自然语言处理技术

第5章 隐马尔科夫模型

杨沐昀 孙承杰

哈工大教育部-微软语言语音重点实验室
MOE-MS Joint Key Lab of NLP and Speech (HIT)

主要内容

- 马尔科夫模型
- 隐马尔科夫模型
- 隐马尔科夫模型的应用

马尔科夫(Markov)模型：概述

- 马尔科夫模型是一种统计模型，广泛的应用在语音识别，词性自动标注，音字转换，概率文法等各个自然语言处理的应用领域。
- Markov(1856~1922)，苏联数学家。切比雪夫的学生。在概率论、数论、函数逼近论和微分方程等方面卓有成就。
- 经过长期发展，尤其是在语音识别中的成功应用，使它成为一种通用的统计工具。
- N元语言模型，是Markov模型的应用。

马尔科夫(Markov)模型：概述

- 随机过程又称为随机函数，是随时间随机变化的过程。马尔科夫模型描述了一类重要随机过程。
- 一个系统有N个有限状态$S = \{s_1, s_2, \dots, s_N\}$，随时间推移，系统将由某一状态转移到另一状态。
- $Q = (q_1, q_2, \dots, q_T)$为随机变量序列，其取值为状态集S中的某个状态，在时间t的状态为q_t。

马尔科夫(Markov)模型：概述

- 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，该概率为：

$$P(q_t = s_j | q_{t-1} = s_i, q_{t-2} = s_k, \ldots)$$

- 离散的一阶马尔科夫链：系统在时间t的状态只与时间t-1的状态有关。

$$P(q_t = s_j | q_{t-1} = s_i, q_{t-2} = s_k, \ldots) = P(q_t = s_j | q_{t-1} = s_i)$$

马尔科夫(Markov)模型：概述

- 马尔科夫模型：只考虑独立于时间t的随机过程

$$P\left(q_t = s_j \middle| q_{t-1} = s_i\right) = a_{ij}, 1 \le i.j \le N$$

- 状态转移概率a_{ij}必须满足以下条件：
  - $a_{ij} \ge 0$
  - $\sum_{j=1}^{N} a_{ij} = 1$
- N个状态的一阶马尔科夫过程有N^2，可以表示成为一个状态转移矩阵。

马尔科夫(Markov)模型：举例

- 一段文字中名词，动词，形容词三类词性出现的情况可以由三个状态的马尔科夫模型描述：
- 状态s_1：名词
- 状态s_2：动词
- 状态s_3：形容词

马尔科夫(Markov)模型：举例

- 假设状态之间的转移矩阵如下：

$$A = [a_{ij}] = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} s_1 & s_2 & s_3 \end{array} \\ \begin{bmatrix} 0.3 & 0.5 & 0.2 \\ 0.5 & 0.3 & 0.2 \\ 0.4 & 0.2 & 0.4 \end{bmatrix} \end{array}$$

- 如果在该文字中某句子的第一个词为名词，那么该句子中三类词出现顺序为O=“名动形名”的概率。

马尔科夫(Markov)模型：举例

- $P(O|M) = P(s_1, s_2, s_3, s_1|M)$

$= \mathrm{P}(s_1) \cdot \mathrm{P}(s_2|s_1) \cdot \mathrm{P}(s_3|s_2) \cdot \mathrm{P}(s_1|s_3)$

$= 1 \times a_{12} \times a_{23} \times a_{31}$

$= 0.5 \times 0.2 \times 0.4$

$= 0.04$

$$A = [a_{ij}] = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \end{array} \begin{array}{c} \begin{array}{ccc} s_1 & s_2 & s_3 \end{array} \\ \begin{bmatrix} 0.3 & 0.5 & 0.2 \\ 0.5 & 0.3 & 0.2 \\ 0.4 & 0.2 & 0.4 \end{bmatrix} \end{array}$$

- 系统初始化式时可以定义一个初始状态概率向量

$$\pi_i \geq 0, \sum_{i=1}^{N} \pi_i = 1$$

马尔科夫(Markov)模型：有限状态机

- 马尔科夫模型可视为随机的有限状态机。
- 圆圈表示状态，状态之间的转移用带箭头的弧表示，弧上的数字为状态转移的概率。
- 初始状态用标记为start的输入箭头表示。
- 假设任何状态都可作为终止状态。
- 对每个状态来说，发出弧上的概率和为1。

马尔科夫(Markov)模型：有限状态机

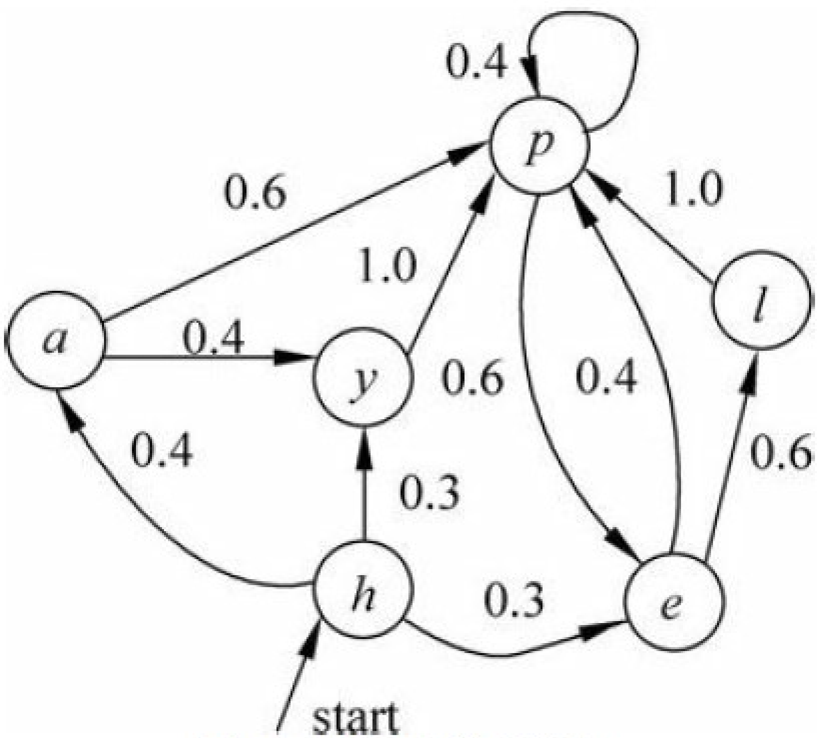


图6-4 马尔可夫模型的例子

$$
\begin{aligned}
&P(h,e,l,p) \\
&= P(h) \times P(e|h) \times P(l|e) \times P(p|l) \\
&= 1.0 \times 0.3 \times 0.6 \times 1.0 \\
&= 0.18
\end{aligned}
$$

隐马尔可夫模型：概述

- 隐马尔可夫模型创建于20世纪70年代，是美国数学家鲍姆等人提出来的。
- 该模型是一个双重随机过程，我们不知道具体的状态序列，只知道状态转移的概率，即模型的状态转换过程是不可观察的（隐蔽的），可观察事件的随机过程是隐蔽状态过程的随机函数。

