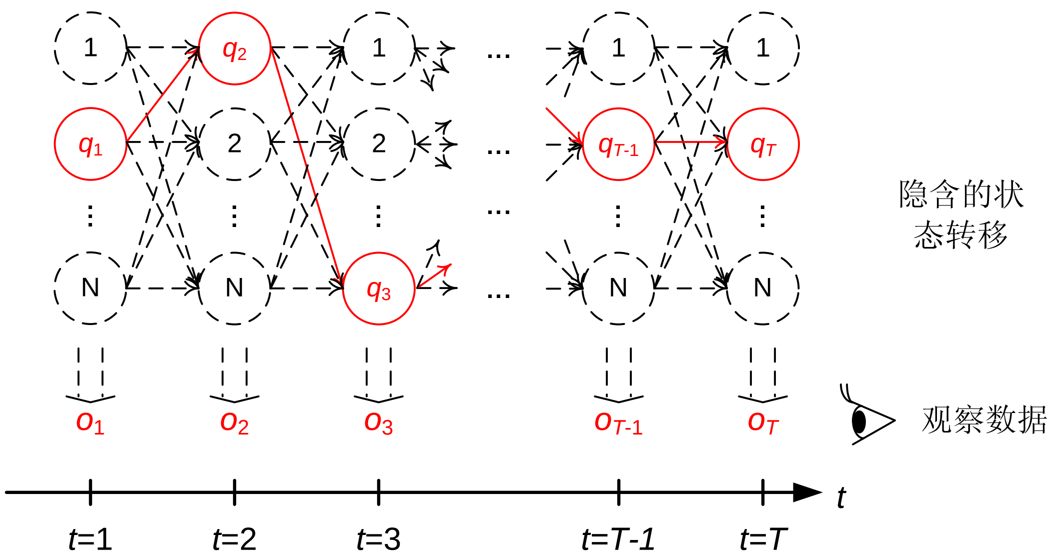
EM 算法与隐马尔可夫模型(HMM)参数估计

备忘录，里面跳过部分的推导可以在参考文献找到（未完待续）

uingrd@gmail.com

1. HMM 模型



状态序列记作：

$$Q := q_1, q_2, \cdots, q_T \tag{1}$$

其中每个状态的取值范围为：

$$q_t \in \{1, 2, \cdots, N\} \tag{2}$$

表示有 N 个状态。

观察序列记作：

$$O := o_1, o_2, \cdots, o_N \tag{3}$$

状态转移概率记作：

$$a_{i,j} := p(q_{t+1} = j \mid q_t = i) \tag{4}$$

初始状态概率记作：

$$\pi_i = p(q_1 = i) \tag{5}$$

状态/输出关系（概率）记作：

$$b_i(o) := p(o_t = o \mid q_t = i) \tag{6}$$

（上面表达式虽然有 t，但 $b_i(o)$ 和 t 无关，或者说不随 t 改变）

关于 HMM 模型有假设：

1) 当前状态仅仅取决于上一状态

$$p(q_{t+1} \mid q_t, q_{t-1}, \cdots, q_1, o_t, o_{t-1}, \cdots, o_1) = p(q_{t+1} \mid q_t) \tag{7}$$

2) 当前输出仅仅取决于当前状态

$$p(o_t \mid q_t, q_{t-1}, \cdots, q_1, o_{t-1}, o_{t-1}, \cdots, o_1) = p(o_t \mid q_t) \tag{8}$$

关于 HMM 的参数记作：

$$\begin{cases} \mathbf{A} := \{a_{i,j}\}\big|_{i,j=1,2,\cdots,M} \\ \mathbf{B} := \{b_i(o)\}\big|_{i=1,2,\cdots,M} \\ \boldsymbol{\pi} := \{\pi_i\}\big|_{i=1,2,\cdots,M} \end{cases} \tag{9}$$

上面 3 个参数集合统一记作：

$$\Theta := \{\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}\} \tag{10}$$

