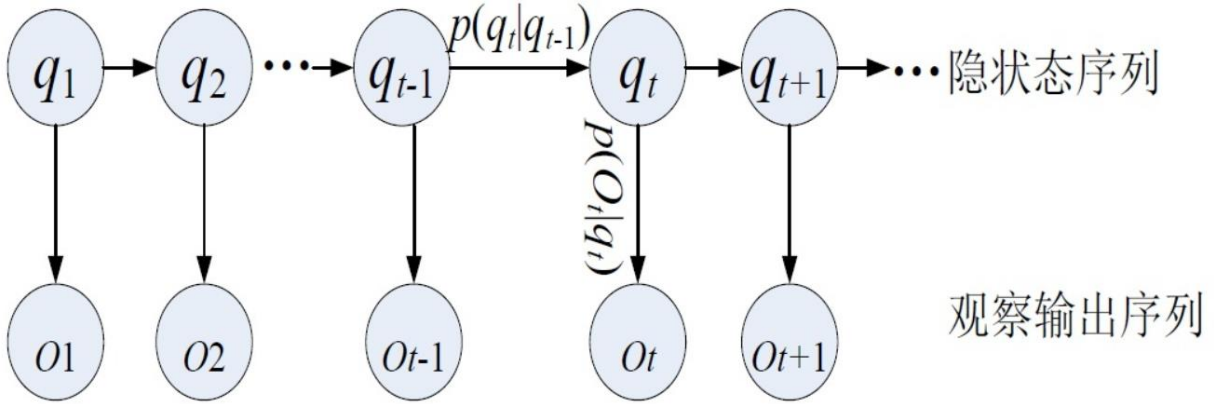
机器学习

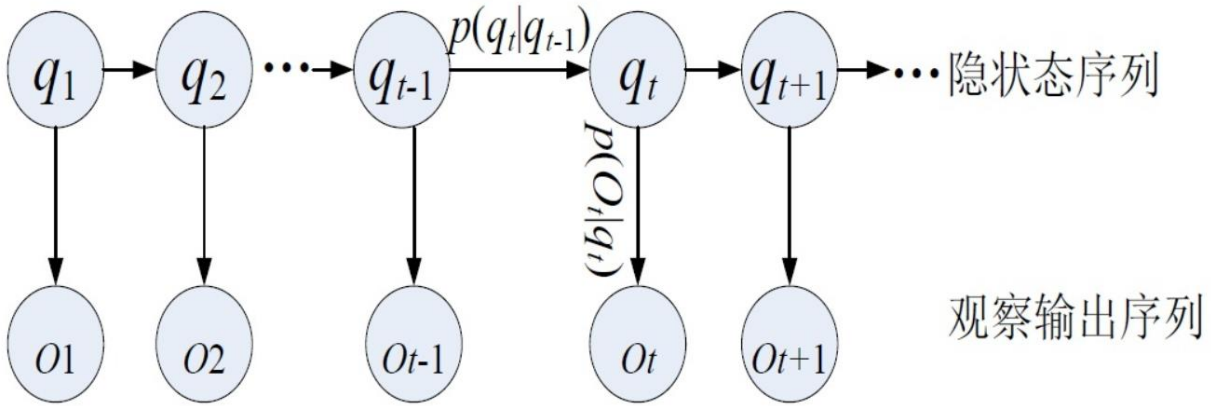
苏州大学计算机科学与技术学院
自然语言处理实验室
主讲：周夏冰
邮箱：zhouxiabing@suda.edu.cn

隐马尔科夫模型

HMM

- 隐马尔可夫模型是关于时序的概率模型;
- 描述由一个隐藏的马尔可夫链随机生成不可观测的状态随机序列(state sequence)，再由各个状态生成一个观测而产生观测随机序列(observation sequence)的过程,序列的每一个位置又可以看作是一个时刻。
- 贝叶斯网络





基因序列分析、蛋白质结构预测

状态序列：基因序列中的功能区域（如编码区、非编码区）
观测序列：DNA或RNA的核苷酸序列

语音识别

状态序列：语音中的音素或音节序列
观测序列：语音信号的声学特征

词性标注、命名实体识别

状态序列：词性标签（如名词、动词）或命名实体类型（如人名、地名）
观测序列：文本中的单词序列

- 需要模型可解释性（如医疗、金融）。
- 数据规模小或标注成本高。
- 实时性要求高（如嵌入式设备）。

HMM

- 组成
  - 初始概率分布，状态转移概率分布，观测概率分布
  - Q：所有可能状态的集合 $Q = \{q_1, q_2, \cdots, q_N\}$
  - V：所有可能观测的集合 $V = \{v_1, v_2, \cdots, v_N\}$
  - I: 长度为T的状态序列 $I = \{i_1, i_2, \cdots, i_N\}$
  - O：对应的观测序列 $O = \{o_1, o_2, \cdots, o_N\}$