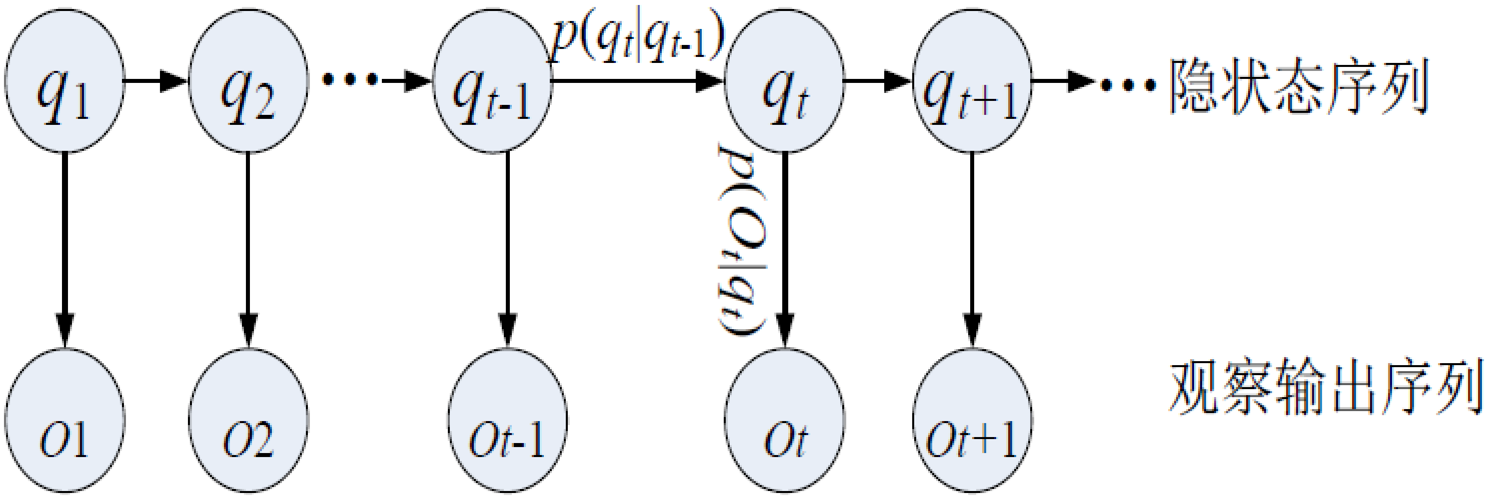
隐马尔可夫模型：例子

- 假定一暗室中有N个口袋，每个口袋中有M种不同颜色的球。
- 一个实验员根据某一概率分布随机地选取一个初始口袋，从中根据不同颜色的球的概率分布，随机地取出一个球，并向室外的人报告该球的颜色。
- 再根据口袋的概率分布选择另一个口袋，根据不同颜色的球的概率分布从中随机选择另外一个球。重复进行这个过程。

隐马尔可夫模型：例子

- 对于暗室外边的人来说，可观察的过程只是不同颜色的球的序列，而口袋的序列是不可观察的。
- 每个口袋对应于HMM中的状态，球的颜色对应于HMM中状态的输出符号。
- 从一个口袋转向另一个口袋对应于状态转换，从口袋中取出球的颜色对应于从一个状态输出的观察符号。

隐马尔可夫模型：图解



HMM 图解

隐马尔可夫模型：组成部分

1.模型中状态的数目N（上例中口袋的数目）；

2.从每个状态可能输出的不同符号的数目M（上例中球的不同颜色的数目）；

3.状态转移概率矩阵$A=\{a_{ij}\}$（a_{ij}为实验员从一个口袋（状态s_i）转向另一个口袋（s_j）取球的概率)。其中：

$$a_{ij}=P(q_t=s_j|q_{t-1}=s_i), 1\le i,j\le N$$

$$a_{ij}\ge 0$$

$$\sum_{j=1}^{N} a_{ij}=1$$

隐马尔可夫模型：组成部分

4.从状态s_j观察到符号v_k的概率分布矩阵$B = \{b_j(k)\}$

（ $b_j(k)$ 为实验员从第j个口袋中取出第k种颜色的球的概率），其中：

$$b_j(k) = P\left(O_t = v_k \middle| q_t = s_j\right), 1 \le j \le N; 1 \le k \le M$$

$$b_j(k) \ge 0$$

$$\sum_{k=1}^{M} b_j(k) = 1$$

隐马尔可夫模型：组成部分

5.初始状态概率分布$\pi=\{\pi_i\}$，其中：

$$\pi_i = P(q_1 = s_i), 1 \le i \le N$$

$$\pi_i \ge 0$$

$$\sum_{i=1}^{N} \pi_i = 1$$

隐马尔可夫模型：组成部分

- 一般地，一个HMM记为一个五元组μ＝（S，K，A，B，π），其中，S为状态的集合，K为输出符号的集合，π，A和B分别是初始状态的概率分布、状态转移概率和符号发射概率。
- 为了简单，有时也将其记为三元组μ＝（A，B，π）

隐马尔可夫模型：三个基本问题

1.估值问题：给定一个观察序列 $O = O_1 O_2 \ldots O_T$ 和模型μ = (A，B，π)，如何快速地计算出给定模型μ情况下，观察序列O的概率，即$P(O|\mu)$?

2.序列问题：给定一个观察序列 $O = O_1 O_2 \ldots O_T$ 和模型μ = (A，B，π),如何快速有效的选择在一定意义下“最优”的状态序列 $Q = q_1 q_2 \ldots q_T$，使得该状态序列“最好的解释”观察序列?

3.参数估计问题：给定一个观察序列$O = O_1 O_2 \ldots O_T$，如何根据最大似然估计来求模型的参数值？即如何调节模型μ = (A，B，π)的参数，使得$P(O|\mu)$最大?

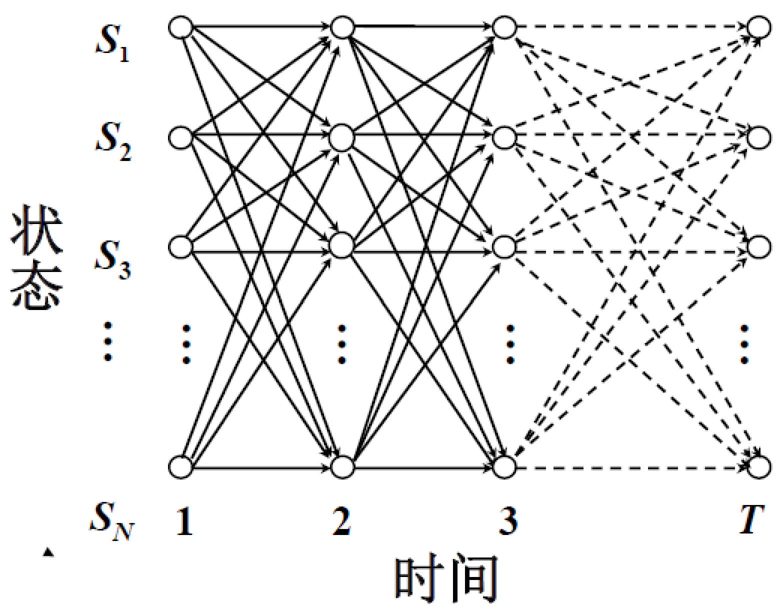
隐马尔可夫模型：求解观察序列的概率

- 给定观察序列$\mathrm{O} = O_1 O_2 \ldots O_T$和模型$\mu = (A, B, \pi)$，快速的计算出给定模型$\mu$情况下观察序列O的概率，即$P(O|\mu)$。

- 对于给定的状态序列$Q = q_1 q_2 \ldots q_T$, $P(O|\mu) =?$
- $p(O|\mu) = \sum_Q p(O, Q|\mu) = \sum_Q p(Q|\mu) \cdot p(O|Q, \mu)$
- $p(Q|\mu) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \ldots a_{q_{t-1} q_T}$
- $p(O|Q, \mu) = b_{q_1}(O_1) b_{q_2}(O_2) \ldots b_{q_T}(O_T)$

隐马尔可夫模型：求解观察序列的概率

- 存在的困难：如果模型μ有N个不同的状态，时间长度为T，那么有N^T个可能的状态序列，搜索路径成指数级组合爆炸。



隐马尔可夫模型：前向算法

- 解决办法：动态规划，前向算法。
- 基本思想：定义前向变量$\alpha_t(i)$，前向变量$\alpha_t(i)$是在时间t，HMM输出了序列$O_1O_2 \dots O_t$，并且位于状态s_i的概率。
- $\alpha_t(i) = P(O_1O_2 \dots O_t, q_t = s_i|\mu)$
- 如果可以高效的计算$\alpha_t(i)$，就可以高效的求得$p(O|\mu)$。