注意，下面的公式为了突出概率模型参数 Θ，把前面的概率符号 $p(\bullet)$ 都改写成条件概率的模式：$p(\bullet \mid \Theta)$。

2. 概率计算的递推算法

考虑计算概率：

$$p(O \mid \Theta) := p(o_1, o_2, \cdots, o_T \mid \Theta) \tag{11}$$

注意：计算 $p(O \mid \Theta)$ 的一个意义是用于得到 Θ 的估计，即 $\Theta^* = \arg\max_{\Theta} P(O \mid \Theta)$。

直接计算 $p(O \mid \Theta)$ 公式为：

$$p(O|\Theta) := p(o_1, o_2, \cdots, o_T | \Theta) = \sum_{q_1=1}^{N} \sum_{q_2=1}^{N} \cdots \sum_{q_T=1}^{N} \left[\left(\prod_{t=1}^{T} p(q_t|\Theta) b_{q_t}(o_t) \right) \left(\prod_{t=1}^{T-1} a_{q_t q_{t+1}} \right) \right] \tag{12}$$

（证明参照附录）上面的表达式表示计算 $p(O|\Theta)$ 需要大量的计算，直接计算很困难，但可以通过下面的递推算法大大减少计算量。

定义：

$$\alpha_i(t) := p(o_1, o_2, \cdots, o_t, q_t = i | \Theta) \tag{13}$$

于是有下面的前向递归：

$$\begin{cases} \alpha_i(1) = \pi_i b_i(o_1) \\ \alpha_j(t+1) = \left[\sum_{i=1}^{N} \alpha_i(t) a_{i,j} \right] b_j(o_{t+1}) \\ p(O|\Theta) = \sum_{i=1}^{N} \alpha_i(T) \end{cases} \tag{14}$$

通过式(14)中第二个等式的递推运算得到 $p(O|\Theta)$。关于(14)的证明参照附录。

另一个计算(11)的递推方法是后向递推法，定义：

$$\beta_i(t) := p(o_{t+1}, o_{t+2}, \cdots, o_T | q_t = i, \Theta) \tag{15}$$

递推公式为：

$$\begin{cases} \beta_i(T) = 1 \\ \beta_i(t) = \sum_{j=1}^{N} a_{i,j} b_j(o_{t+1}) \beta_j(t+1) \\ p(O|\Theta) = \sum_{i=1}^{N} \beta_i(1) \pi_i b_i(o_1) \end{cases} \tag{16}$$

关于(16)的证明参照附录。

3. Viterbi 算法

这一节讨论已知 O 条件下对 Q 的估计，即：

$$\max_{q_1, q_2, \cdots, q_T} p(Q|O, \Theta) \tag{17}$$

先定义：

$$\delta_t(i)=\max_{q_1,q_2,\cdots,q_{t-1}} p\left(o_1,o_2,\cdots,o_t,q_1,q_2,\cdots,q_{t-1},q_t=i\,|\,\Theta\right) \tag{18}$$

表示根据t时刻以及之前的观测值$\{o_1,o_2,\cdots,o_t\}$，并假定t时刻状态$q_t=i$时，对之前的状态序列$\{q_1,q_2,\cdots,q_{t-1}\}$的最优估计。这样可以得到下面的递推算法：

1) 递推初始

$$\begin{cases}\delta_1(i)=\pi_i b_i(o_1)\\ \phi_1(i)=0\end{cases},\quad 1\le i\le N \tag{19}$$