HMM

- 组成
  - 状态转移矩阵 $A = [a_{ij}]_{N\times N}$

    $a_{ij} = P(i_{t+1} = q_j | i_t = q_i)$
  - 观测概率 $B = [b_j(k)]_{N\times M}$
    - $b_j(k) = P(o_t = v_k | i_t = q_j)$
  - 初始概率 $\pi = (\pi_i)$
    - $\pi_i = P(i_1 = q_i)$

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

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

HMM

- 例：盒子和球模型
  - 状态集合：Q={盒子1，盒子2，盒子3，盒子4}， N=4
  - 观测集合：V={红球，白球} M=2
  - 初始化概率分布： $\pi = (0.25, 0.25, 0.25, 0.25)^T$
  - 状态转移矩阵： 观测矩阵：

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 \\ 0 & 0.4 & 0 & 0.6 \\ 0 & 0 & 0.5 & 0.5 \end{bmatrix} \qquad B = \begin{bmatrix} 0.5 & 0.5 \\ 0.3 & 0.7 \\ 0.6 & 0.4 \\ 0.8 & 0.2 \end{bmatrix}$$

  - 观测结果： O=(红，红，白，白，红) T=5

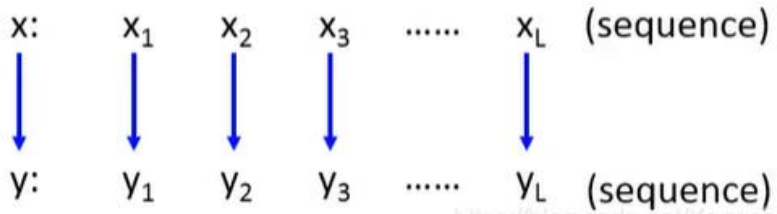
HMM—序列标注

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.2767	0.0006	0.0031	0.0453	0.0449	0.0510	0.2026
NNP	0.3777	0.0110	0.0009	0.0084	0.0584	0.0090	0.0025
MD	0.0008	0.0002	0.7968	0.0005	0.0008	0.1698	0.0041
VB	0.0322	0.0005	0.0050	0.0837	0.0615	0.0514	0.2231
JJ	0.0366	0.0004	0.0001	0.0733	0.4509	0.0036	0.0036
NN	0.0096	0.0176	0.0014	0.0086	0.1216	0.0177	0.0068
RB	0.0068	0.0102	0.1011	0.1012	0.0120	0.0728	0.0479
DT	0.1147	0.0021	0.0002	0.2157	0.4744	0.0102	0.0017

Sequence Labeling

$$f : X \rightarrow Y$$

Sequence Sequence



	Janet	will	back	the	bill
NNP	0.000032	0	0	0.000048	0
MD	0	0.308431	0	0	0
VB	0	0.000028	0.000672	0	0.000028
JJ	0	0	0.000340	0	0
NN	0	0.000200	0.000223	0	0.002337
RB	0	0	0.010446	0	0
DT	0	0	0	0.506099	0

名词	动词	副词	形容词	形容词	结构助词	名词
中国	是	非常	繁荣	稳定	的	国家

HMM

- 1、概率计算问题
  - 给定： $\lambda=(A,B,\pi)$ $O=(o_1,o_2,\cdots,o_T)$
  - 计算： $P(O|\lambda)$
- 2、学习问题
  - 已知： $O=(o_1,o_2,\cdots,o_T)$
  - 估计： $\lambda=(A,B,\pi)$
- 3、预测问题（解码）
  - 已知： $\lambda=(A,B,\pi)$ $O=(o_1,o_2,\cdots,o_T)$
  - 求：使 $P(I|O)$ 最大的状态序列 $I=(i_1,i_2,\cdots,i_T)$

隐马尔可夫模型的三个基本问题

- 1、概率计算问题
  - 给定：$\lambda = (A, B, \pi), O = (o_1, o_2, \cdots, o_T)$
  - 计算：$P(O|\lambda)$
- 2、学习问题
- 3、预测问题（解码）

