在https://github.com/Casey1203/ml-ease/blob/master/hmm/HMM%E6%A8%A1%E5%9E%8B-%E5%AE%9A%E4%B9%89.md文章中给出了有关HMM模型应用的三个问题，接下来介绍第一个问题：概率计算。

# 2. 概率计算算法

## 2.1 直接计算法

给定模型$\lambda$和观测序列$O=(o_1,o_2,\dots,o_T)$，计算$O$出现的概率$P(O|\lambda)$，直接按照概率公式计算，枚举出所有可能的状态序列$I=(i_1,i_2,\dots,i_T)$，长度为$T$。然后求各状态序列$I$与观测序列$O=(o_1,o_2,\dots,o_T)$的联合概率分布$P(O,I|\lambda)$，再把所有可能的状态序列积分掉，得到$P(O|\lambda)$。

$O$和$I$同时出现的联合概率$P(O,I|\lambda)=P(O|I,\lambda)P(I|\lambda)$。根据给定的参数$\lambda=(A,B,\pi)$，求$P(O|I,\lambda)$和$P(I|\lambda)$。

$$P(I|\lambda)=\pi_{i_1}a_{i_1i_2}\dots a_{i_ti_{t+1}}\dots a_{i_{T-1}i_T}$$

其中$a_{i_t}a_{i_{t+1}}$为状态从$i_t$转移到$i_{t+1}$的概率。

$$P(O|I,\lambda)=b_{i_1}(o_1)b_{i_2}(o_2)\dots b_{i_t}(o_t)\dots b_{i_T}(o_T)$$

其中$b_{i_t}(o_t)$为在状态$i_t$的情况下，观测到$o_t$的概率。

$$\begin{aligned}P(O,I|\lambda)&=P(O|I,\lambda)P(I|\lambda)\\&=\pi_{i_1}b_{i_1}(o_1)a_{i_1i_2}b_{i_2}(o_2)\dots a_{i_ti_{t+1}}b_{i_{t+1}}(o_{t+1})\dots a_{i_{T-1}i_T}b_{i_T}(o_T)\end{aligned}$$

最后，因为我们不关心状态序列，因此要把$I$给积分掉，因此

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

注意，$\sum_{i_1,i_2,\dots,i_T}$的含义是$\sum_{i_1}\sum_{i_2}\dots\sum_{i_T}$，这里有$T$个$\sum$，每个$\sum$有$N$种情况，表示$N$种状态。因此$\sum_{i_1,i_2,\dots,i_T}$的复杂度是$N^T$。同时$\sum_{i_1,i_2,\dots,i_T}\pi_{i_1}b_{i_1}(o_1)a_{i_1i_2}b_{i_2}(o_2)\dots a_{i_ti_{t+1}}b_{i_{t+1}}(o_{t+1})\dots a_{i_{T-1}i_T}b_{i_T}(o_T)$这个连乘有$2T$个数，因此总体的复杂度是$O(TN^T)$，是很庞大的计算量，因此该算法在实际情况下是不可行的。

## 2.2 前向算法和后向算法（动态规划）

首先定义前向概率和后向概率。

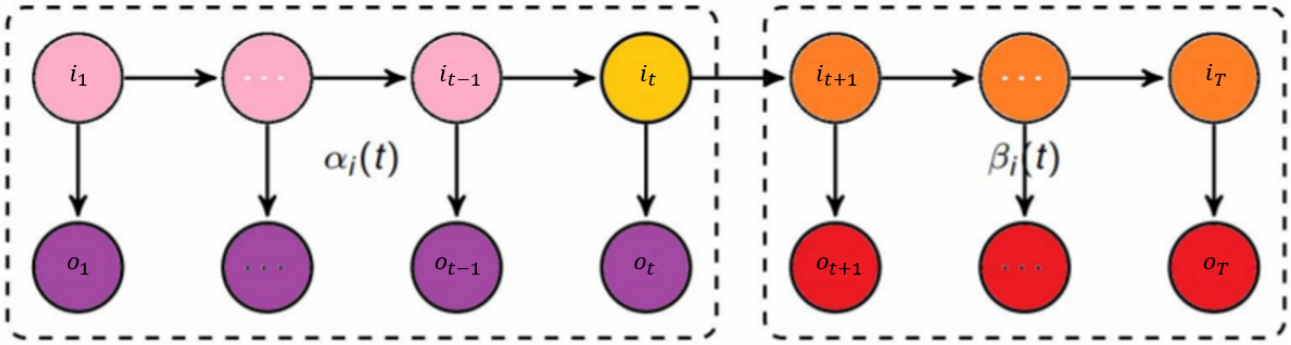
前向概率：给定HMM的参数$\lambda$，定义到$t$时刻的部分观测序列为$o_1,o_2,\dots,o_t$，且状态位于$q_i$，发生这样的事件的概率定义作前向概率