2) 递推

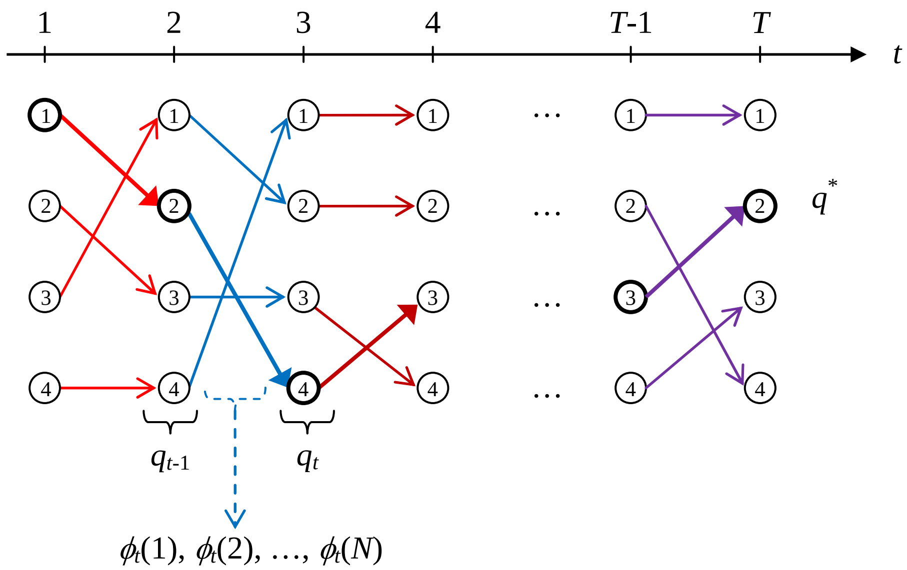
$$\begin{cases}\delta_t(j)=\max\limits_{1\le i\le N}\left(\delta_{t-1}(i)a_{i,j}\right)b_j(o_t)\\ \phi_t(j)=\arg\max\limits_{1\le i\le N}\left(\delta_{t-1}(i)a_{i,j}\right)\end{cases},\quad 2\le t\le T\text{，}\ 1\le j\le N \tag{20}$$

3) 递推终止

$$\begin{cases}p^*=\max\limits_{1\le i\le N}\left(\delta_T(i)\right)\\ q^*=\arg\max\limits_{1\le i\le N}\left(\delta_T(i)\right)\end{cases} \tag{21}$$

步骤的证明参考附录。

上面步骤中$\phi_t(j)$的值代表递推进行到t时刻时，连接$q_t=j$状态的最优q_{t-1}状态编号。(21)中的q^*是最后一刻决定的路线终点状态编号。如下图：



最后注意到：

$$
\begin{aligned}
p^* &= \max_{1\le i\le N} \max_{q_1,q_2,\cdots,q_{T-1}} p\left(o_1,o_2,\cdots,o_T,q_1,q_2,\cdots,q_{T-1},q_T=i\,|\,\Theta\right) \\
&= \left(\max_{1\le i\le N} \max_{q_1,q_2,\cdots,q_{T-1}} p\left(q_1,q_2,\cdots,q_{T-1},q_T=i\,|\,o_1,o_2,\cdots,o_T,\Theta\right)\right) p\left(o_1,o_2,\cdots,o_T\,|\,\Theta\right) \\
&= \left(\max_{q_1,q_2,\cdots,q_T} p\left(q_1,q_2,\cdots,q_{T-1},q_T\,|\,o_1,o_2,\cdots,o_T,\Theta\right)\right) p\left(o_1,o_2,\cdots,o_T\,|\,\Theta\right) \\
&= \left(\max_{q_1,q_2,\cdots,q_T} p\left(Q\,|\,O,\Theta\right)\right) p\left(O\,|\,\Theta\right)
\end{aligned}
\tag{22}
$$

这意味着 Viterbi 算法实际上实现了(17)，$\max_{q_1,q_2,\cdots,q_T} p\left(Q\,|\,O,\Theta\right)$，即：在已知观测值$O$的条件下估计状态序列$Q$。

算法总结

这里考虑到实际计算中的精度问题（概率在计算中接近 0，在双精度表示中被舍入成 0），因此下面的算法用 log 计算概率。

1) 递推初始

$$
\begin{cases}
\tilde{\delta}_1(i) = \log \pi_i + \log b_i(o_1) \\
\phi_1(i) = 0
\end{cases}
,\quad 1 \le i \le N
\tag{23}
$$

2) 递推

$$
\begin{cases}
\tilde{\delta}_t(j) = \max_{1\le i\le N}\left(\log \delta_{t-1}(i) + \log a_{i,j}\right) + \log b_j(o_t) \\
\phi_t(j) = \arg\max_{1\le i\le N}\left(\log \delta_{t-1}(i) + \log a_{i,j}\right)
\end{cases}
,\quad 2 \le t \le T,\quad 1 \le j \le N
\tag{24}
$$

3) 递推终止

$$
\begin{cases}
p^* = \max_{1\le i\le N}\left(\tilde{\delta}_T(i)\right) \\
q^* = \arg\max_{1\le i\le N}\left(\tilde{\delta}_T(i)\right)
\end{cases}
\tag{25}
$$

