隐马尔可夫模型

隐马尔科夫模型是概率图模型中的一种，主要用于序列数据的建模，如语音识别，序列标注等。

1. 马尔可夫模型

考虑一个不相互独立的随机变量组成的序列，序列中每个变量的取值依赖于前一个变量。也就是说，给定序列中的一个元素，序列中未来的元素与过去的元素是条件独立的。

对于序列$X=(X_1,\dots X_T)$，其中X_i 取值于状态空间$S=\{s_1,\dots s_N\}$，如果X满足有限视野和时间不变性两条性质，即：

- 有限视野：$P(x_{t+1}=s_k|X_1,\dots X_t)=P(X_{t+1}=s_k|X_t)$
- 时间不变性：$P(x_{t+1}=s_k|X_t)=P(X_2=s_k|X_1)$

则序列X称为一个马尔可夫链。一个马尔可夫链可以用转移矩阵，即状态之间相互转化的概率来描述：

$$a_{ij}=P(X_{t+1}=s_j|X_t=s_i)$$

其中，$a_{ij}\geq 0,\forall i,j$且$\sum_{j=1}^N a_{ij}=1,\forall i$。

除此以外，还需要指定π，表示马尔可夫链中不同初始状态的概率。

$$\pi_i=P(X_1=s_i)$$

其中，$\sum_{i=1}^N \pi_i=1$。

一个马尔可夫链中，状态序列$X_1,\dots X_N$的概率等于各个状态转移概率的乘积，即：

$$P(X_1,\dots X_T)=P(X_1)P(X_2|X_1)\dots P(X_T|X_{T-1})=\pi_{X_1}\prod_{t=1}^{T-1}a_{XtXt+1}$$

在实际应用中，最重要的一点是能否将一个过程转化成一个马尔可夫模型。《统计自然语言处理》第九章中有如下例子：一个$n\geq 3$的n-gram模型不是马尔可夫模型，因为不符合有限视野性质。但如果将适量的历史信息转化成状态空间，即状态是n-1 gram的，那n-gram模型就被重构成了一个马尔可夫模型。

2. 隐马尔可夫模型

在隐马尔可夫模型（HMM）中，状态序列是不可见的，但能观测到由状态序列生成的观测序列。当一个系统中，表层事件可能是由底层事件引发的时候，如词性标注任务中，可以认为文本中的词语是由词性标注序列生成的，HMM能有效地对其进行建模。

一个HMM中的模型参数μ可以用五元组(S,K,π,A,B)表示。其中，S和K分别是状态的集合和发射输出的字母表，π,A,B分别是初始状态的概率，状态转移矩阵和发射概率。HMM分为弧发射和状态发射两种：弧发射HMM中，t时刻观测到的输出取决于t时刻和$t+1$时刻的状态；而状态发射HMM中，t时刻观测到的输出只依赖于t时刻的状态。发射概率$B=\{b_{ijk}\},i,j\in S,k\in K$，表示从状态$i$转移到状态$j$时，观测到发射输出$k$的概率。令

$\forall j', j'', b_{ij'k} = b_{ij''k}$，可以将状态发射HMM看成一个弧发射HMM。

3. 隐马尔可夫模型的三个基本问题

隐马尔可夫模型有三个基本问题：

- 给出一个模型μ的参数，如何计算某个观测到的输出序列O的的概率？
- 给出观测到的输出序列O和模型参数μ，如何选择一个状态序列X,使其能够最好地解释观测到的输出序列？
- 给定观测到的输出序列O，如何找到一个最好地解释这个观测序列的模型？

3.1 计算输出序列的概率

给定观测到的输出序列$O = o_1, \ldots o_T$和模型$\mu = (A, B, \pi)$，输出序列的概率为$P(O|\mu)$。计算这个概率可以使用动态规划算法完成。前向(forward)过程和后向(backward)过程都可以用于计算这一概率。

- 前向过程从前向后计算：定义$\alpha_i(t)$表示在t时刻以状态S_i结束的概率，即:

$$\alpha_i(t) = P(o_1 o_2 \ldots o_{t-1}, X_t = i|\mu)$$

则有：

$$\begin{aligned} \alpha_i(1) =& \pi_i \\ \alpha_j(t+1) =& \sum_{i=1}^{N} \alpha_i(t) a_{ij} b_{ijo_t} \end{aligned}$$

求出$\alpha_i(T+1)$后，求和可得输出序列O的概率：

$$P(O|\mu) = \sum_{i=1}^{N} \alpha_i(T+1)$$

- 后向过程从后向前计算：定义$\beta_i(t)$表示给定时刻t的状态为S_i时，观测到输出序列的剩余部分的概率，即：

$$\beta_i(t) = P(o_t \ldots o_T|X_t = i, \mu)$$

则有：

$$\begin{aligned} \beta_i(T+1) &= 1 \\ \beta_i(t) &= \sum_{j=1}^{N} a_{ij} bijo_t \beta_j(t+1) \end{aligned}$$

