马尔可夫模型(MM)

设系统有N个状态$S=\{S_1, S_2, \cdots, S_N\}$，随着时间的推移，系统将从某一状态转移到另一状态。设$Q=\{q_1, q_2, \cdots, q_T\}$为一随机变量序列，$q_t \in S$，$t=1,2,\cdots,T$，则系统在时间t处于状态$S_j$的概率取决于其在时间1,2,…,t-1的状态

$$P(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \cdots)$$

即现在状态只与历史状态有关

一阶马尔可夫过程：系统在时间t的状态只与其在t-1时间状态有关

$$P(q_t = S_j | q_{t-1} = S_i)$$

二元语法模型便是如此

马尔可夫模型：初始状态$\pi = (\pi_1, \pi_2, \cdots, \pi_N)$

π_i表示系统在初始时处于S_i状态的概率

以当下状态为横轴，下一时刻状态为纵轴，有：

	S_i	S_j	S_k	…
S_i	a_{ii}	a_{ji}	a_{ki}	
S_j	a_{ij}	a_{jj}	a_{kj}	
S_k	a_{ik}	a_{jk}	a_{kk}	
⋮				

…状态转移概率矩阵

$$P(q_{t+1} = S_j | q_t = S_i) = a_{ij}$$

a_{ij}表示从S_i变成S_j的状态转移概率

$a_{ij} \geq 0 \quad \sum_{j=1}^{N} a_{ij} = 1$ (列和必为1，行和无限制)

马尔可夫模型形式化定义：$M=(S, A, \pi)$

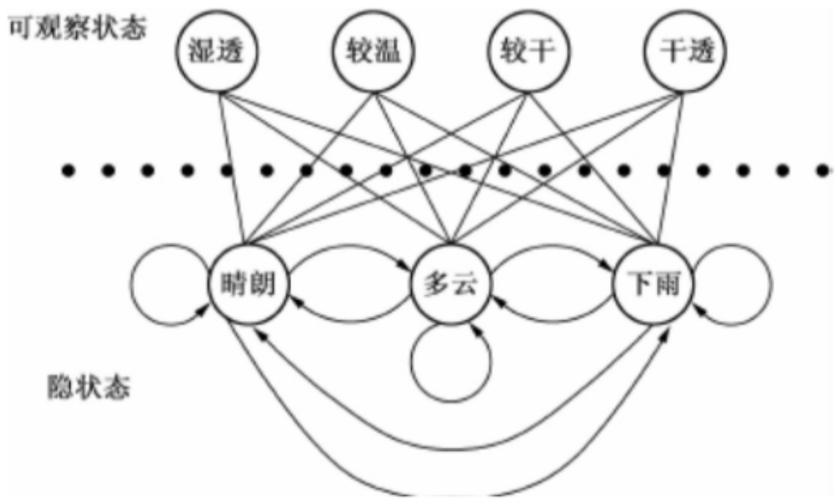
$S=\{S_1, S_2, \cdots, S_N\}$ 状态集合

$A=\{a_{ij}\}$ 状态转移概率矩阵

$\pi=\{\pi_1,\pi_2,\cdots\}$ 初始概率矩阵

隐马尔可夫模型（HMM）

这是一种具有隐藏状态的马尔可夫链，其核心特征是：状态不可直接观测，只能通过观测序列推断



形式定义：$\mu=(S,V,A,B,\pi)$ 简记为 $\mu=(A,B,\pi)$

S：状态集合 $\{S_1,\cdots,S_N\}$

V：观察值集合 $\{O_1,\cdots,O_M\}$

A：转移概率矩阵 $[a_{ij}]$

B：发射概率矩阵 $[b_j(k)]$

π：初始状态矩阵 $\pi=[\pi_i]$

其中B表示每个隐藏状态生成每个观测值的概率，$b_j[k]=P(O_t=O_k|q_t=S_j)$ 即状态为 S_j 时，生成观测 O_k 的概率

HMM使用场景：语音识别 同性标注 拼音输入法 文字识别 ...