隐马尔可夫模型：前向算法

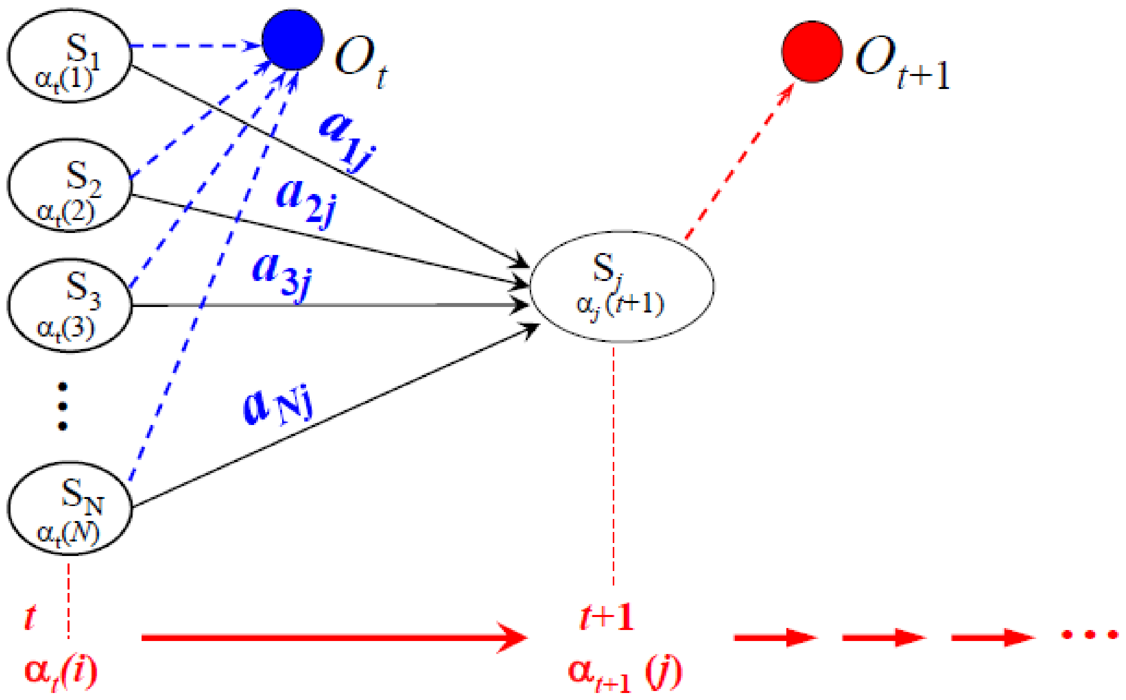
- $p(O|\mu)$是在到达状态q_T时观察到序列$O = O_1O_2 \dots O_T$的概率：

$$p(O|\mu) = \sum_{s_i} p(O_1O_2 \dots O_T, q_T = S_i|\mu) = \sum_{i=1}^{N} \alpha_T(i)$$

- 在时间t+1的前向变量可以根据在时间t时的前向变量$\alpha_t(1)$，$\alpha_t(2), \dots, \alpha_t$ (N)的值来归纳计算

$$\alpha_{t+1}(j) = (\sum_{i=1}^{N} \alpha_t(i)\, a_{ij})b_j(O_{t+1})$$

隐马尔可夫模型：前向算法



$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{N} \alpha_t(i)\, a_{ij}\right) b_j(O_{t+1})$$

隐马尔可夫模型：前向算法

- 前向算法

1.初始化：$\alpha_1(i) = \pi_i b_i(O_1), 1 \le i \le N$。

2.归纳计算

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{N} \alpha_t(i)\, a_{ij}\right) b_j(O_{t+1}), 1 \le t \le T-1$$

3.求和终结

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

隐马尔可夫模型：前向算法

- 时间复杂度：
  - 每计算一个$\alpha_t(i)$必须考虑从t-1时的所有N个状态转移到状态s_i的可能性，时间复杂度为O(N)，对应每一个时刻t，要计算N个前向变量：$\alpha_{\mathrm{t}}(1)$, $\alpha_{\mathrm{t}}(2)$,…$\alpha_{\mathrm{t}}(N)$，所以，时间复杂度为：O(N)XN=O(N^2),
  - 又因为t=1,2…T,所以前向算法总的复杂度为O(N^2T)

隐马尔可夫模型：后向算法

- 后向变量$\beta_t(i)$是在给定模型$\mu = (A, B, \pi)$，并且在时间t状态为s_i的条件下，HMM输出观察序列$O_{t+1} \dots O_T$的概率。

- $\beta_t(i) = \mathrm{P}(O_{t+1} \dots O_T | q_t = s_i, \mu)$

隐马尔可夫模型：后向算法

- 与计算前向变量一样，可以用动态规划的算法计算后向变量。

1. 从时刻t到t+1，模型由状态S_i转移到状态S_j，并从S_j输出O_{t+1}
2. 在时间t+1，状态为S_j的条件下，模型输出观察序列$O_{t+2}O_{t+3}\ldots O_T$

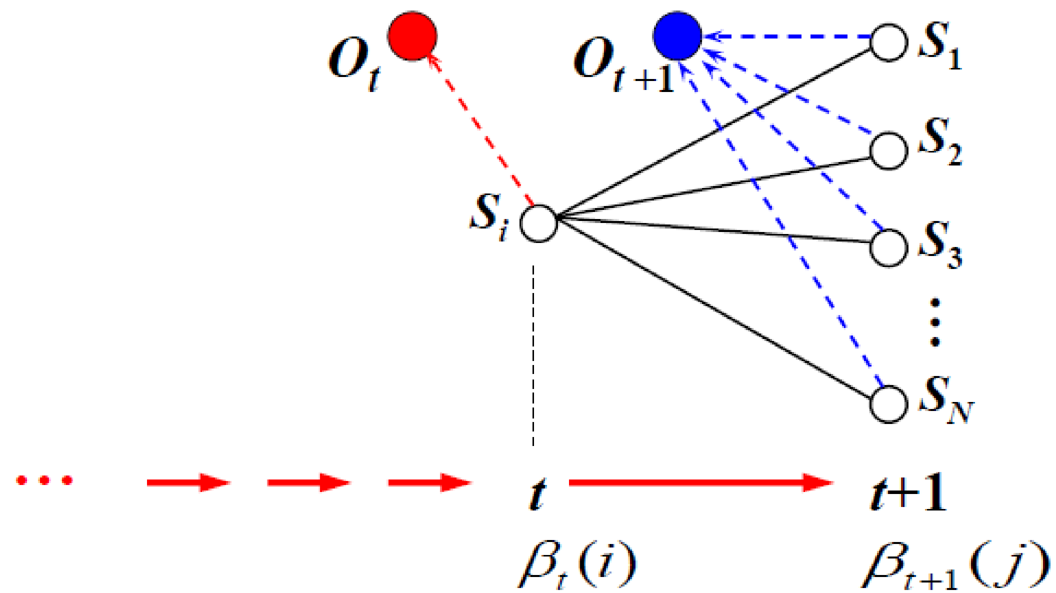
隐马尔可夫模型：后向算法

- 第一步的概率：$\mathrm{a}_{ij} \times b_j(O_{t+1})$
- 第二步的概率按后向变量的定义为$\beta_{t+1}(i)$
- 可得到如下归纳关系：

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

- 归纳顺序为：$\beta_{\mathrm{T}}(x), \beta_{\mathrm{T}-1}(x), \dots, \beta_1(x)$

隐马尔可夫模型：后向算法



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

隐马尔可夫模型：后向算法

1.初始化：$\beta_T(i) = 1, 1 \leq i \leq N$

2.归纳计算：

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

3.求和终结：

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

时间复杂度：$\mathrm{O}(N^2T)$

隐马尔可夫模型：维特比算法

- 维特比算法用于求解HMM中的第二个问题，给定一个观察序列$\mathrm{O} = O_1 O_2 \dots O_T$和模型$\mu = (A, B, \pi)$，如何快速有效的选择在一定意义下最优的状态序列$Q = q_1 q_2 \dots q_T$，使得该状态序列“最好的解释”观察序列。

- 对于最优状态序列的一种理解：状态序列中的每个状态都单独的具有概率，对于每个时刻$t (1 \le t \le T)$，寻找q_t使得$\gamma_t(i) = p(q_t = S_i | O, \mu)$最大。

隐马尔可夫模型：维特比算法

$$\gamma_t(i) = p(q_t = S_i | O, \mu) = \frac{p(q_t = S_i, O | \mu)}{p(O | \mu)}$$

- $p(q_t = S_i, O|\mu)$表示模型的输出序列O，并在时间t到达状态i的概率。

隐马尔可夫模型：维特比算法

- 分解过程：
  - 模型在时间t到达状态i，并且输出$O = O_1 O_2 \dots O_t$。根据前向变量的定义，实现这一步的概率为$\alpha_t(i)$。
  - 从时间t，状态S_i出发，模型输出$O = O_{t+1} O_{t+2} \dots O_T$，根据后向变量定义，实现这一步的概率为$\beta_t(i)$。
  - 因此：

$$p(q_t = S_i, O|\mu) = \alpha_t(i) \times \beta_t(i)$$

隐马尔可夫模型：维特比算法

- 而$p(O|\mu)$与时间t的状态无关，因此：

$$p(O|\mu) = \sum_{i=1}^{N} \alpha_t(i) \times \beta_t(i)$$

- 因此：$\gamma_t(i) = \frac{\alpha_t(i)\times\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\times\beta_t(i)}$