求出$\beta_i(1)$后，求和可得输出序列O的概率：

$$P(O|\mu) = \sum_{i=1}^{N} \pi_i \beta_i(1)$$

- 更一般地，结合前向和后向过程，可得：

$$\begin{aligned} P(O, X_t = i|\mu) =& P(o_1 \ldots o_T, X_t = i|\mu) \\ =& P(o_1 \ldots o_t - 1, X_t = i|\mu) \times P(o_t \ldots o_T|X_t = i, \mu) \\ =& \alpha_i(t)\beta_i(t) \end{aligned}$$

因此，对于任意t，均有：

$$P(O|\mu) = \sum_{i=1}^{N} \alpha_i(t)\beta_i(t)$$

3.2 确定最佳状态序列

给定一个观测序列O和模型参数μ，确定最佳状态序列X意味着找到给定O和μ的条件下，概率最大的X，即：

$$argmax_X P(X|O,\mu)$$

由于O是已知的，上式等同于：

$$argmax_X P(X,O|\mu)$$

一个计算这个状态序列X的有效算法称为Viterbi算法，它也是一个基于动态规划的算法。定义

$$\delta_j(t) = max_{X_1,\ldots X_{t-1}} P(X_1,\ldots X_{t-1}, o_1,\ldots o_{t-1}, X_t = j|\mu)$$

即在给定O和μ时，t时刻到达状态j的最可能路径的概率。易知：

$$\delta_j(1) = \pi_j$$
$$\delta_j(t+1) = max_{1\leq i\leq N} \delta_i(t) a_{ij} b_{ijo_t}$$

同时，用$\psi_j(t+1) = argmax_{1\leq i\leq N} \delta_i(t) a_{ij} b_{ijo_t}$记录每个节点的入弧，方便最终通过回溯的方法解码出最优路径。动态规划计算完成后，最可能的状态序列可以通过以下回溯算法得到：

$$\hat{X}_{T+1} = argmax_{1\leq i\leq N} \delta_i(T+1)$$
$$\hat{X}_t = \psi_{\hat{X}_{t+1}}(t+1)$$

3.3 隐马尔可夫模型的参数估计

隐马尔可夫模型的参数估计问题指：给定一个特定的观测序列O，希望确定模型μ的参数值，使得$P(O|\mu)$最大。目前，没有解析的算法来选择μ，一般使用前向-后向(Backward-Forward)算法来求局部最优解。这个算法是EM算法的一个特例。

算法的大致执行过程如下:

- 使用某个模型（可能是随机初始化的）算出观测序列的概率。
- 查看计算过程，发现某个状态转移或者输出发射出现的次数最多。增加它们的概率以得到一个新的模型。
- 新的模型能为观测序列给出更高的概率。

定义$p_t(i,j)$为给定观测序列的情况下，在t时刻经过了弧(i,j)的概率。则有：

$$\begin{aligned} p_t(i,j) =& P(X_t = i, X_{t+1} = j|O,\mu) \\ =& \frac{P(X_t = i, X_{t+1} = j, O|\mu)}{P(O|\mu)} \\ =& \frac{\alpha_i(t) a_{ij} b_{ijo_t} \beta_j(t+1)}{\sum_{m=1}^{N} \alpha_m(t)\beta_m(t)} \\ =& \frac{\alpha_i(t) a_{ij} b_{ijo_t} \beta_j(t+1)}{\sum_{m=1}^{N}\sum_{n=1}^{N} \alpha_m(t) a_{mn} b_{mno_t} \beta_n(t+1)} \end{aligned}$$

令$\gamma_i(t)=\sum_{j=1}^{N} p_t(i,j)$，则有:

- $\sum_{t=1}^{T} \gamma_i(t)$为状态$i$发出的转移的期望数目；
- $\sum_{t=1}^{T} p_t(i,j)$为从状态i到状态j的转移的期望数目;

因此，我们可以从预先选择（或随机选择）的模型$\mu=(A,B,\pi)$开始，用模型μ运行观测序列O来估计模型中各个参数值的期望。不断重复这个过程，直到模型收敛到μ的最优值。模型参数值的估计方法如下：

- $\pi_i=\gamma_i(1)$
- $a_{ij}=\frac{\sum_{t=1}^{T} p_t(i,j)}{\sum_{t=1}^{T} \gamma_i(t)}$
- $b_{ijo_t}=\frac{\sum_{t:o_t=k} p_t(i,j)}{\sum_{t=1}^{T} p_t(i,j)}$

  换言之，对模型μ的参数进行重新估计后，可以得到模型$\hat{\mu}$，根据Baum的证明结果，有：

$$P(O|\hat{\mu}) \geq P(O|\mu)$$

值得一提的是，这个算法并不能保证得到最好的模型，因为重新估计参数这一过程会停留在局部极值点上。