注意：上面的$\tilde{\delta}_t(i) = \log \delta_t(i)$

4. EM 算法估计 HMM 参数

考虑估计 HMM 的参数Θ。引用 EM 的读书笔记里的下界函数：

$$
\begin{aligned}
C\left(\Theta,\Theta^{(k)}\right) &:= \Sigma_{\mathbf{y}\in Y}\left\{p\left(\mathbf{y}\,|\,\mathbf{x},\Theta^{(k)}\right)\log\left[p\left(\mathbf{y}\,|\,\Theta\right)p\left(\mathbf{x}\,|\,\mathbf{y},\Theta\right)\right]\right\} \\
&= \Sigma_{\mathbf{y}\in Y}\left\{p\left(\mathbf{y}\,|\,\mathbf{x},\Theta^{(k)}\right)\log\left[p\left(\mathbf{x},\mathbf{y}\,|\,\Theta\right)\right]\right\}
\end{aligned}
\tag{26}
$$

应用到 HMM 里面，O是可见的观测数据，Q是不可见的，于是递推公式使用下界函数变为：

$$C\left(\Theta,\Theta^{(k)}\right):=\Sigma_Q\left\{p\left(Q\mid O,\Theta^{(k)}\right)\log\left[p\left(O,Q\mid\Theta\right)\right]\right\} \tag{27}$$

由于：

$$\begin{aligned}\arg\max_{\Theta} C\left(\Theta,\Theta^{(k)}\right)&=\arg\max_{\Theta}\Sigma_Q\left\{p\left(Q\mid O,\Theta^{(k)}\right)\log\left[p\left(O,Q\mid\Theta\right)\right]\right\}\\&=\arg\max_{\Theta} p\left(O\mid\Theta^{(k)}\right)\Sigma_Q\left\{p\left(Q\mid O,\Theta^{(k)}\right)\log\left[p\left(O,Q\mid\Theta\right)\right]\right\}\\&=\arg\max_{\Theta}\Sigma_Q\left\{p\left(Q,O,\Theta^{(k)}\right)\log\left[p\left(O,Q\mid\Theta\right)\right]\right\}\\&=\arg\max_{\Theta} D\left(\Theta,\Theta^{(k)}\right)\end{aligned} \tag{28}$$

我们可以等价地考虑对：

$$D\left(\Theta,\Theta^{(k)}\right):=\Sigma_Q\left\{p\left(Q,O\mid\Theta^{(k)}\right)\log\left[p\left(O,Q\mid\Theta\right)\right]\right\} \tag{29}$$

不断优化来求得Θ（可能收敛到局部极大点），即，使用下面的递推式：

$$\Theta^{(k+1)}:=\arg\max_{\Theta} D\left(\Theta,\Theta^{(k)}\right) \tag{30}$$

求$\Theta^{(k)}$的收敛点。

下面把 HMM 的 $p\left(O,Q\mid\Theta\right)$写出来，即：

$$p\left(O,Q\mid\Theta\right)=\pi_{q_0}\prod_{t=1}^{T}a_{q_{t-1}q_t}b_{q_t}\left(o_t\right) \tag{31}$$

注意：上面的表达式把状态计数从$t=0$开始(q_0)，而实际上是从$t=1$，这一差别

是为了后面的推导简单。考虑从$t=0$的状态q_0开始运行模型，这相当于少观察一个o_0，直接根据$\left\{o_1,o_2,\cdots,o_T\right\}$估计 HMM 的参数。

把(31)带入到$D\left(\Theta,\Theta^{(k)}\right)$得到：

$$D\left(\Theta,\Theta^{(k)}\right)$$
$$:=\Sigma_Q\left\{p\left(Q,O\mid\Theta^{(k)}\right)\log\left[p\left(O,Q\mid\Theta\right)\right]\right\}$$
$$=\Sigma_Q\left\{p\left(Q,O\mid\Theta^{(k)}\right)\log\left[\pi_{q_0}\prod_{t=1}^{T}a_{q_{t-1}q_t}b_{q_t}\left(o_t\right)\right]\right\}$$
$$=\sum_Q p(Q,O\mid\Theta^{(k)})\log\pi_{q_0}+\sum_Q\left(p(Q,O\mid\Theta^{(k)})\sum_{t=1}^{T}\log a_{q_{t-1}q_t}\right)+\sum_Q\left(p(Q,O\mid\Theta^{(k)})\sum_{t=1}^{T}\log b_{q_t}\left(o_t\right)\right) \tag{32}$$