- t时刻的最优状态为：

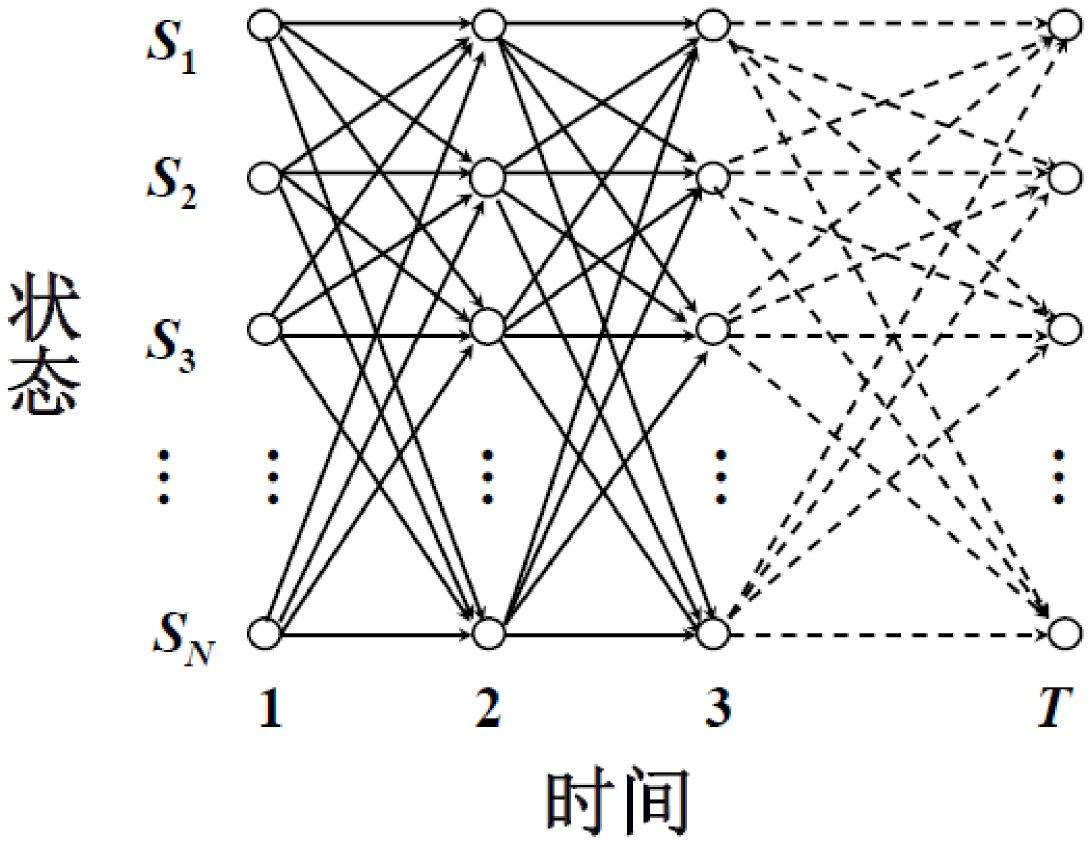
$$\hat{q}_t = \arg\max_{1\le i\le N}(\gamma_t(i))$$

隐马尔可夫模型：维特比算法

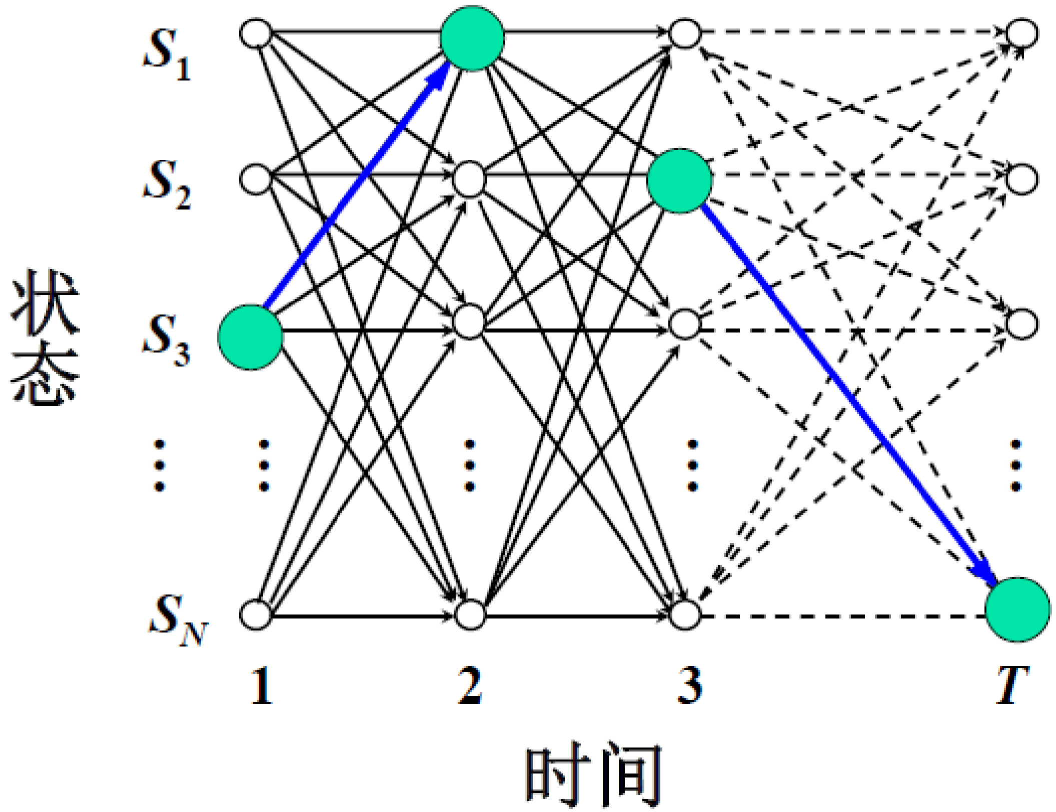
- 存在问题：

- 每一个状态单独最优不一定整体的状态序列最优，可能两个最优的状态 $\widehat{q}_t$ 和 $\widehat{q}_{t+1}$ 之间的转移概率为0.

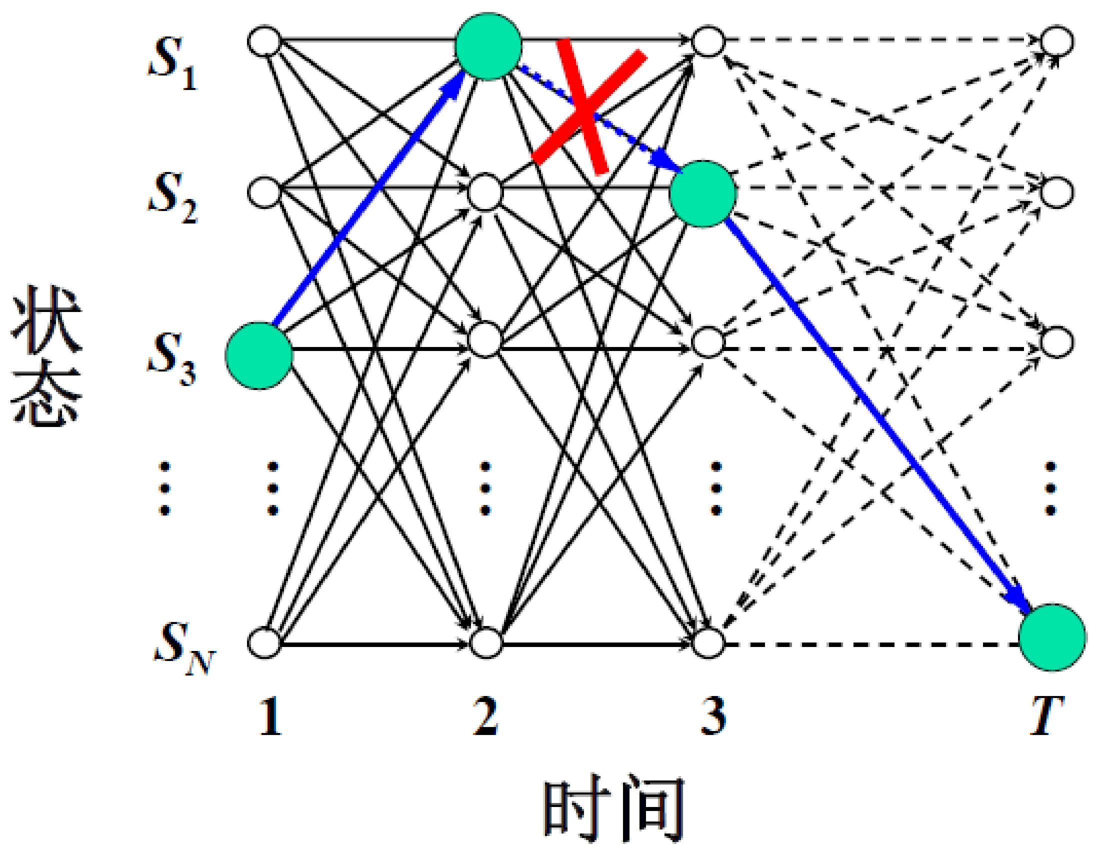
隐马尔可夫模型：维特比算法



隐马尔可夫模型：维特比算法



隐马尔可夫模型：维特比算法



隐马尔可夫模型：维特比算法

- 对于最优的另一种解释：在给定模型μ和观察序列O的条件下，使得$P(Q|O,\mu)$最大。

$$Q' = argmax_Q P(Q|O,\mu)$$

- 维特比算法运用动态规划的搜索算法求解最优状态序列。
- 定义一个维特比变量$\delta_t(i)$
  - $\delta_t(i)$是在时间t时，HMM沿着某一条路径到达状态s_i，并输出观察序列$O_1 O_2 \ldots O_t$的最大概率。