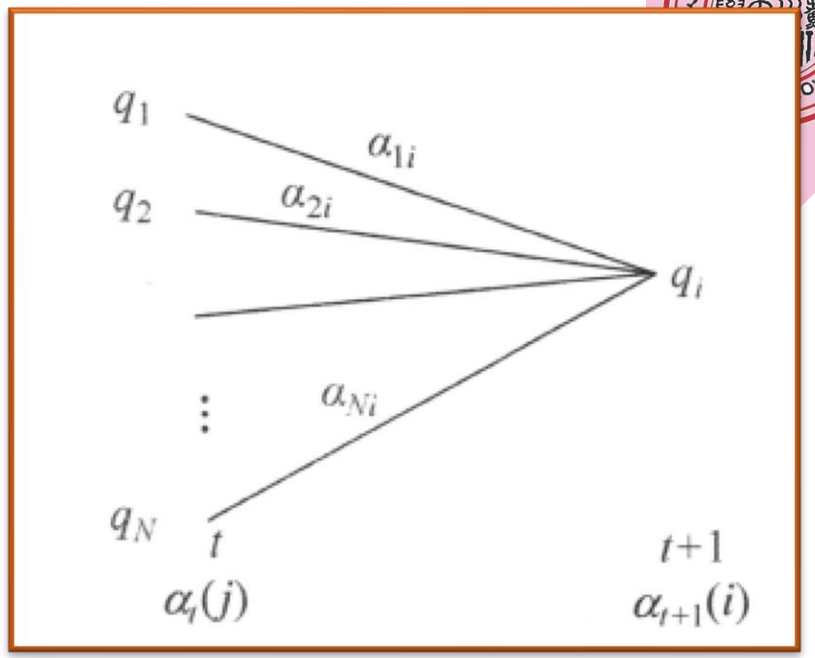
概率计算算法—直接计算法

- 思想：列举所有可能的状态序列，分别计算观测概率进行求和
- 对于状态$I=(i_1,i_2,\cdots,i_T)$，我们可以得到
  - $P(O,I|\lambda)=P(O|I,\lambda)P(I|\lambda)$
  - $P(O|I,\lambda)=b_{i_1}(o_1)b_{i_2}(o_2)\cdots b_{i_T}(o_T)$
  - $P(I|\lambda)=\pi_{i_1}a_{i_1i_2}a_{i_2i_3}\cdots a_{i_{T-1}i_T}$
  - $P(O|\lambda)=\sum_I P(O|I,\lambda)P(I|\lambda)=\sum_{i_1,i_2,\cdots,i_T}\pi_{i_1}b_{i_1}(o_1)a_{i_1i_2}b_{i_2}(o_2)a_{i_2i_3}\cdots a_{i_{T-1}i_T}b_{i_T}(o_T)$
- 所有状态计算量太大

HMM——概率计算

- 前向/后向算法
  - 前向概率：$\alpha_t(i) = P(o_1, o_2, \cdots, o_t, i_t = q_i|\lambda)$
    - 初始：

$$\alpha_1(i) = \pi_i b_i(o_1), i = 1, 2, \cdots, N$$



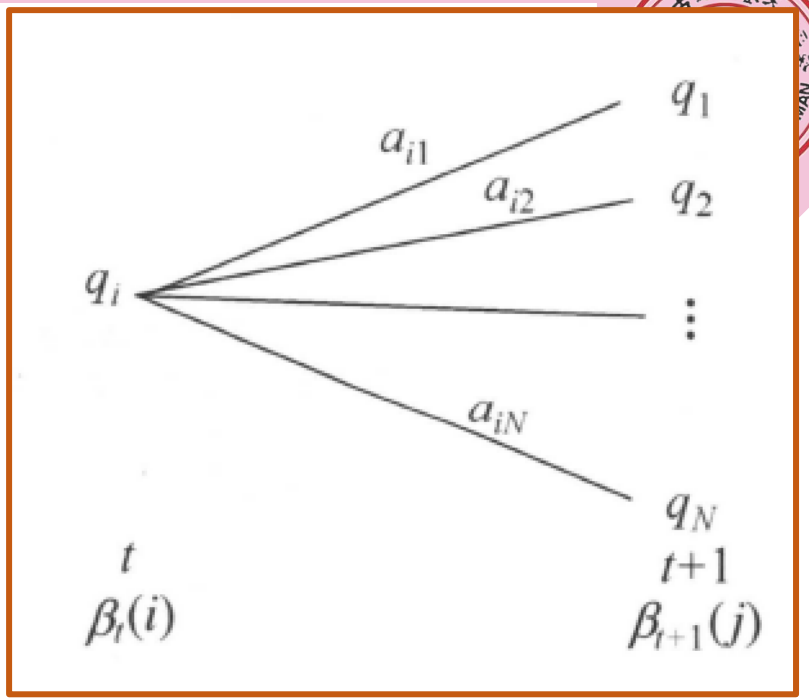
    - 递推：

$$\alpha_{t+1}(i) = [\sum_{j=1}^{N} \alpha_t(j) a_{ji}] b_i(o_{t+1}), i = 1, 2, \cdots, N$$

    - 终止：

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

HMM——概率计算



- 前向/后向算法
  - 后向概率：$\boldsymbol{\beta_t(i) = P(o_{t+1}, o_{t+2}, \cdots, o_T | i_t = q_i, \lambda)}$
    - 初始

$$\boldsymbol{\beta_T(i) = 1, i = 1, 2, \cdots, N}$$

    - 递推：

$$\boldsymbol{\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), i = 1, 2, \cdots, N}$$

    - 终止：

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

HMM——概率计算

- 例 考虑盒子和球模型，状态集合$Q = \{1,2,3\}$，观测集合V={红，白}，转移矩阵$A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$，观测矩阵$B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}$，初始矩阵$\pi = [0.2, 0.4, 0.4]$。

  设T=3，O=(红，白，红)，计算$P(O|\lambda)$

- 前向算法：
  - 初始值

$$\alpha_1(1) = \pi_1 b_1(o_1) = 0.10$$
$$\alpha_1(2) = \pi_2 b_2(o_1) = 0.16$$
$$\alpha_1(3) = \pi_3 b_3(o_1) = 0.28$$

