, ..., i_T^*\}$，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q_i的概率为:

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

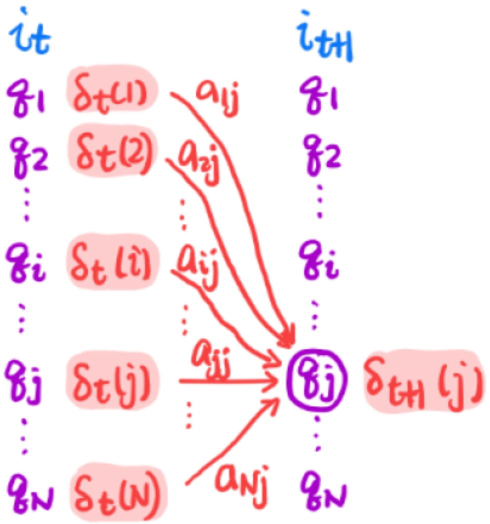
Viterbi算法

- Viterbi算法实际是用动态规划解HMM预测问题，用DP求概率最大的路径(最优路径)，这是一条路径对应一个状态序列。
- 定义变量$\delta_t(i)$: 在时刻t状态为i的所有路径中，概率的最大值



- 定义:

$$\delta_t(i) = \max_{i_1, i_2, \ldots, i_{t-1}} P(i_t = i, i_{t-1}, \ldots, i_1, o_t, \ldots, o_1 | \lambda)$$



$$\begin{aligned}\delta_{t+1}(i) &= \max_{i_1, i_2, \ldots, i_t} P(i_{t+1} = i, i_t, \ldots, i_1, o_{t+1}, \ldots, o_1 | \lambda) \\ &= \max_{1 \le j \le N} (\delta_{t+1}(j) a_{ji}) b_{io_{t+1}}\end{aligned}$$

终止: $P^* = \max_{1 \le j \le N} \delta_T(i)$

考察盒子球模型，观测向量O =“红白红”, 试求最优状态序列

$$\pi = (0.2, 0.4, 0.4)^T$$

$$A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$$

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

初始化: 在$t=1$时，对于每一个状态i，求状态为i观测到o_1为红的概率，记此概率为$\delta_1(t)$

$$\delta_1(i)=\pi_i b_{io_1}=\pi_i b_{i红}$$

求得$\delta_1(1)=0.1, \delta_1(2)=0.16, \delta_1(3)=0.28$
在$t=2$时，对每个状态i，求在$t=1$时状态为j观测为红，并且在$t=2$时状态为i观测为白的路径的最大概率，记概率为$\delta_2(t)$，则:

$$\delta_{t+1}(i)=\max_{1\leq j\leq 3}(\delta_1(j)a_{ji})b_{io_2}=\max_{1\leq j\leq 3}(\delta_1(j)a_{ji})b_{i白}$$

求得

$$\begin{aligned}\delta_2(1) &= \max_{1\leq j\leq 3}(\delta_1(j)a_{j1})b_{i白}\\ &= \max\{0.10\times 0.5, 0.16\times 0.3, 0.28\times 0.2\}\times 0.5\\ &= 0.028\end{aligned}$$

同理: $\delta_2(2)=0.0504, \delta_2(3)=0.042$

同理，求得$\delta_3(1) = 0.00756, \delta_3(2) = 0.01008, \delta_3(3) = 0.0147$

从而，最大是$\delta_3(3) = 0.0147$，根据每一步的最大，得到序列是 $(3, 3, 3)$