HMM 三个基本问题：

评估：给定模型 $\mu=(A,B,\pi)$，计算某个观测序列 $O=O_1O_2\cdots O_T$ 的概率 $P(O|\mu)$

预测：给定观测序列 $O=O_1O_2\cdots O_T$，如何有效地确定一个状态序列 $Q=q_1q_2\cdots q_T$ 以便最好地解释观测序列

训练：给定观测序列 $O=O_1O_2\cdots O_T$，如何找到一个能够最好地解释这个观察序列模型，即如何调节参数 $\mu=(A,B,\pi)$ 使 $P(O|\mu)$ 最大化

评估

$$P(O)=\sum_Q P(O,Q) \Rightarrow P(O,Q)=P(Q)P(O|Q)$$

$$P(Q)=P(q_1q_2\cdots q_T)\approx P(q_1|q_0)P(q_2|q_1)\cdots P(q_T|q_{T-1})=\pi_{q_1}a_{q_1q_2}a_{q_2q_3}\cdots a_{q_{T-1}q_T}$$

转移概率

$$P(O|Q)=P(o_1o_2\cdots o_T|Q)\overset{\text{独立性假设}}{\approx}P(o_1|Q)P(o_2|Q)\cdots P(o_T|Q)=\prod_{t=1}^{T}P(o_t|Q)=\prod_{t=1}^{T}P(o_t|q_1q_2\cdots q_T)\approx\prod_{t=1}^{T}P(o_t|q_t)=b_{q_1}(o_1)\times b_{q_2}(o_2)\times\cdots\times b_{q_T}(o_T)$$

发射概率

$$\Rightarrow P(O,Q)=\prod_{t=1}^{T}[a_{q_{t-1}q_t}b_{q_t}(o_t)]$$

时间复杂度 $O(N^T)$

N 是状态数

T 是观测状态数

注意到一阶马尔可夫过程与动态规划思想一致，

我们也可以用动态规划来优化算法

t 时刻生成观测序列 $O=O_1O_2\cdots O_t$ 且达到状态 S_i（记为 $\alpha_t(i)$）的概率可由 t-1 时刻所有状态 $\alpha_{t-1}(j)$ 结合状态转移概率 a_{ji} 和观测概率 $b_i(O_t)$ 得到

$$\alpha_t(i) = \left(\sum_{j=1}^{N} \alpha_{t-1}(j)\, a_{ji}\right) \cdot b_i(O_t)$$

再累加最后一步所有状态便可知全局解

$$P(O|u) = \sum_{i=1}^{N} \alpha_T(i)$$

其中 $\alpha_t(i) = P(O_1 O_2 \cdots O_t, q_t = S_i)$ 表示从时刻 t 的第 i 个状态 S_i 结束的所有路径概率之和，称之为前向变量

- 前向算法：$O(N^2 T)$

  ① 初始化：$\alpha_1(i) = \pi_i b_i(O_1)$

  ② 递归计算：$\alpha_t(i) = \left(\sum_{j=1}^{N} \alpha_{t-1}(j)\, a_{ji}\right) b_i(O_t)$

  ③ 求和：$P(O|u) = \sum_{i=1}^{N} \alpha_T(i)$

- 后向算法：

  引入后向变量 $\beta_t(i)$，表示在时间 t 状态为 S_i 的条件下，输出后续观测序列 $O_{t+1} O_{t+2} \cdots O_T$ 的概率

  $$\beta_t(i) = P(O_{t+1} O_{t+2} \cdots O_T | q_t = S_i)$$

  即从时刻 t 的第 i 个状态开始的所有路径概率之和

递归关系 $\beta_t(i)=\sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$

a_{ij}：i转移 ；$b_j(O_{t+1})$：状态j被观测作为O_{t+1} ；$\beta_{t+1}(j)$：后向变量

① 初始化 $\beta_T(i)=1$

② 递归 $\beta_t(i)=\sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$

③ 求和 $P(O|\mu)=\sum_{i=1}^{N} \pi_i b_i(O_1) \beta_1(i)$

- 前后结合：
  - 前：$P(O)=\sum \alpha_T(i)$
  - 后：$P(O)=\sum \pi_i b_i(O_1) \beta_1(i)$