HMM——概率计算

- 例 考虑盒子和球模型，状态集合$Q=\{1,2,3\}$，观测集合V={红，白}，转移矩阵$A=\begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\end{bmatrix}$，观测矩阵$B=\begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\\ 0.7 & 0.3\end{bmatrix}$，初始矩阵$\pi=[0.2,\ 0.4, 0.4]$。

  设T=3，O=(红，白，红)，计算$P(O|\lambda)$

- 前向算法：

  - 递推计算

$$\alpha_2(1)=\left[\sum_{i=1}^{3}\alpha_1(i)a_{i1}\right]b_1(o_2)=0.0770$$

$$\alpha_2(2)=\left[\sum_{i=1}^{3}\alpha_1(i)a_{i2}\right]b_2(o_2)=0.1104$$

$$\alpha_2(3)=\left[\sum_{i=1}^{3}\alpha_1(i)a_{i3}\right]b_3(o_2)=0.0606$$

$$\alpha_3(1)=\left[\sum_{i=1}^{3}\alpha_2(i)a_{i1}\right]b_1(o_3)=0.04187$$

$$\alpha_3(2)=\left[\sum_{i=1}^{3}\alpha_2(i)a_{i2}\right]b_2(o_3)=0.03551$$

$$\alpha_3(3)=\left[\sum_{i=1}^{3}\alpha_2(i)a_{i3}\right]b_3(o_3)=0.05284$$

HMM——概率计算

- 例 考虑盒子和球模型，状态集合$Q = \{1,2,3\}$，观测集合V={红，白}，转移矩阵$A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$，观测矩阵$B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}$，初始矩阵$\pi = [0.2,\ 0.4, 0.4]$。

  设T=3，O=(红，白，红)，计算$P(O|\lambda)$

- 前向算法：
  - 终止:$P(O|\lambda) = \sum_{i=1}^{3} \alpha_3(i) = 0.13022$

概率计算算法—后向算法

- 例 考虑盒子和球模型，状态集合$Q=\{1,2,3\}$，观测集合V={红，白}，转

移矩阵$A=\begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\end{bmatrix}$，观测矩阵$B=\begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\\ 0.7 & 0.3\end{bmatrix}$，初始矩阵$\pi=[0.2,$

$0.4,0.4]$。设T=3，O=(红，白，红)，计算$P(O|\lambda)$

- **后向算法**：
  - 初始值

$$\beta_3(1)=1$$
$$\beta_3(2)=1$$
$$\beta_3(3)=1$$

概率计算算法—后向算法

- 例 考虑盒子和球模型，状态集合$Q=\{1,2,3\}$，观测集合V={红，白}，转移矩阵$A=\begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\end{bmatrix}$，观测矩阵$B=\begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\\ 0.7 & 0.3\end{bmatrix}$，初始矩阵$\pi=[0.2, 0.4, 0.4]$。设T=3，O=(红，白，红)，计算$P(O|\lambda)$

- **前向算法**：
  - **递推计算**

$$\beta_2(1)=\sum_{j=1}^{3}a_{1j}b_j(o_3)\beta_3(j)=0.54 \qquad \beta_1(1)=\sum_{j=1}^{3}a_{1j}b_j(o_2)\beta_2(j)=0.2451$$

$$\beta_2(2)=\sum_{j=1}^{3}a_{2j}b_j(o_3)\beta_3(j)=0.49 \qquad \beta_1(2)=\sum_{j=1}^{3}a_{2j}b_j(o_2)\beta_2(j)=0.2622$$

$$\beta_2(3)=\sum_{j=1}^{3}a_{3j}b_j(o_3)\beta_3(j)=0.57 \qquad \beta_1(3)=\sum_{j=1}^{3}a_{3j}b_j(o_2)\beta_2(j)=0.2277$$

概率计算算法—后向算法

- 例 考虑盒子和球模型，状态集合$Q=\{1,2,3\}$，观测集合V={红，白}，转移矩阵$A=\begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\end{bmatrix}$，观测矩阵$B=\begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\\ 0.7 & 0.3\end{bmatrix}$，初始矩阵$\pi=[0.2, 0.4, 0.4]$。设T=3，O=(红，白，红)，计算$P(O|\lambda)$

- **前向算法**：
  - 终止:$P(O|\lambda)=\sum_{i=1}^{3}\pi_i b_i(o_1)\beta_1(i)=0.13022$

概率计算算法

- 根据前向和后向概率的定义，可以将观测概率$P(O|\lambda)$统一写成
  - $P(O|\lambda)=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j),\quad t=1,2,\cdots,T-1$

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

$$\beta_{t+1}(j)=P(o_{t+2},o_{t+3},\cdots,o_T|i_{t+1}=q_j,\lambda)$$

$$P(i_{t+1}=q_j|i_t=q_i,\lambda)$$

$$P(o_{t+1}|i_{t+1}=q_j,\lambda)$$

$$\sum_i P(o_1,\cdots,o_t,i|\lambda)\sum_j P(j|i)P(o_{t+1},\cdots,o_T|j,\lambda)=\sum_i P(o_1,\cdots,o_t,i,\lambda)P(o_{t+1},\cdots,o_T|i,\lambda)=P(o_1,\cdots,o_T|\lambda)$$