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

这里的记号看起来可能比较混乱，再回顾一下，时间到$t$的观测序列是$O_{1,t}=(o_1,o_2,\dots,o_t)$，同时状态序列为$I=(i_1,i_2,\dots,i_t)$，状态的取值范围是$Q=\{q_1,q_2,\dots,q_N\}$，这里仅考虑状态序列$I$中第$t$时刻的状态$i_t$取到了$Q$中的元素$q_i$。

后向概率：给定HMM的参数$\lambda$，定义在时刻$t$的状态为$q_i$的条件下，从$t+1$时刻到$T$时刻，观测到的部分观测序列$o_{t+1},o_{t+2},\dots o_T$，定义发生这样的时间的概率为后向概率

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




在这幅图中可以清楚的看到，第一行是状态序列，第二行是观测序列。图中的左半边，为观测到$(o_1, o_2, \dots, o_t)$，同时状态$i_t$位于$q_i$，即图中黄色的点。图中的右半边，为给定了$i_t = q_i$的条件下，观测到了$o_{t+1}, o_{t+2}, \dots o_T$的情况。

有了上面两个概率的定义，接下来可以介绍前向算法和后向算法了。

前向算法

初值：在$t = 1$时刻的前向概率

$$\alpha_1(i) = P(o_1, i_1 = i|\lambda) = \pi_i b_i(o_1), \quad i = 1, 2, \dots, N$$

这个式子表明，在$t = 1$时刻，状态位于$i$，且观测到的状态为$o_1$的概率，$\pi_i$为初始时刻下，选择状态$i$的概率，$b_i(o_1)$表示在状态$i$下，选择观测$o_1$的概率。

递推：对于$t = 1, 2, \dots, T-1$，

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

这是前向概率的递推公式，要求第$t+1$时刻状态位于$i$，而$t$时刻位于什么状态可以不关心。假定时刻$t$位于状态$j$，只需要在$t+1$时刻将状态转移到$i$即可，因此这样的事件发生的概率为$\alpha_t(j)a_{ji}$，表示在时刻$t$观测到了$o_1, o_2, \dots, o_t$并且时刻$t$处于状态$q_j$，而在时刻$t+1$到达状态$i$的联合概率。因为$t$时刻可以在任意一种状态，因此需要把$j$给积分掉，因此得到$\sum_{j=1}^{N} \alpha_t(j)a_{ji}$。由于前向概率的定义是在$t+1$时刻，还要看到观测$o_{t+1}$，因此还需要乘以$b_i(o_{t+1})$，表示第$t+1$时刻，从状态$i$得到观测$o_{t+1}$。

终止：有了递推公式，则可以计算$T$时刻的状态位于$i$，并且看到了观测$o_T$的概率$\alpha_T(i)$。由于我们并不关心最终的状态位于哪个状态，因此需要把$i$积分掉，因此

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

前向算法是基于“状态序列的路径结构”递推计算$P(O|\lambda)$的算法。具体为，在$t = 1$时刻，计算$\alpha_1(i)$有$N$个值，分别表示$t = 1$时刻位于$N$个状态的某一个。然后在计算下一个时刻$t+1$的状态的前向概率$\alpha_{t+1}(i)$时，有$N$个前向概率需要计算，均利用了前一个时刻$t$的$N$个前向状态$\alpha_t(j)$，因此相邻两次计算需要有$N^2$的复杂度。因为序列长度为$T$，因此需要执行$T$次这样相邻的递推计算，因此总体的复杂度为$O(TN^2)$。