隐马尔可夫模型：维特比算法

- $\delta_t(i) = max_{q_1,q_2\ldots q_{t-1}} P(q_1, q_2 \ldots q_t = s_i, O_1 O_2 \ldots O_t|\mu)$

- 与前向变量类似，$\delta_t(i)$有如下递归关系：

$$\delta_{t+1}(i) = max_j[\delta_t(j) \cdot a_{ji}] \cdot b_i(O_{t+1})$$

- 维特比算法设置了变量$\varphi_t(i)$用来记录最优路径上状态s_i的前一个（在时间$t-1$的）状态

隐马尔可夫模型：维特比算法

- 步1 初始化:

$$\delta_1(i) = \pi_i b_i(O_1)\ , \qquad 1 \le i \le N$$
$$\varphi_1(i) = 0$$

- 步2 归纳计算:

$$\delta_t(j) = max_{1\le i\le N}[\delta_{t-1}(i)\cdot a_{ji}]\cdot b_j(O_t)\ , \quad 2 \le t \le T; 1 \le j \le N$$

记忆回退路径:

$$\varphi_t(j) = \mathrm{arg}max_{1\le i\le N}[\delta_{t-1}(i)\cdot a_{ji}]\cdot b_j(O_t)\ ,\ 2 \le t \le T; 1 \le j \le N$$

- 步3 终结:

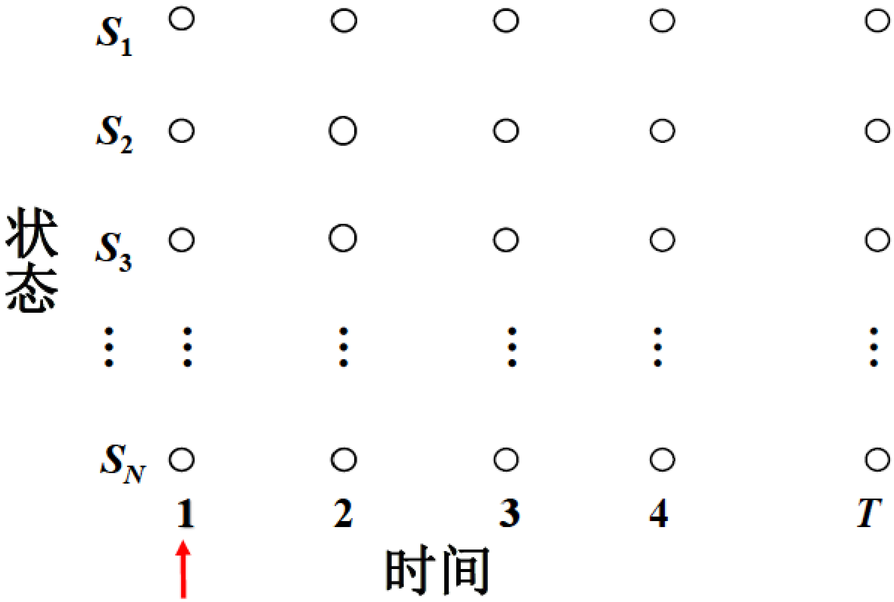
$$\widehat{\mathrm{Q}}_T = \underset{1\le i\le N}{\operatorname{argmax}}\, \delta_{\mathrm{T}}(i)$$
$$\hat{P}(\widehat{\mathrm{Q}}_T) = \max_{1\le i\le N} \delta_{\mathrm{T}}(i)$$

- 步4 路径（状态序列回溯）

$$\hat{q}_t = \varphi_{t+1}(\hat{q}_{t+1})\ , \qquad t = T-1, \mathrm{T}-2, \cdots, 1$$