得 $P(O)=\sum \alpha_t(i) \beta_t(i)$

$$
\begin{aligned}
P(O) &= P(o_1 \cdots o_T, q_t = s_1) + P(o_1 \cdots o_T, q_t = s_2) + \cdots + P(o_1 \cdots o_T, q_t = s_N) \\
&= \sum_{i=1}^{N} P(o_1 \cdots o_T, q_t = s_i)
\end{aligned}
$$

$$
\begin{aligned}
& P(o_1 \cdots o_T, q_t = s_i) \\
&= P(o_1 \cdots o_t, q_t = s_i, o_{t+1} \cdots o_T) \\
&= P(o_1 \cdots o_t, q_t = s_i) \times P(o_{t+1} \cdots o_T \mid o_1 \cdots o_t, q_t = s_i) \\
&\approx P(o_1 \cdots o_t, q_t = s_i) \times P(o_{t+1} \cdots o_T \mid q_t = s_i) \\
&= \alpha_t(i) \beta_t(i)
\end{aligned}
$$

$$
\begin{aligned}
\alpha_t(i) &= P(o_1 o_2 \cdots o_t, q_t = s_i) \\
\beta_t(i) &= P(o_{t+1} o_{t+2} \cdots o_T \mid q_t = s_i)
\end{aligned}
$$

$$
P(O) = \sum_{i=1}^{N} \alpha_t(i) \beta_t(i), \quad 1 \le t \le T
$$

预测：给定模型(A, B, π)和观测序列O，求概率最大的隐藏状态序列Q

- 维特比变量：$\delta_t(i)$表示到t时刻、第i个状态s_i的节点的所有可能路径里概率最大的那

求路径的概率值

$$\delta_t(i) = \max_{q_1, q_2, \cdots, q_{t-1}} P(q_1, q_2, \cdots, q_{t-1}, q_t = S_i, o_1 o_2 \cdots o_t)$$

max：所有P里的最大值

$q_t = S_i$：t时刻状态为S_i　$o_1 \cdots o_t$：前t个观测序列

状态转移：

① 初始化：$\delta_1(i) = \pi_i b_i(o_1)$

② 递推：$\delta_t(i) = [\max_j(\delta_{t-1}(j) \times a_{ji})] \times b_i(o_t)$

到状态i的所有路径里概率最大的一条乘状态i被观测为o_t的概率

- 反向指针：$\psi_t(i)$表示到S_i的最优路径里S_i的前驱节点

$$\psi_t(i) = \arg\max_j[(\delta_{t-1}(j) \times a_{ji}) \times b_i(o_t)]$$

arg：取索引，也就是S_i的前驱S_j的j

注：其实$\psi_t(i)$值与$b_i(o_t)$无关，$\psi_t(i) = \arg\max_j(\delta_{t-1}(j) \times a_{ji})$

- Viterbi算法

① 初始化

$\delta_1(i) = \pi_i b_i(o_1)$

$\psi_1(i) = 0$

② 递归

$\delta_t(i) = [\max_j(\delta_{t-1}(j) a_{ji})] b_i(o_t)$

$\psi_t(i) = \arg\max_j[(\delta_{t-1}(j) a_{ji}) b_i(o_t)]$

③ 路径回溯

$q_T = \arg\max_i \delta_T(i)$

$$q_{t-1} = \psi_t(q_t)$$

训练：给一组观测序列 $O = O_1 O_2 \cdots O_T$，调节模型参数 $u = (A, B, \pi)$
使 $P(O|u)$ 最大化，即 $\arg\max_u P(O|u)$

情况1：状态序列 $Q = q_1 \cdots q_T$ 已知

方法：最大似然估计，直接用 O, Q 得到 $A、B、\pi$

$$\pi_i = \frac{\text{状态}i\text{作为初始状态次数}}{\text{总序列数}} = \delta(q_1, s_i)$$

$$a_{ij} = \frac{\text{状态}i\text{转移到}j\text{的次数}}{\text{状态}i\text{出现总次数}} = \frac{C(s_i s_j)}{C(s_j)}$$

$$= \frac{\sum_{t=1}^{T} \delta(q_t, s_i) \times \delta(q_{t+1}, s_j)}{\sum_{t=1}^{T} \delta(q_t, s_i)}$$

$$b_j(k) = \frac{\text{状态}j\text{生成观察状态}v_k\text{的次数}}{\text{状态}j\text{出现总次数}} = \frac{C(s_j v_k)}{C(s_j)}$$

$$= \frac{\sum_{t=1}^{T} \delta(q_t, s_j) \times \delta(O_t, v_k)}{\sum_{t=1}^{T} \delta(q_t, s_j)}$$