下面给一个例子来运用前向算法

给定HMM模型$\lambda = (A, B, \pi)$，它们分别是

$$\pi = \begin{pmatrix} 0.2 \\ 0.4 \\ 0.4 \end{pmatrix} \quad A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix} \quad B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}$$

状态集合是$Q = \{1, 2, 3\}$，观测集合是$V = \{\text{红}，\text{白}\}$，假设序列长度为$T = 3$，观测到了$O = (\text{红}，\text{白}，\text{红})$，用前向算法求$P(O|\lambda)$。

初值

$$\begin{aligned} \alpha_1(1) &= \pi_1 b_1(o_1) = 0.2 \times 0.5 = 0.1 \\ \alpha_1(2) &= \pi_2 b_2(o_1) = 0.4 \times 0.4 = 0.16 \\ \alpha_1(3) &= \pi_3 b_3(o_1) = 0.4 \times 0.7 = 0.28 \end{aligned}$$

递推

$T = 2$：

$$\begin{aligned} &\alpha_2(1) = \left(\sum_{j=1}^{N} \alpha_1(j) a_{j1}\right) b_1(o_2) \\ &= (0.1 \times 0.5 + 0.16 \times 0.3 + 0.28 \times 0.2) \times 0.5 \\ &= 0.077 \\ &\alpha_2(2) = \left(\sum_{j=1}^{N} \alpha_1(j) a_{j2}\right) b_2(o_2) \\ &= (0.1 \times 0.2 + 0.16 \times 0.5 + 0.28 \times 0.3) \times 0.6 \\ &= 0.1104 \\ &\alpha_2(3) = \left(\sum_{j=1}^{N} \alpha_1(j) a_{j2}\right) b_3(o_2) \\ &= (0.1 \times 0.3 + 0.16 \times 0.2 + 0.28 \times 0.5) \times 0.3 \\ &= 0.0606 \end{aligned}$$

$T = 3$：

$$\begin{aligned} &\alpha_3(1) = \left(\sum_{j=1}^{N} \alpha_2(j) a_{j1}\right) b_1(o_3) \\ &= (0.077 \times 0.5 + 0.1104 \times 0.3 + 0.0606 \times 0.2) \times 0.5 \\ &= 0.04187 \\ &\alpha_3(2) = \left(\sum_{j=1}^{N} \alpha_2(j) a_{j2}\right) b_2(o_3) \\ &= (0.077 \times 0.2 + 0.1104 \times 0.5 + 0.0606 \times 0.3) \times 0.4 \\ &= 0.03551 \\ &\alpha_3(3) = \left(\sum_{j=1}^{N} \alpha_2(j) a_{j3}\right) b_3(o_3) \\ &= (0.077 \times 0.3 + 0.1104 \times 0.2 + 0.0606 \times 0.5) \times 0.7 \\ &= 0.05284 \end{aligned}$$

最终，$P(O|\lambda) = 0.04187 + 0.03551 + 0.05284 = 0.13022$

后向算法

和前向算法的思路相反，后向算法先从时刻$T$开始。回顾后向概率的定义

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

初值：$t = T$时刻，$\beta_T(i) = 1, \quad i = 1, 2, \ldots, N$，表示$T$时刻，状态位于$i$。由于后面已经没有观测了，因此对于一个只要前提，不要结论的概率等于1。