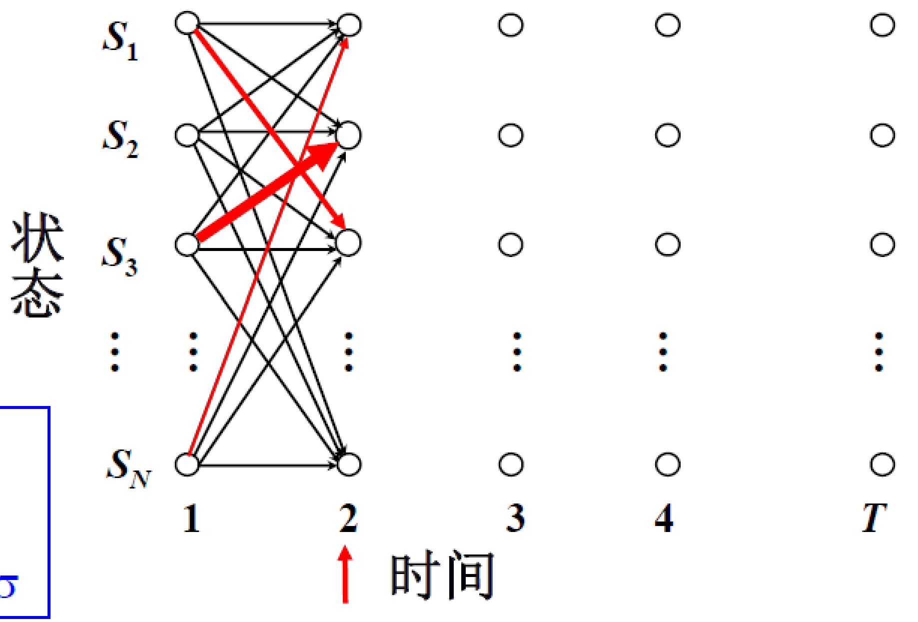
隐马尔可夫模型：维特比算法

图解
Viterbi
搜索
过程



隐马尔可夫模型：维特比算法

图解
Viterbi
搜索
过程



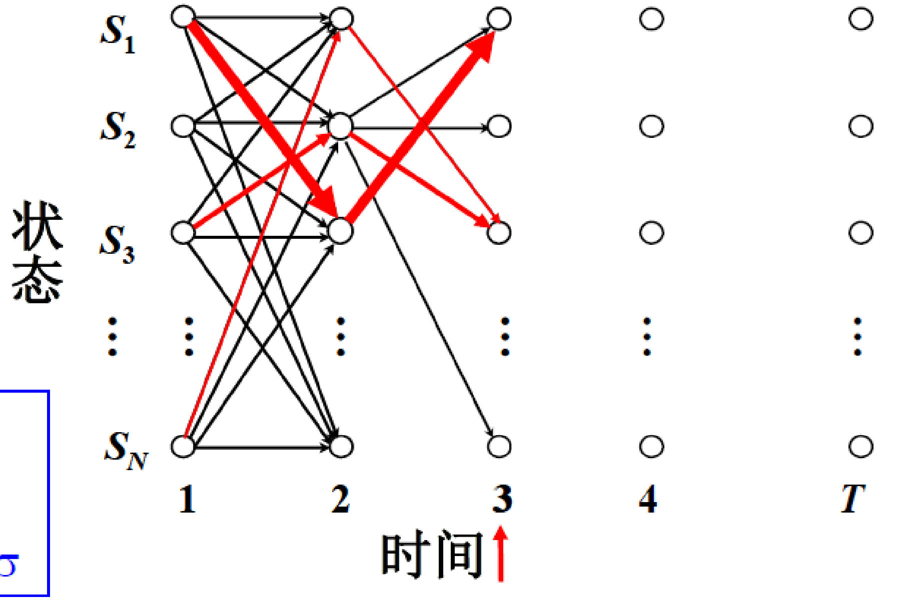
剪枝策略：

❶ $\delta_t(j) \geq \Delta$

❷ $\text{NPath} \leq \sigma$

隐马尔可夫模型：维特比算法

图解
Viterbi
搜索
过程



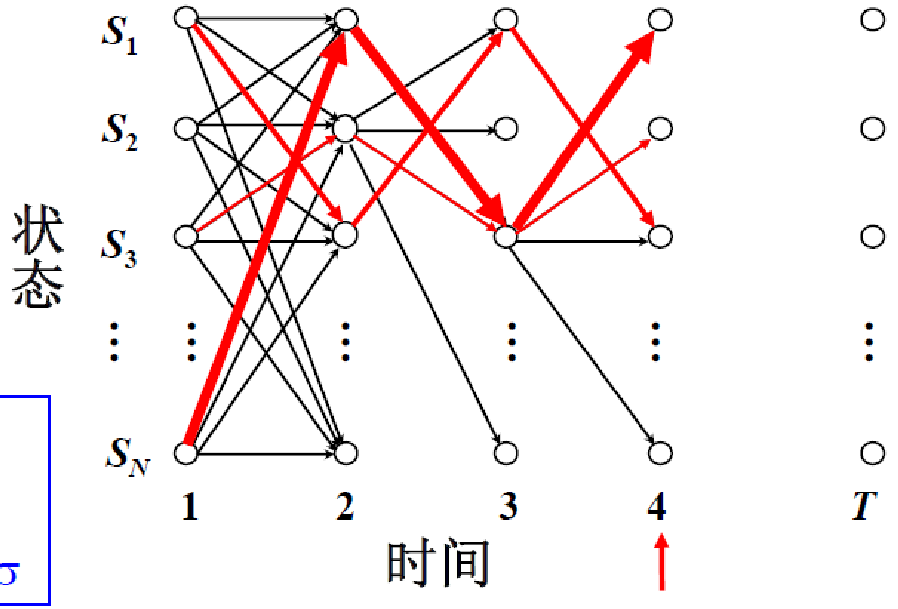
剪枝策略：

❶ $\delta_t(j) \geq \Delta$

❷ $\mathrm{NPath} \leq \sigma$

隐马尔可夫模型：维特比算法

图解
Viterbi
搜索
过程



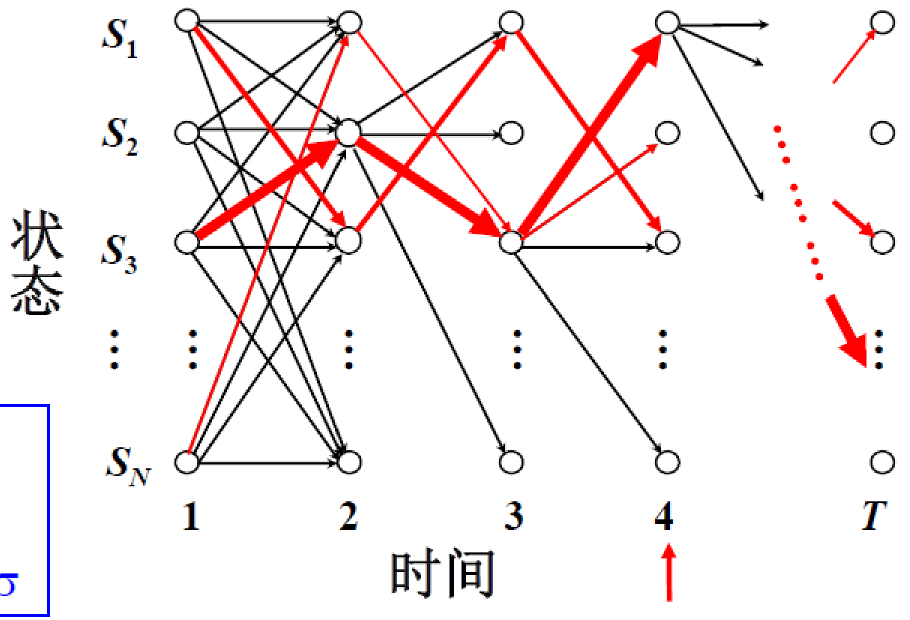
剪枝策略：

❶ $\delta_t(j) \geq \Delta$

❷ NPath $\leq \sigma$

隐马尔可夫模型：维特比算法

图解
Viterbi
搜索
过程



剪枝策略：

❶ $\delta_t(j) \geq \Delta$

❷ $NPath \leq \sigma$

隐马尔可夫模型：参数估计

- 参数估计是HMM面临的第三个问题，给定观察序列 $\mathrm{O} = O_1 O_2 \dots O_T$，如何调节模型$\mu = (A, B, \pi)$的参数，使得$P(O|\mu)$最大。

$$argmax_{\mu} P(O_{training}|\mu)$$

- 模型的参数是指构成μ的$\pi_i, a_{ij}, b_j(k)$。

隐马尔可夫模型：参数估计

- 如果产生观察序列O的状态$Q = q_1 q_2 \ldots q_T$已知，可以用最大似然估计来计算μ的参数：

$$\pi_i' = \delta(q_1, S_i)$$

$$a_{ij}' = \frac{Q\text{中从状态}q_i\text{转移到}q_j\text{的次数}}{Q\text{中所有从状态}q_i\text{转移到另一状态（包括}q_j\text{自身）的总数}} = \frac{\sum_{t=1}^{T-1} \delta(q_t, S_i) \times \delta(q_{t+1}, S_j)}{\sum_{t=1}^{T-1} \delta(q_t, S_i)}$$

- 其中，$\delta(x, y)$为克罗奈克(kronecker)函数，当x=y时，$\delta(x, y)$=1，否则$\delta(x, y)$=0

隐马尔可夫模型：参数估计

- 类似的
- $$b_j'(k) = \frac{\text{Q中从状态}q_j\text{输出符号}v_k\text{的次数}}{\text{Q到达}q_j\text{的总次数}}$$

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

- 其中，v_k是模型输出符号集中的第k个符号。

隐马尔可夫模型：参数估计

- 期望值最大化算法（EM）
  - 初始化时随机的给模型的参数赋值，遵循限制规则，例如：从某一状态出发的转移概率总和为1，得到模型μ_0，然后可以从μ_0得到从某一状态转移到另一状态的期望次数，然后以期望次数代替公式中的次数，得到模型参数的新估计，由此得到新的模型μ_1，从μ_1又可以得到模型中隐变量的期望值，由此重新估计模型参数。循环这个过程，参数收敛于最大似然估计。

隐马尔可夫模型：参数估计

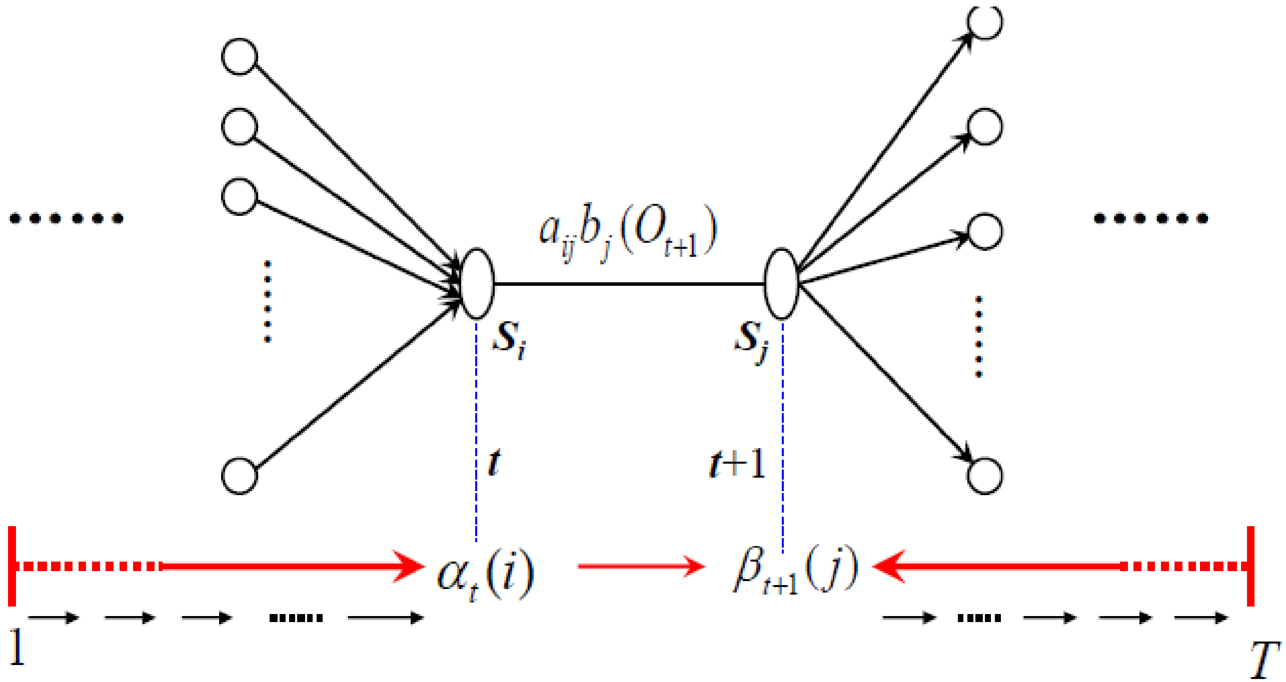
- 给定模型μ和观察序列$O = O_1 O_2 \ldots O_T$，在时间t位于状态S_i，时间t+1位于状态S_j的概率：
- $\xi_t(i,j) = p\left(q_t = S_i, q_{t+1} = S_j \middle| O, \mu\right)$