上面的表达式分成 3 项，第一项只有 π_q，第二项只有 $a_{i,j}$，第三项只有 $b_q\left(o\right)$，因此可以分别对他们三项进行优化得到 $D\left(\Theta,\Theta^{(k)}\right)$ 的最大值。

首先处理第一项，

$$\sum_Q p(O,Q\mid\Theta^{(k)})\log\pi_{q_0}=\sum_{i=1}^{N}p(O,q_0=i\mid\Theta^{(k)})\log\pi_i \tag{33}$$

在约束：

$$\Sigma_{i=1}^{N}\pi_i=1 \tag{34}$$

下极大化：

$$\max_{\{\pi_1,\pi_2,\cdots,\pi_N\}}\sum_{i=1}^{N}p(O,q_0=i\mid\Theta^{(k)})\log\pi_i \tag{35}$$

可以得到 $\left\{\pi_1,\pi_2,\cdots,\pi_N\right\}$ 解：

$$\pi_i^{(k+1)}=\frac{p\left(O,q_0=i\mid\Theta^{(k)}\right)}{p\left(O\mid\Theta^{(k)}\right)} \tag{36}$$

类似的，从(32)的第二项得到：

$$\sum_Q\left(p(Q,O\mid\Theta^{(k)})\sum_{t=1}^{T}\log a_{q_{t-1}q_t}\right)=\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\log a_{i,j}\,p(O,q_{t-1}=i,q_t=j\mid\Theta^{(k)}) \tag{37}$$

在约束：

$$\Sigma_{j=1}^{N}a_{i,j}=1 \tag{38}$$

极大化(37)得到

$$a_{i,j}^{(k+1)}=\frac{\Sigma_{i=1}^{T}p\left(O,q_{t-1}=i,q_t=j\mid\Theta^{(k)}\right)}{\Sigma_{i=1}^{T}p\left(O,q_{t-1}=i\mid\Theta^{(k)}\right)} \tag{39}$$

从(32)的第三项得到：

$$\sum_{Q}\left(p(Q,O\mid\Theta^{(k)})\sum_{t=1}^{T}\log b_{q_t}\left(o_t\right)\right)=\sum_{i=1}^{N}\sum_{t=1}^{T}\log b_i(o_t)p(O,q_t=i\mid\Theta^{(k)}) \tag{40}$$

这里需要优化的是对应每一个状态的分布函数$b_i\left(o\right)$的形状，当$b_i\left(o\right)$的分布仅仅由若干个参数决定，比如是有限个离散取值的分布，或者是混合高斯分布，那样通过有限个参数优化可以确定极大化(40)的$b_i\left(o\right)$。

比如考虑$b_i\left(o_t\right)$取有限个离散值，$o_t\in\left\{v_1,v_2,\cdots,v_L\right\}$，这时对$b_i\left(o\right)$的约束成为：

$$\Sigma_{l=1}^{L}b_i\left(v_l\right)=1 \tag{41}$$

在此约束下(40)的极大化解为：

$$b_i^{(k+1)}\left(l\right)=\frac{\Sigma_{t=1}^{T}p\left(O,q_t=i\mid\Theta^{(k)}\right)\delta_{o_t,v_l}}{\Sigma_{t=1}^{T}p\left(O,q_t=i\mid\Theta^{(k)}\right)} \tag{42}$$

当$b_i\left(o_t\right)$是混合高斯模型时，

$$b_i\left(o\right)=\Sigma_{l=1}^{L}\alpha_{i,l}b_{i,l}\left(o\right) \tag{43}$$

其中$b_i\left(o\right)$表示处于i状态时，观察到的数据o的分布，(43)表示$b_i\left(o\right)$是来自于M个独立的随机数生成器$b_{i,l}\left(o\right)$的输出混合，混合比例（即选择开关的切换概率）为$\alpha_{i,l}$。