注：这里的 δ 是克罗内克函数 $\delta(a,b) = \begin{cases} 1 & a=b \\ 0 & a \neq b \end{cases}$
不是维特比变量

情况2：状态序列 Q 未知

同前无任何方法来求取全局最优解，但可以用 Baum-Welch 算法（前向后向算法）来求局部最优解，这种算法是 EM 算法的特例，下面介绍 EM 算法

EM算法：一种迭代优化算法，用于在有在隐变量的情况下进行参数估计。它通过交替执行两个步骤来优化参数：

E步：在当前参数下，计算隐变量（状态序列）的期望

即 Expectation

M步：通过E步的期望更新参数，以求最大化似然函数

即 Maximization.

Baum-Welch算法：① 初始化：随机给参数赋值，但要满足以下约束：

$$\sum_{i=1}^{N} \pi_i = 1 \qquad \sum_{j=1}^{N} a_{ij} = 1 \qquad \sum_{k=1}^{M} b_i(k) = 1$$

由此得到模型 u_0，令 $i=0$，执行②

② EM计算：

E：由 u_i 参数计算 $\gamma_t(i)$（时刻t处于状态i的概率）

$$\gamma_t(i) = \tilde{P}(q_t = S_i | O, u)$$

$$= \frac{\tilde{P}(q_t = S_i, O | u)}{\tilde{P}(O|u)}$$

$$= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}$$

$\tilde{P}(q_t = S_i | O, u)$：给定观测序列 O，模型 u，求时刻t处于状态 S_i 的概率

$\tilde{P}(q_t = S_i, O | u)$：在模型 u 下，时刻t处于状态 S_i 且观测序列为O的概率

$\tilde{P}(O|u)$：在模型 u 下观测到 O 的边缘概率

$\alpha_t(i)$：前向概率　　$\beta_t(i)$：后向概率

M：由 $\gamma_t(i)$ 估计 π_i，a_{ij}，$b_i(k)$，得到 u_{i+1}

$$\pi_i' = P(q_1 = S_i | O, u) = \gamma_1(i)$$

$$a'_{ij} = \frac{\sum_{t=1}^{T-1} \tilde{P}(q_t = s_i, q_{t+1} = s_j \mid O, u)}{\sum_{t=1}^{T-1} \tilde{P}(q_t = s_i \mid O, u)}$$

$$= \frac{\sum_{t=1}^{T-1} \frac{\tilde{P}(q_t = s_i, q_{t+1} = s_j, O \mid u)}{\tilde{P}(O \mid u)}}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$= \frac{\sum_{t=1}^{T-1} \alpha_t(i)\, a_{ij}\, b_j(O_{t+1})\, \beta_{t+1}(j)}{\sum_{j=1}^{N} \alpha_t(j)\, \beta_t(j)}$$

$$b'_i(k) = \frac{\sum_{t=1}^{T-1} \tilde{P}(q_t = s_i, O_t = v_k \mid O, u)}{\sum_{t=1}^{T-1} \tilde{P}(q_t = s_i \mid O, u)}$$

$$= \frac{\sum_{t=1}^{T} \gamma_t(i) \times \delta(O_t, v_k)}{\sum_{t=1}^{T} \gamma_t(i)}$$

③ 循环 i++，执行②，直到 π，A，B 收敛

A、B、π 的初始化：刚提到，算法开始时我们会对 A B π 随机赋值

而算法只能找到这个起点附近的局部最优解

要想得到全局最优解，起点就必须在其附近

在实践中我们发现，好的初始估计对 B 十分重要

而 A、π 随机赋值就够了