$$= \frac{p(q_t=S_i, q_{t+1}=S_j, O|\mu)}{p(O|\mu)}$$

$$= \frac{\alpha_t(i)\times a_{ij}b_j(O_{t+1})\times\beta_{t+1}(j)}{p(O|\mu)}$$

$$= \frac{\alpha_t(i)\times a_{ij}b_j(O_{t+1})\times\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)\times a_{ij}b_j(O_{t+1})\times\beta_{t+1}(j)} \qquad (6\text{-}24)$$

隐马尔可夫模型：参数估计



$$\alpha_t(i) \times a_{ij}b_j(O_{t+1}) \times \beta_{t+1}(j)$$

隐马尔可夫模型：参数估计

- 因此，给定模型μ和观察序列$O = O_1 O_2 \ldots O_T$，在时间t位于状态S_i的概率为：

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \quad (6\text{-}25)$$

- 因此，模型μ的参数可由下面的公式重新估计：

1.q_1为S_i的概率：

$$\pi_i = \gamma_1(i) \quad (6\text{-}26)$$

隐马尔可夫模型：参数估计

2.

$$a'_{ij}=\frac{Q\text{中从状态}q_i\text{转移到}q_j\text{的期望次数}}{Q\text{中所有从状态}q_i\text{转移到另一状态（包括}q_j\text{自身）的期望总数}}$$

$$=\frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \qquad (6\text{-}27)$$

隐马尔可夫模型：参数估计

3.

$$b_j'(k) = \frac{\text{Q中从状态}q_j\text{输出符号}v_k\text{的期望次数}}{\text{Q到达}q_j\text{的期望次数}}$$

$$= \frac{\sum_{t=1}^{T} \gamma_t(j) \times \delta(O_t, v_k)}{\sum_{t=1}^{T} \gamma_t(j)} \qquad (6\text{-}28)$$

隐马尔可夫模型：前向后向算法

步1　初始化：随机地给参数π_i，a_{ij}，b_j（k）赋值，使其满足如下约束：

$$\sum_{i=1}^{N} \pi_i = 1$$

$$\sum_{j=1}^{N} a_{ij} = 1, \quad 1 \leqslant i \leqslant N$$

$$\sum_{k=1}^{M} b_j(k) = 1, \quad 1 \leqslant j \leqslant N$$

由此得到模型μ_0。令i＝0，执行下面的EM估计。

步2　EM计算：

E-步骤： 由模型μ_i根据式（6-24）和式（6-25）计算期望值ξ_t（i, j）和γ_t（i）；

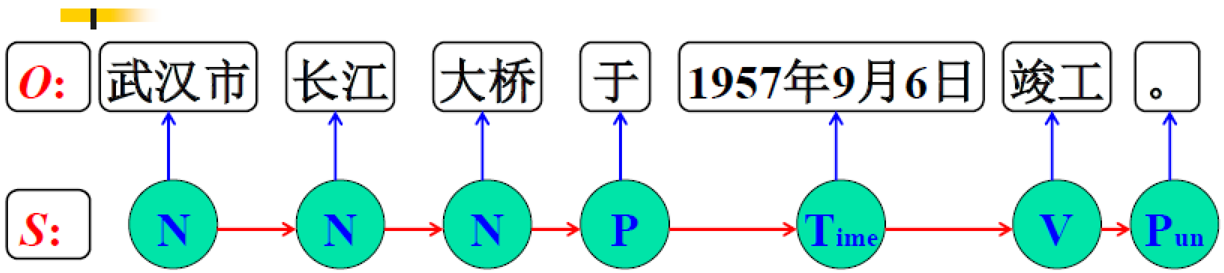
M-步骤： 用E-步骤得到的期望值，根据式（6-26）、（6-27）和（6-28）重新估计参数π_i，a_{ij}，b_j（k）的值，得到模型μ_{i+1}。

步3　循环计算：

令i＝i＋1。重复执行EM计算，直到π_i，a_{ij}，b_j（k）收敛。

隐马尔可夫模型：应用举例

- 词性标注问题。
- 例如：武汉市长江大桥于1957年9月6日竣工。



隐马尔可夫模型：应用举例

- 用HMM解决问题必须考虑的几个问题：
- 1. 如何确定状态、观察及各自的数目？
- 2. 参数估计：初始状态概率、状态转移概率、输出概率如何确定？

隐马尔可夫模型：应用举例

- 对于汉语分词：如果将汉语分词的结果作为观察序列

$\mathrm{O} = O_1 O_2 \ldots O_T$，那么则需求解$O' = argmax_O P(O|\mu)$。

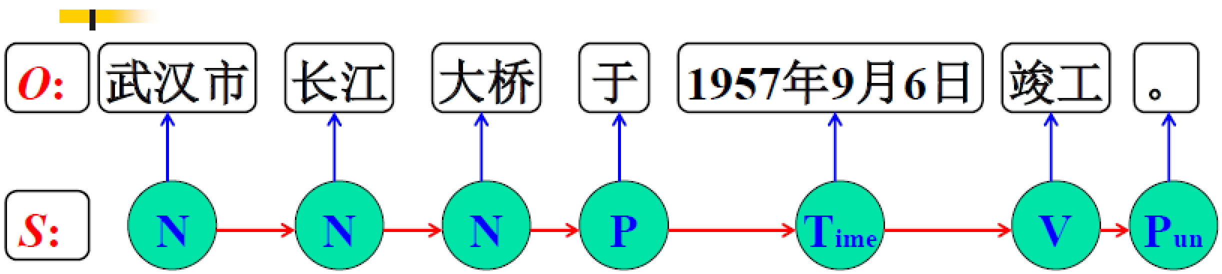
- 对于词性标注问题：则需要求解的是：

$$Q' = argmax_Q P(Q|O,\mu)。$$

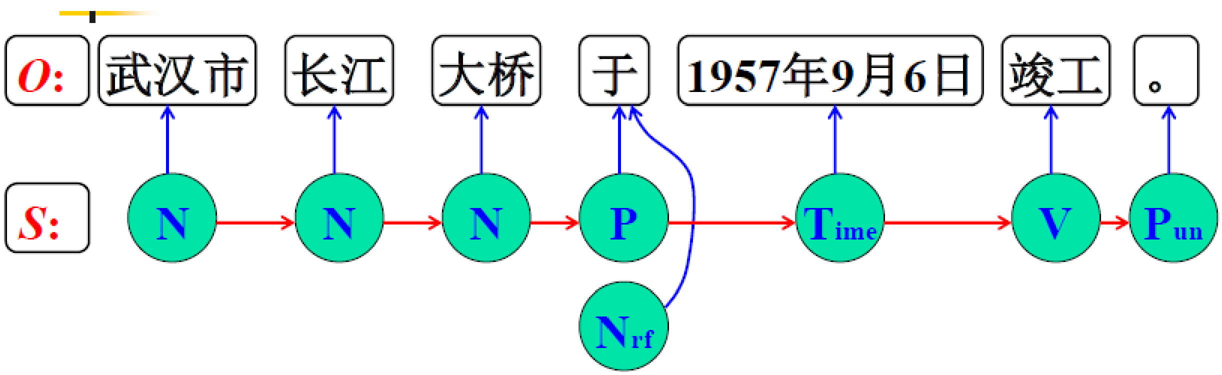
隐马尔可夫模型：应用举例

- 进一步解释：
  - 估计HMM模型$\mu = (A, B, \pi)$的参数。
  - 对于任意给定的一个输入句子及其可能的输出序列O，求所有可能的O中使概率$p(O|\mu)$最大的解。
  - 快速地选择“最优”的状态序列(词性序列)，使其最好地解释观察序列。