一些概率与期望值的计算

- 利用前向和后向概率，可以得到单个状态和两个状态的概率计算公式
  - 时刻t处于状态q_i的概率
    - $\gamma_t(i) = P(i_t = q_i|O,\lambda) = \frac{P(i_t = q_i, O|\lambda)}{P(O|\lambda)}$
    - $\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)}$

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

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

  - 时刻t处于状态q_i且时刻t+1处于状态q_j
    - $\xi_t(i,j) = P(i_t = q_i, i_{t+1} = q_j|O,\lambda) = \frac{P(i_t=q_i, i_{t+1}=q_j, O|\lambda)}{P(O|\lambda)}$
    - $\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(i)}$

（1）观测O下状态i出现的期望

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

（2）观测O下状态i转移的期望

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

（3）观测O下状态i转移状态j的期望

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

隐马尔可夫模型的三个基本问题

- 1、概率计算问题
- **2、学习问题**
  - 已知：$O=(o_1,o_2,\cdots,o_T)$
  - 估计：$\lambda=(A,B,\pi)$
- 3、预测问题（解码）

HMM——学习算法

- 监督学习 ⇒ 观测序列+状态序列
- Baum-Welch算法 ⇒ 观测序列

学习算法——监督学习方法

- 假设已给训练数据包含S个长度相同的观测序列和对应的状态序列 $\{(O_1, I_1), (O_2, I_2) \cdots, (O_S, I_S)\}$

➢ 状态转移矩阵

- $a_{ij} = \frac{A_{ij}}{\sum_{k=1}^{N} A_{ik}}, \quad i = 1,2,\cdots,N; \quad j = 1,2,\cdots,N$

➢观测概率

- $b_j(k) = \frac{B_{jk}}{\sum_{l=1}^{M} B_{jl}}, \ j = 1,2,\cdots,N; \quad k = 1,2,\cdots,M$

➢初始概率

学习算法——无监督学习方法

- **Baum-Welch 算法**
- 假设已给训练数据包含S个长度为T的观测序列 $\{O_1, O_2, \cdots, O_S\}$，目标是学习隐马尔可夫模型中的参数 $\lambda = (A, B, \pi)$
  - 隐变量：状态序列

学习算法——无监督学习方法

- **EM算法**
  - **E-step:** 记$\boldsymbol{\theta^i}$为第$\boldsymbol{i}$次迭代参数$\boldsymbol{\theta}$的估计值，在第$\boldsymbol{i+1}$次迭代的E步，计算

$$\boldsymbol{Q(\theta,\theta^i)=\sum_Z log\,P(Y,Z|\theta)\,P(Z|Y,\theta^i)}$$

  - **M-step:** 求使$\boldsymbol{Q(\theta,\theta^i)}$极大化的$\boldsymbol{\theta}$，确定第i+1次迭代的参数的估计值$\boldsymbol{\theta^{i+1}}$

$$\boldsymbol{\theta^{i+1}=argmax_\theta Q(\theta,\theta^i)}$$

学习算法——BW 算法

- 观测序列：$(O_1, \cdots, O_n)$
- $\lambda = (\pi, A, B)$
- $\boldsymbol{Q(\lambda, \bar{\lambda}) = \sum_I P(I|O, \bar{\lambda}) \, log \, P(O, I|\lambda)}$
- $\boldsymbol{P(O, I|\lambda) = \pi_{i_1} b_{i_1}(O_1) a_{i_1 i_2} b_{i_2}(O_2) \cdots a_{i_{T-1} i_T} b_{i_T}(O_T)}$

学习算法——BW 算法

- 观测序列：$(O_1, \cdots, O_n)$
- $\lambda = (\pi, A, B)$
- $\boldsymbol{Q(\lambda, \bar{\lambda}) = \sum_I P(I|O, \bar{\lambda}) \log P(O, I|\lambda)}$
- $\boldsymbol{P(O, I|\lambda) = \pi_{i_1} b_{i_1}(O_1) a_{i_1 i_2} b_{i_2}(O_2) \cdots a_{i_{T-1} i_T} b_{i_T}(O_T)}$
- $\boldsymbol{\log P(O, I|\lambda) = \log \pi_{i_1} + \sum_{t=1}^{T} \log b_{i_t}(O_t) + \sum_{t=1}^{T-1} \log a_{i_t i_{t+1}}}$

学习算法——BW 算法

- 观测序列：$(O_1, \cdots, O_n)$
- $\lambda = (\pi, A, B)$
- $Q(\lambda, \bar{\lambda}) = \sum_I P(I|O, \bar{\lambda}) \log P(O, I|\lambda)$
- $P(O, I|\lambda) = \pi_{i_1} b_{i_1}(O_1) a_{i_1 i_2} b_{i_2}(O_2) \cdots a_{i_{T-1} i_T} b_{i_T}(O_T)$
- $\log P(O, I|\lambda) = \log \pi_{i_1} + \sum_{t=1}^{T} \log b_{i_t}(O_t) + \sum_{t=1}^{T-1} \log a_{i_t i_{t+1}}$
- $P(I|O, \bar{\lambda}) = \frac{P(I, O|\bar{\lambda})}{P(O|\bar{\lambda})}$