其中的每个随机数生成器$b_{i,l}\left(o\right)$的密度函数为：

$$b_{i,l}\left(o\right)=\left(2\pi\right)^{-\frac{K}{2}}\mid\mathbf{V}_{i,l}\mid^{-1/2}\exp\left(-\frac{1}{2}\left(o-\mu_{i,l}\right)^T\mathbf{V}_{i,l}^{-1}\left(o-\mu_{i,l}\right)\right) \tag{44}$$

其中K是高斯混合输出的随机向量的维数、$\mathbf{V}_{i,l}$和$\mu_{i,l}$分别是i状态的第l个高斯随机数发生器的协方差矩阵和均值。 对于这种情况，(29)的形式不一样，因为要加入混合高斯模型的混合变量作为未知变量。具体内容略去（请查阅参考文献）

（上述约束极大化算法用拉格朗日乘子法，细节请查看参考文献）

HMM 模型参数估计算法总结：

这里考虑到计算效率，用到下面几个公式：（请参阅参考文献以了解推导的细节）

$$\gamma_i(t) := p(q_t = i \mid O, \Theta) = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N} \alpha_j(t)\beta_j(t)} \tag{45}$$

$$\xi_{i,j}(t) := p(q_t = i, q_{t+1} = j \mid O, \Theta) = \frac{\gamma_i(t) a_{i,j} b_j(o_{t+1}) \beta_j(t+1)}{\beta_i(t)} \tag{46}$$

而 $\alpha_i(t)$ 和 $\beta_i(t)$ 的计算由(14)和(16)给出。下面公式的 $\alpha_i^{(k)}(t)$ 和 $\beta_i^{(k)}(t)$ 计算也是通过(14)和(16)进行，只不过计算用的 HMM 参数来自于第 k 步递推结果 $\Theta^{(k)}$。

参数识别的递推公式为：

$$\pi_i^{(k+1)} = \gamma_i^{(k)}(1) \tag{47}$$

$$a_{i,j}^{(k+1)} = \frac{\sum_{t=1}^{T-1} \xi_{i,j}^{(k)}(t)}{\sum_{t=1}^{T-1} \gamma_i^{(k)}(t)} \tag{48}$$

$$b_i^{(k+1)}(l) = \frac{\sum_{t=1}^{T} \delta_{o_t, v_l} \gamma_i^{(k)}(t)}{\sum_{t=1}^{T} \gamma_i^{(k)}(t)} \tag{49}$$

（注意：根据(36)递推公式(47)应该是 $\pi_i^{(k+1)} = \gamma_i^{(k)}(0)$，但实际计算）

5. 附录

5.1. (12)的证明

$$\begin{aligned}
& p(O \mid \Theta) := p(o_1, o_2, \cdots, o_T \mid \Theta) \\
& = \sum_{Q \in \mathbf{Q}} p(q_1, q_2, \cdots, q_T, o_1, o_2, \cdots, o_T \mid \Theta) \\
& = \sum_{q_1=1}^{N} \sum_{q_2=1}^{N} \cdots \sum_{q_T=1}^{N} p(q_1, q_2, \cdots, q_T, o_1, o_2, \cdots, o_T \mid \Theta) \\
& = \sum_{q_1=1}^{N} \sum_{q_2=1}^{N} \cdots \sum_{q_T=1}^{N} p(q_1 \mid \Theta) a_{q_1 q_2} p(q_2 \mid \Theta) a_{q_2 q_3} \cdots a_{q_{T-1} q_T} p(q_T \mid \Theta) b_{q_1}(o_1) b_{q_2}(o_2) \cdots b_{q_T}(o_T) \\
& = \sum_{q_1=1}^{N} \sum_{q_2=1}^{N} \cdots \sum_{q_T=1}^{N} \left[\left(\prod_{t=1}^{T} p(q_t \mid \Theta) b_{q_t}(o_t) \right) \left(\prod_{t=1}^{T-1} a_{q_t q_{t+1}} \right) \right]
\end{aligned}$$

(50)