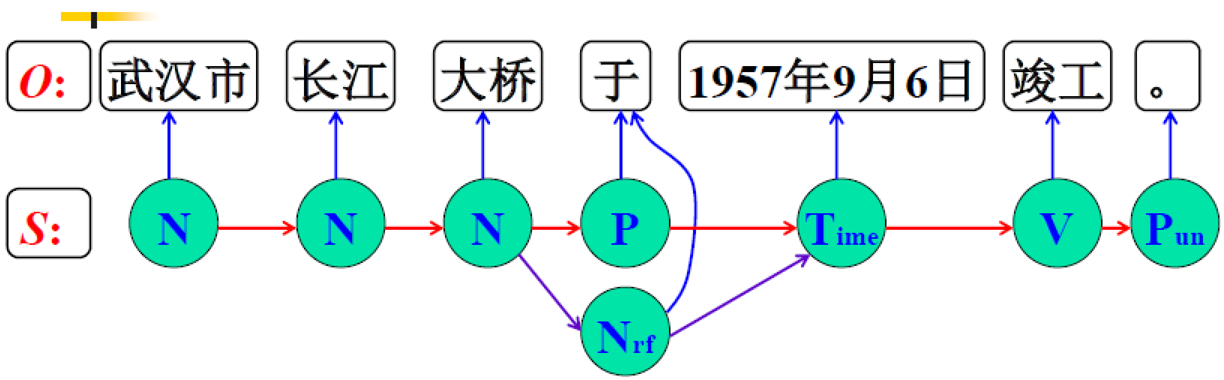
隐马尔可夫模型：应用举例



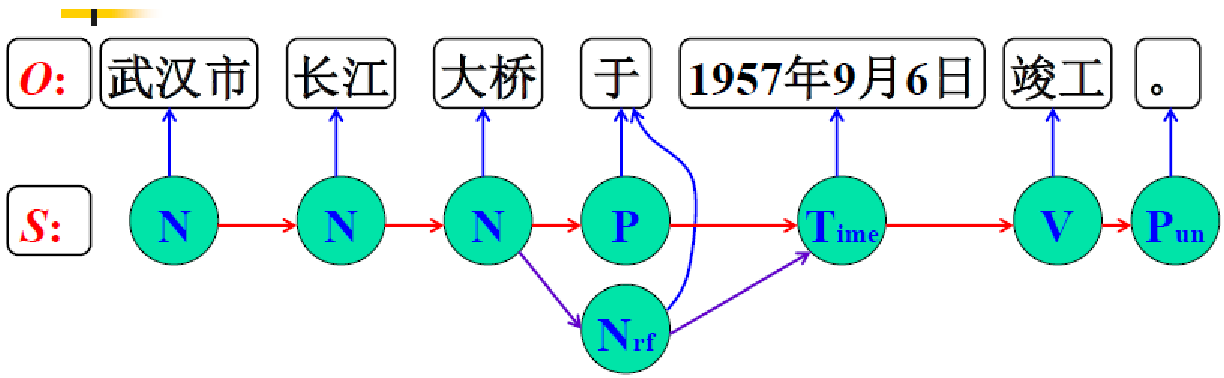
隐马尔可夫模型：应用举例



隐马尔可夫模型：应用举例



隐马尔科夫模型：应用举例



a.武汉市/N 长江/N 大桥/N 于/P 1957年9月6日/Time 竣工/V 。/Pun

b.武汉市/N 长江/N 大桥/N 于/N_{rf} 1957年9月6日/Time 竣工/V 。/Pun

隐马尔科夫模型：应用举例

- 问题1：模型参数
  - 观察序列：单词序列。
  - 状态序列：词类标记序列。
  - 状态数目N：为词类标记符号的个数，如Upenn LDC汉语树库中有33个词类，北大语料库词类标记符号106个等。
  - 输出符号数M：每个状态可输出的不同词汇个数，如汉语介词P约有60个，连词C约有110个，即状态P和C对应的输出符号数为60、110.

隐马尔可夫模型：应用举例

- 参数估计：
- 如果无任何标注语料：需要一部有词性标注的词典，采用无指导学习方法：
  - 获取词类个数（状态数）
  - 获取对应每种词类的词汇数（输出符号数）
  - 利用EM迭代算法获取初始状态概率、状态转移概率和输出符号概率。

隐马尔可夫模型：应用举例

- 若有大规模分词和词性标注语料：有指导学习方法。

咱们/rr 中国/ns 这么/rz 大{da4}/a 的{de5}/ud 一个/mq 多/a 民族/n 的{de5}/ud 国家/n 如果/c 不/df 团结/a ，/wd 就/d 不/df 可能/vu 发展/v 经济/n ，/wd 人民/n 生活/n 水平/n 也/d 就/d 不/df 可能/vu 得到/v 改善/vn 和{he2}/c 提高/vn 。/wj

- 可以从这些标注语料中抽取所有的词汇和词类标记，并用最大似然估计方法计算各种概率。

隐马尔可夫模型：应用举例

$$\bar{\pi}_{\mathrm{pos}_i} = \frac{\mathrm{POS}_i\text{出现在句首的次数}}{\text{所有句首的个数}}$$

$$\bar{a}_{ij} = \frac{\text{从词类}\mathrm{POS}_i\text{转移到}\mathrm{POS}_j\text{的次数}}{\text{所有从状态}\mathrm{POS}_i\text{转移到另一}\mathrm{POS}(\text{包括}\mathrm{POS}_j)\text{的总数}}$$

$$\bar{b}_j(k) = \frac{\text{从状态}\mathrm{POS}_j\text{输出词汇}w_k\text{的次数}}{\text{状态}\mathrm{POS}_j\text{出现的总次数}}$$

隐马尔可夫模型：应用举例

- 问题2：如何获取观察序列?
  - 借助其他工具，获得n-best的粗切分。

本地主叫通话时长1400分钟。

→ 本地/ 主叫/ 通话/ 时长/ 1400/ 分钟/ 。

本/ 地主/ 叫/ 通话/ 时/ 长/ 1400/ 分钟/ 。

本/ 地主/ 叫/ 通话/ 时长/ 1400/ 分钟/ 。

负责任 → 负/ 责任

负责/ 任

负/ 责/ 任

隐马尔可夫模型：应用举例

- 分词实验：以“负责任”为例
- 利用部分人民日报语料。

词类 词	A	C	Q	NF	NG	NL	V	VN	总计
负责	4	0	0	0	0	0	177	50	231
任	0	4	11	59	2	4	98	0	178
其他	34469	25475	24232	11453	4550	25670	184488	42674	
总计	34473	25479	24243	11512	4552	25674	184763	42724	

隐马尔可夫模型：应用举例

$O_1 = w_1\ w_2 =$ 负责/ 任 $\qquad p(O_1|\mu) = 5.4 \times 10^{-6}$

$O_2 = w_1\ w_2 =$ 负/ 责任 $\qquad p(O_2|\mu) = 9.3 \times 10^{-4}$

$O_3 = w_1\ w_2\ w_3 =$ 负/ 责/ 任 $\qquad p(O_3|\mu) = 4.3 \times 10^{-6}$

$$\boxed{p(O_2|\mu) > p(O_1|\mu) > p(O_3|\mu)}$$

- 第二种切分结果可能性较大：负/责任。

隐马尔可夫模型：应用举例

- 分词性能测试：
  - 封闭测试：《人民日报》1998年1月份的部分切分和标注语料，约占训练语料的1/10，计78396个词，含中国人名1273个。（人名识别前）准确率：90.34%。
  - 开放测试：《人民日报》1998年2月份的部分切分和标注语料，也占训练语料的1/10，共82347个词，含中国人名2316个。（人名识别前）准确率：86.32%。

隐马尔可夫模型：应用举例

- 词性标注：$Q' = argmax_Q P(Q|O,\mu)$。
  - 采用有指导的参数估计方法。
  - 训练语料：北京大学标注的《人民日报》2000年1、2、4月的语料。
  - 封闭测试：2000年2月20-29日的标注语料，词性标注的精确率为：95.16%；
  - 开放测试：2000年3月1-7日的语料，词性标注的精确率为88.45%。