HMM模型--概率计算

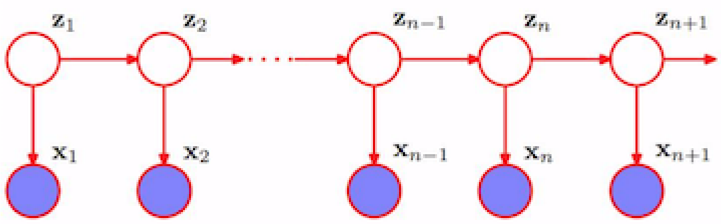
隐马尔科夫模型（HMM）可用于标注问题，在语音识别、NLP、生物信息（DNA）、模式识别等领域被实践证明是有效的算法。

1. HMM的定义

1.1 标记说明

HMM是关于时序的概率模型，描述一个过程：由一个隐藏的马尔科夫链随机生成不可观测的状态随机序列，同时每个状态会产生一个观测，进而形成观测随机序列。

由隐藏的马尔科夫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，形成的观测的序列称为观测序列（observation sequence）。这两个序列的每个位置可以看作一个时刻。



图中$z_1, z_2 \ldots z_{n+1}$是状态序列，$x_1, x_2 \ldots x_{n+1}$是观测序列。

HMM由初始状态分布π，状态转移概率分布A以及观测概率分布B确定。以三元符号表示，即

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

接着，定义Q是所有可能的状态集合，N是可能的状态数。

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

定义V是可能的观测的集合，M是可能的观测数。

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

接下来定义状态序列I和对应的观测序列O，它们的长度都为T

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

状态转移概率分布A是一个$N \times N$的矩阵，即$A = [a_{ij}]_{N \times N}$

其中，$a_{ij} = P(i_{t+1} = q_j | i_t = q_i)$表示在$t$时刻处于状态$q_i$的条件下，在$t+1$时刻下转移到了状态$q_j$的概率。

观测概率分布B是一个$N \times M$的矩阵，即$B = [b_{i,k}]_{N \times M}$

其中，$b_{ik} = b_{q_i}(v_k) = P(o_t = v_k | i_k = q_i)$表示在$t$时刻处于状态$q_i$的条件下，生成观测$v_k$的概率。

初始状态分布π是一个$N \times 1$的向量，其中$\pi_i = P(i_1 = q_i)$表示在1时刻处于状态q_i的概率。

举个例子说明观测和状态：在做语音识别系统的时候，系统听到的是人发出的声音，严格来说系统看到的是声音的波形，这是观测。这个语音识别系统的目的是为了把人发出的声音识别成对应的文字，这个文字就是状态，即隐藏在波形背后对应的文字。

再举个例子，在做中文分词的工作时，系统看到的是一句话，则这句话就是观测。这句话里包含有词和字，系统要判断应该在哪个字后面切一刀，那么系统就要判断每个字是否为一个词的结尾，即这个字是否应该被切开，那是/否切开，这就是状态。

为了对$\lambda=(A,B,\pi)$有更加深刻的认识，下面举《统计学习方法》上的一个例子来说明。

假设有4个合资，每个盒子里都装有红白两种颜色的球，盒子里的红白球数目见下表。

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

首先以等概率随机抽取一个盒子。接着从这个盒子中随机抽取一个球，记录颜色后放回。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,那么下一个盒子一定是盒子2；如果当前盒子时盒子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。如果当前的盒子是盒子4，那么各以0.5的概率停留在盒子4或转移到盒子3，确定完盒子后，再从里面随机抽取一个球，并记录颜色后放回。如此重复下去。

在这个例子中，有四个盒子，盒子对应着状态。因此状态集合

$$Q=\{盒子1,盒子2,盒子3,盒子4\},\quad N=4$$

球的颜色对应着观测，观测集合

$$V=\{红,白\},\quad M=2$$

观测序列和状态序列的长度，取决于这样抽取球的次数重复了几次。

因此，初始概率分布

$$\pi=(0.25,0.25,0.25,0.25)^T$$

状态转移概率分布

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

观测概率分布为

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

1.2 HMM的两个假设

1. 齐次马尔科夫性假设：假设隐藏的马尔科夫链在任意时刻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。

$$P(i_t|i_{t-1},o_{t-1,\ldots,i_1,o_1})=P(i_t|i_{t-1}),\quad t=1,2,\ldots,T$$

1. 观测独立性假设：假设任意时刻的观测，只依赖于该时刻的马尔科夫链的状态，与其他观测及状态无关。

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

1.3 HMM的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$O=(o_1,o_2,\ldots,o_T)$，计算在模型λ的控制下观测序列O出现的概率$P(O|\lambda)$。
2. 学习问题：已知观测序列$O=(o_1,o_2,\ldots,o_T)$，估计模型$\lambda=(A,B,\pi)$参数，使得在该模型下观测序列的概率$P(O|\lambda)$最大。$P(O|\lambda)$是似然函数，因此可以套用极大似然估计的方法来估计参数λ。
3. 预测问题：也称为解码。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$I=(i_1,i_2,\ldots,i_T)$，即给定观测序列，求最有可能的对应的状态序列I。

第一个问题，是计算概率的问题。以1.1小节末尾的例子来说，假设我重复取5次，得到的观测结果是

$$O=(\text{红，红，白，白，红})$$

求得到这个观测结果的概率有多大。

第二个问题，在未知模型参数λ的时候，给定一个观测序列，找到能够使得看到这个观测序列的概率最大的模型参数。这个学习过程中，含有隐状态I，即挑选到的盒子序列，因此本质是EM算法的过程。

第三个问题，已知模型的参数λ（也可以认为是经过了第二个问题后，学到了一个参数λ），以及观测序列，想要知道最有可能得到这个观测序列的状态序列I。这其实就是分词的过程，给定一个训练好的HMM模型（λ），以及一句话（O），想要找到最合适的分词结果（I）。