5.2. (14)的证明：

第一个表达式：

$$\begin{aligned}\alpha_i(1) &:= p(o_1, q_1 = i \mid \Theta) \\ &= p(q_1 = i \mid \Theta)\, p(o_1 \mid q_1 = i, \Theta) \\ &= \pi_i b_i(o_1)\end{aligned} \tag{51}$$

第二个表达式：

$$\begin{aligned}&\left[\sum_{i=1}^{M} \alpha_i(t) a_{i,j}\right] b_j(o_{t+1}) \\ &= \left[\sum_{i=1}^{M} p(o_1, o_2, \cdots, o_t, q_t = i \mid \Theta)\, p(q_{t+1} = j \mid q_t = i, \Theta)\right] b_j(o_{t+1}) \\ &\overset{(a)}{=} \left[\sum_{i=1}^{M} p(o_1, o_2, \cdots, o_t, q_t = i \mid \Theta)\, p(q_{t+1} = j \mid o_1, o_2, \cdots, o_t, q_t = i, \Theta)\right] b_j(o_{t+1}) \\ &= \left[\sum_{i=1}^{M} p(o_1, o_2, \cdots, o_t, q_t = i, q_{t+1} = j \mid \Theta)\right] b_j(o_{t+1}) \\ &= p(o_1, o_2, \cdots, o_t, q_{t+1} = j \mid \Theta)\, b_j(o_{t+1}) \\ &= p(o_1, o_2, \cdots, o_t, q_{t+1} = j \mid \Theta)\, p(o_{t+1} \mid q_{t+1} = j, \Theta) \\ &\overset{(b)}{=} p(o_1, o_2, \cdots, o_t, q_{t+1} = j \mid \Theta)\, p(o_{t+1} \mid o_1, o_2, \cdots, o_t, q_{t+1} = j, \Theta) \\ &= p(o_1, o_2, \cdots, o_t, o_{t+1}, q_{t+1} = j \mid \Theta) \\ &= \alpha_j(t+1)\end{aligned} \tag{52}$$

其中等号(a)用到了(7)，等号(b)用到了(8)

第三个表达式：

$$\begin{aligned}&\sum_{i=1}^{N} \alpha_i(T) \\ &= \sum_{i=1}^{N} p(o_1, o_2, \cdots, o_T, q_T = i \mid \Theta) \\ &= p(o_1, o_2, \cdots, o_T \mid \Theta) \\ &= p(O \mid \Theta)\end{aligned} \tag{53}$$

5.3. (16)的证明：

注意，根据定义(15)，$\beta_i(T)$的定义不存在，否则$\beta_i(T)$将变成$\beta_i(T) := p(\mid q_T = i, \Theta)$（只有条件概率的条件，没有随机变量的数值），因此人为定义：

$$\beta_i\left(T\right)=1 \tag{54}$$

第二个表达式的证明：

$$\begin{aligned}
&\sum_{j=1}^{N} a_{i,j} b_j\left(o_{t+1}\right)\beta_j\left(t+1\right)\\
&=\sum_{j=1}^{N} a_{i,j}\, p\left(o_{t+1}\mid q_{t+1}=j,\Theta\right) p\left(o_{t+2},o_{t+3},\cdots,o_T\mid q_{t+1}=j,\Theta\right)\\
&\overset{(a)}{=}\sum_{j=1}^{N} a_{i,j}\, p\left(o_{t+1}\mid q_{t+1}=j,\Theta\right) p\left(o_{t+2},o_{t+3},\cdots,o_T\mid o_{t+1},q_{t+1}=j,\Theta\right)\\
&=\sum_{j=1}^{N} a_{i,j}\, p\left(o_{t+1},o_{t+2},o_{t+3},\cdots,o_T\mid q_{t+1}=j,\Theta\right)\\
&=\sum_{j=1}^{N} p\left(q_{t+1}=j\mid q_t=i,\Theta\right) p\left(o_{t+1},o_{t+2},o_{t+3},\cdots,o_T\mid q_{t+1}=j,\Theta\right)\\
&\overset{(b)}{=}\sum_{j=1}^{N} p\left(q_{t+1}=j\mid q_t=i,\Theta\right) p\left(o_{t+1},o_{t+2},o_{t+3},\cdots,o_T\mid q_{t+1}=j,q_t=i,\Theta\right)\\
&=\sum_{j=1}^{N} p\left(o_{t+1},o_{t+2},o_{t+3},\cdots,o_T,q_{t+1}=j\mid q_t=i,\Theta\right)\\
&=p\left(o_{t+1},o_{t+2},\cdots,o_T\mid q_t=i,\Theta\right)\\
&=\beta_i\left(t\right)
\end{aligned} \tag{55}$$