学习算法——BW 算法

- $Q = \sum_I \left[P(I, O|\bar{\lambda}) \log \pi_{i_1} + P(I, O|\bar{\lambda}) \sum_{t=1}^{T} \log b_{i_t}(O_t) + P(I, O|\bar{\lambda}) \sum_{t=1}^{T-1} \log a_{i_t i_{t+1}} \right]$

状态个数

- $\sum_I P(I, O|\bar{\lambda}) \log \pi_{i_1} = \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_T=1}^{N} P(O, i_1, i_2, \cdots, i_T|\bar{\lambda}) \log \pi_{i_1}$
- $\sum_{i_2=1}^{N} \cdots \sum_{i_T=1}^{N} P(O, i_1 = i, i_2, \cdots, i_T|\bar{\lambda}) \log \pi_i$
- 边缘分布： $= P(O, i_1 = i\,|\bar{\lambda}) \log \pi_i$

边缘概率

- 对于离散随机变量 X 和 Y，其联合概率为 $P(X=x,Y=y)$。边缘概率 $P(X=x)$ 的计算方式为：
  - $P(X=x)=\sum_y P(X=x,Y=y)$
- 将所有可能的 Y 值对应的联合概率相加，消去 Y 的影响，得到 X 的概率分布

$$\sum_{i_2=1}^{N}\cdots\sum_{i_T=1}^{N} P(O,i_1=i,i_2,\cdots,i_T|\bar{\lambda})\log\pi_i \quad \Rightarrow \quad P(O,i_1=i\,|\bar{\lambda})\log\pi_i$$

学习算法——BW 算法

- $Q = \sum_I \left[P(I, O|\bar{\lambda}) \log \pi_{i_1} + P(I, O|\bar{\lambda}) \sum_{t=1}^{T} \log b_{i_t}(O_t) + P(I, O|\bar{\lambda}) \sum_{t=1}^{T-1} \log a_{i_t i_{t+1}} \right]$

状态个数

- $\sum_I P(I, O|\bar{\lambda}) \log \pi_{i_1} = \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_T=1}^{N} P(O, i_1, i_2, \cdots, i_T|\bar{\lambda}) \log \pi_{i_1}$
- $\sum_{i_2=1}^{N} \cdots \sum_{i_T=1}^{N} P(O, i_1 = i, i_2, \cdots, i_T|\bar{\lambda}) \log \pi_i$
- 边缘分布： $= P(O, i_1 = i\,|\bar{\lambda}) \log \pi_i$
- $\sum_I P(I, O|\bar{\lambda}) \log \pi_{i_1} = \sum_{i=1}^{N} P(O, i_1 = i\,|\bar{\lambda}) \log \pi_i$

学习算法——BW 算法

- $Q = \sum_I \left[P(I, O|\bar{\lambda}) \log \pi_{i_1} + P(I, O|\bar{\lambda}) \sum_{t=1}^{T} \log b_{i_t}(O_t) + P(I, O|\bar{\lambda}) \sum_{t=1}^{T-1} \log a_{i_t i_{t+1}} \right]$
- $\sum_I P(I, O|\bar{\lambda}) \log \pi_{i_1} = \sum_{i=1}^{N} P(O, i_1 = i|\bar{\lambda}) \log \pi_i$
- 有：$\sum_{i=1}^{N} \pi_i = 1$
- 拉格朗日乘子法，拉格朗日函数为：
  - $\sum_{i=1}^{N} P(O, i_1 = i|\bar{\lambda}) \log \pi_i + \lambda(\sum_{i=1}^{N} \pi_i - 1)$

学习算法——BW 算法

- $\sum_{i=1}^{N} P(O, i_1 = i \mid \bar{\lambda}) \log \pi_i + \lambda(\sum_{i=1}^{N} \pi_i - 1)$
- 求导
- $P(O, i_1 = i \mid \bar{\lambda}) + \lambda \pi_i = 0 \quad (*)$
- $\sum_{i=1}^{N} P(O, i_1 = i \mid \bar{\lambda}) + \sum_{i=1}^{N} \lambda \pi_i = 0$
- $\lambda = -P(O|\bar{\lambda})$

边缘概率

- $\pi_i = \frac{P(O, i_1 = i \mid \bar{\lambda})}{P(O|\bar{\lambda})}$ $\gamma_1(i)$