递推：对于$t=T-1,\dots,1$

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

这是后项概率的递推公式。因为后项概率要求在第$t$时刻位于$i$状态，且不需要获得观测$o_t$，因此在计算从$t+1$时刻的后项概率推出$t$时刻的后项概率时，只需考虑在时刻$t$到时刻$t+1$转移的所有可能的$N$个状态$j$的转移概率$a_{ij}$，以及在$t+1$时刻，$j$状态下得到的观测$o_{t+1}$的概率$b_j(o_{t+1})$。此时已经完成了从$t$时刻的状态$i$，完成了向$t+1$时刻的状态$j$转变，同时在状态$j$观测到了$o_{t+1}$，概率为$a_{ij}b_j(o_{t+1})$。之后再考虑$t+1$时刻在状态$j$的后项概率$\beta_{t+1}(j)$。由于我们不关心$t+1$时刻位于哪个状态$j$，我们只关心$t$时刻在状态$i$就行，因此还要把$j$积分掉。

终止：

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

在$t=1$时刻，后项概率表示状态位于$i$，得到了$o_2,o_3,\dots,o_T$的观测序列。因此还需要乘上一开始进入状态$i$的概率$\pi_i$和在状态$i$下观测到$o_1$的概率$b_i(o_1)$。

最后，由于$t=1$时刻位于哪个状态我们不考虑，因此需要把$i$给积分掉。

## 2.3 前向后向概率的关系

把前向概率和后向概率相乘，有以下式子

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

这个式子的含义是观测到了观测序列$O=(o_1,o_2,\dots,o_T)$的同时，$t$时刻位于状态$i$的联合概率。

利用这个联合概率，可以求

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

这是给定HMM模型参数，以及观测的情况下，$t$时刻位于状态$i$的概率，定义为$\gamma_t(i)$。

因为$P(O|\lambda)=\sum_i^N P(i_t=i,O|\lambda)$中把状态$i$给积分掉了，因此

$$\begin{aligned}\gamma_t(i)&=\frac{P(i_t=i,O|\lambda)}{\sum_{i=1}^{N}P(i_t=i,O|\lambda)}\\&=\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}\end{aligned}$$

$\gamma_t(i)$的意义在于，它表示在$t$时刻下，出现状态$i$的概率，因此在每个时刻下，选择最大的$\gamma$对应的状态，为在该时刻下最有可能出现的状态$i_t^\star$，从而得到一个状态序列$I=(i_1^\star,i_2^\star,\dots,i_T^\star)$作为预测的结果。

再更进一步，定义$\xi(i,j)$为在$t$时刻位于状态$i$，同时在$t+1$时刻位于状态$j$的联合概率。

因此

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

注意到

$$\begin{aligned}P(i_t=i,i_{t+1}=j,O|\lambda)&=P(i_t=i,O_{1\to t}|\lambda)P(i_{t+1}=j|i_t=i,\lambda)P(o_{t+1}|i_{t+1}=j,\lambda)P(O_{t+2\to T}|i_{t+1}=j,\lambda)\\&=P(i_t=i,O_{1\to t}|\lambda)P(i_{t+1}=j|i_t=i,\lambda)P(O_{t+1\to T}|i_{t+1}=j,\lambda)\\&=P(i_t=i,O_{1\to t}|\lambda)P(O_{t+1\to T},i_{t+1}=j|i_t=i,\lambda)\\&=P(i_t=i,i_{t+1}=j,O_{1\to T}|\lambda)=P(i_t=i,i_{t+1}=j,O|\lambda)\end{aligned}$$

其中，

$$\begin{aligned}P(i_t=i,O_{1\to t}|\lambda)&=\alpha_t(i)\\P(i_{t+1}=j|i_t=i,\lambda)&=a_{ij}\\P(o_{t+1}|i_{t+1}=j,\lambda)&=b_j(o_{t+1})\\P(O_{t+2\to T}|i_{t+1}=j,\lambda)&=\beta_{t+1}(j)\end{aligned}$$

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

进而

$$\xi_t(i,j)=P(i_t=i,i_{t+1}=j|O,\lambda)=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}$$

因为$\gamma_t(i)$和$\xi_t(i,j)$都是针对任意一个时刻$t$的，因此如果对时间$t$求和，可以得到它们的期望，即

1. 在观测$O$出现的情况下，状态$i$出现的期望

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

2. 在观测O出现的情况下，由状态$i$转移的期望

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

3. 在观测O出现的情况下，由状态$i$转移到状态$j$的期望

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