其中等号(a)和(b)来自于(8)。

第三个表达式的证明：

$$
\begin{aligned}
&\sum_{i=1}^{N}\beta_i(1)\pi_i b_i(o_1)=\sum_{i=1}^{N}\beta_i(1)b_i(o_1)\pi_i\\
&=\sum_{i=1}^{N}p(o_2,o_3,\cdots,o_T\mid q_1=i,\Theta)\,p(o_1\mid q_1=i,\Theta)\pi_i\\
&=\sum_{i=1}^{N}p(o_2,o_3,\cdots,o_T\mid o_1,q_1=i,\Theta)\,p(o_1\mid q_1=i,\Theta)\pi_i\\
&=\sum_{i=1}^{N}p(o_1,o_2,o_3,\cdots,o_T\mid q_1=i,\Theta)\pi_i\\
&=\sum_{i=1}^{N}p(o_1,o_2,o_3,\cdots,o_T\mid q_1=i,\Theta)\,p(q_1=i\mid\Theta)\\
&=\sum_{i=1}^{N}p(o_1,o_2,o_3,\cdots,o_T\mid q_1=i,\Theta)\,p(q_1=i\mid\Theta)\\
&=\sum_{i=1}^{N}p(o_2,o_3,\cdots,o_T,q_1=i\mid\Theta)\\
&=p(o_1,o_2,\cdots,o_T\mid\Theta)\\
&=p(O\mid\Theta)
\end{aligned}
\tag{56}
$$

5.4. Viterbi 算法的证明

$$
\begin{aligned}
&\max_{1\le i\le N}\left(\delta_{t-1}(i)a_{i,j}\right)b_j(o_t)\\
&=\max_{1\le i\le N}\left(\max_{q_1,q_2,\cdots,q_{t-1}}p(q_1,q_2,\cdots,q_{t-2},q_{t-1}=i\mid o_1,o_2,\cdots,o_{t-1},\Theta)a_{i,j}\right)b_j(o_t)\\
&\overset{(a)}{=}\max_{1\le i\le N}\left(p\left(q_1^*,q_2^*,\cdots,q_{t-2}^*,q_{t-1}=i\mid o_1,o_2,\cdots,o_{t-1},\Theta\right)p\left(q_t=i\mid q_{t-1}=i,\Theta\right)\right)b_j(o_t)\\
&\overset{(b)}{=}\max_{1\le i\le N}\left(p\left(q_1^*,q_2^*,\cdots,q_{t-2}^*,q_{t-1}=i\mid o_1,o_2,\cdots,o_{t-1},\Theta\right)p\left(q_t=i\mid o_1,o_2,\cdots,o_{t-1},q_{t-1}=i,\Theta\right)\right)b_j(o_t)\\
&=\max_{q_1,q_2,\cdots,q_{t-1}}p(q_1,q_2,\cdots,q_{t-1},q_t=j\mid o_1,o_2,\cdots,o_t,\Theta)=\delta_t(j)
\end{aligned}
\tag{57}
$$

上面(a)中 q_n^* 表示上一个等号里 max 的结果

6. 附录

Jeff A. Bilmes, *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*

$$\log(a+b) = \log\left(a\left(1+\frac{b}{a}\right)\right)$$

$$= \log a + \log\left(1+\frac{b}{a}\right)$$

$$= \log a + \left(\frac{b}{a} - \frac{1}{2}\left(\frac{b}{a}\right)^2 + \frac{1}{3}\left(\frac{b}{a}\right)^3 - \frac{1}{4}\left(\frac{b}{a}\right)^4 + \cdots\right)$$

$$\left|\frac{b}{a}\right| < 1$$

$$\frac{b}{a} = \exp(\log b - \log a)$$