学习算法——BW 算法

- $Q = \sum_I \left[P(I, O|\bar{\lambda}) \log \pi_{i_1} + P(I, O|\bar{\lambda}) \sum_{t=1}^{T} \log b_{i_t}(O_t) + P(I, O|\bar{\lambda}) \sum_{t=1}^{T-1} \log a_{i_t i_{t+1}} \right]$
- $\sum_I P(I, O|\bar{\lambda}) \sum_{t=1}^{T-1} \log a_{i_t i_{t+1}} = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T-1} P(O, i_t = i, i_{t+1} = j|\bar{\lambda}) \log a_{ij}$
- 有：$\sum_{j=1}^{N} a_{ij} = 1$
- 拉格朗日乘子法，拉格朗日函数为：
  - $\sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T-1} P(O, i_t = i, i_{t+1} = j|\bar{\lambda}) \log a_{ij} + \gamma(\sum_{j=1}^{N} a_{ij} - 1)$

学习算法——BW 算法

- $\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1} P(O, i_t = i, i_{t+1} = j \mid \bar{\lambda}) \log a_{ij} + \gamma(\sum_{j=1}^{N} a_{ij} - 1)$
- 求导
- $\sum_{t=1}^{T-1} P(O, i_t = i, i_{t+1} = j \mid \bar{\lambda}) + \gamma a_{ij} = 0$
- $\sum_{j=1}^{N}\sum_{t=1}^{T-1} P(O, i_t = i, i_{t+1} = j \mid \bar{\lambda}) + \sum_{j=1}^{N} \gamma a_{ij} = 0$

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

- $\gamma = -\sum_{t=1}^{T-1} P(O, i_t = i \mid \bar{\lambda})$
- $a_{ij} = \frac{\sum_{t=1}^{T-1} P(O, i_t = i, i_{t+1} = j \mid \bar{\lambda})}{\sum_{t=1}^{T-1} P(O, i_t = i \mid \bar{\lambda})}$

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

学习算法——BW 算法

- $Q = \sum_I \left[P(I, O|\bar{\lambda}) \log \pi_{i_1} + P(I, O|\bar{\lambda}) \sum_{t=1}^{T} \log b_{i_t}(O_t) + P(I, O|\bar{\lambda}) \sum_{t=1}^{T-1} \log a_{i_t i_{t+1}} \right]$
- $\sum_I P(I, O|\bar{\lambda}) \sum_{t=1}^{T} \log b_{i_t}(O_t) = \sum_{i=1}^{N} \sum_{t=1}^{T} P(O, i_t = i\,|\bar{\lambda}) \log b_i(O_t)$
- 有：$\sum_{k=1}^{M} b_i(k) = 1$
- 拉格朗日乘子法，拉格朗日函数为：
  - $\sum_{i=1}^{N} \sum_{t=1}^{T} P(O, i_t = i\,|\bar{\lambda}) \log b_i(O_t) + \beta(\sum_{k=1}^{M} b_i(k) - 1)$

学习算法——BW 算法

$o_t \neq k$
加号前面的部分与求导项无关

- $\sum_{i=1}^{N}\sum_{t=1}^{T} P(O, i_t = i \mid \bar{\lambda}) \log b_i(O_t) + \beta(\sum_{k=1}^{M} b_i(k) - 1)$
- 对$b_i(k)$求导
- $\sum_{t=1}^{T} P(O, i_t = i \mid \bar{\lambda}) I(O_t = k) + \beta b_i(k) = 0$

前提： $O_t = k$

- $\sum_{k=1}^{M}\sum_{t=1}^{T} P(O, i_t = i \mid \bar{\lambda}) I(O_t = k) + \sum_{k=1}^{M} \beta b_i(k) = 0$
- $\beta = -\sum_{t=1}^{T} P(O, i_t = i \mid \bar{\lambda})$
- $b_i(k) = \frac{\sum_{t=1}^{T} P(O, i_t = i \mid \bar{\lambda}) I(O_t = k)}{\sum_{t=1}^{T} P(O, i_t = i \mid \bar{\lambda})}$ $\quad \frac{\sum_{t=1, o_t = k}^{T} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}$

学习算法——BW 算法

- 初始化：$n=0$，给出$a_{ij}^0, b_j(k)^0, \pi_i^0$，得到模型初始参数$\lambda^0=(\pi^0, A^0, B^0)$
- 对于$n=1,2,\cdots,$
  - $\pi_i^{n+1}=\gamma_1(i)$，$a_{ij}^{n+1}=\frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}$，$b_i(k)^{n+1}=\frac{\sum_{t=1,o_t=k}^{T}\gamma_t(i)}{\sum_{t=1}^{T}\gamma_t(i)}$
  - $\gamma_t(i)=\frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)}$，$\xi_t(i,j)=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(i)}$，$\lambda^n=(\pi^n, A^n, B^n)$
- 终止

隐马尔可夫模型的三个基本问题

- 1、概率计算问题
- 2、学习问题
- **3、预测问题（解码）**
  - 已知：$\lambda = (A, B, \pi), O = (o_1, o_2, \cdots, o_T)$
  - 求：使$P(I|O)$最大的状态序列 $I = (i_1, i_2, \cdots, i_T)$

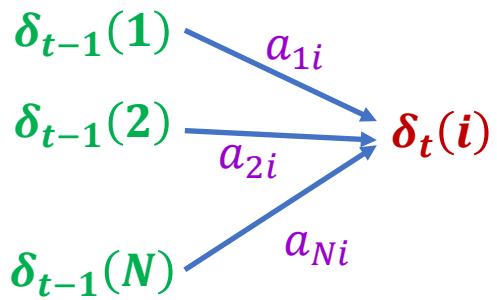
HMM——预测算法

- 维特比算法

HMM——预测算法

- 维特比算法
- 思想：动态规划求概率最大路径
  - 如果最优路径在时刻t通过结点i_t^*，那么这一路径从结点i_t^*到终点i_T^*的部分路径，对于从i_t^* 到i_T^*的所有可能的部分路径来说，必须是最优的。
  - t=1开始递推，前面都选好的基础上，此t时刻状态为i的部分路径的最大概率是哪个

预测算法——维特比算法



- **维特比算法**
  - 在时刻t状态为i的所有单个路径$(i_1, i_2, \cdots, i_t)$中概率最大值为：
    - $\delta_t(i) = \max\limits_{i_1,i_2,\cdots,i_{t-1}} P(i_t = i, i_{t-1}, \cdots, i_1, o_t, \cdots, o_1|\lambda), \quad i = 1,2,\cdots,N$
    - $= \max\limits_{1\le j\le N}[\delta_{t-1}(j)a_{ji}]b_i(o_t), \ i = 1,\cdots,N; t = 1,\cdots,T-1$
  - 定义在时刻t状态为i的所有单个路径中概率最大的路径的第t-1个结点为：
    - $\Psi_t(i) = \arg\max\limits_{1\le j\le N}[\delta_{t-1}(j)a_{ji}], \quad i = 1,2,\cdots,N$

预测算法——维特比算法

- 初始化：$\delta_1(i) = \pi_i b_i(o_1);\ \Psi_1(i) = 0;\quad i = 1,2,\cdots,N$
- 递推：对$t = 2,3,\cdots,T$
  - $\boldsymbol{\delta_t(i) = \max\limits_{1\le j\le N}\left[\delta_{t-1}(j)a_{ji}\right]b_i(o_t),\ i = 1,\cdots,N}$
  - $\boldsymbol{\Psi_t(i) = \arg\max\limits_{1\le j\le N}\left[\delta_{t-1}(j)a_{ji}\right],\quad i = 1,2,\cdots,N}$
- 终止
  - $\boldsymbol{P^* = \max\limits_{1\le i\le N}\delta_T(i)}$
  - $\boldsymbol{i^* = arg\max\limits_{1\le j\le N}[\delta_T(j)]}$

最优路径回溯：$t = T-1, T-2, \cdots, 1$

$$i_t^* = \Psi_{t+1}(i_{t+1}^*)$$

得到的最优路径为：

$$I^* = (i_1^*, i_2^*, \cdots, i_T^*)$$

预测算法——维特比算法

- 例 已知模型参数 $A=\begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\end{bmatrix}$，$B=\begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\\ 0.7 & 0.3\end{bmatrix}$，$\pi=\begin{bmatrix}0.2\\ 0.4\\ 0.4\end{bmatrix}$，

  和观测序列 $O=\{红，白，红\}$，求最优路径$I^*=(i_1^*,i_2^*,i_3^*)$

- 初始化：
  - $\boldsymbol{\delta_1(i)=\pi_i b_i(o_1)=\pi_i b_i(红)\quad i=1,2,3}$
  - $\boldsymbol{\delta_1(1)=0.2\times0.5=0.10;\ \delta_1(2)=0.4\times0.4=0.16;\ \delta_1(3)=0.4\times0.7=0.28}$
  - $\boldsymbol{\Psi_1(i)=0}$

预测算法——维特比算法

- 例 已知模型参数 $A=\begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\end{bmatrix}$，$B=\begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\\ 0.7 & 0.3\end{bmatrix}$，$\pi=\begin{bmatrix}0.2\\ 0.4\\ 0.4\end{bmatrix}$，和观测序列 $O=\{红，白，红\}$，求最优路径$I^*=(i_1^*,i_2^*,i_3^*)$

- $t=2$:
  - $\delta_2(i)=\max\limits_{1\le j\le N}[\delta_1(j)a_{ji}]b_i(o_2)=\max\limits_{1\le j\le N}[\delta_1(j)a_{ji}]b_i(白)$ $\quad i=1,2,3$
  - $\delta_2(1)=\max\limits_{1\le j\le 3}[0.10\times0.5,\ 0.16\times0.3,\ 0.28\times0.2]\times0.5=0.028$ $\quad \Psi_2(1)=3$
  - $\delta_2(2)=\max\limits_{1\le j\le 3}[0.10\times0.2,\ 0.16\times0.5,\ 0.28\times0.3]\times0.6=0.0504$ $\quad \Psi_2(2)=3$
  - $\delta_2(3)=\max\limits_{1\le j\le 3}[0.10\times0.3,\ 0.16\times0.2,\ 0.28\times0.5]\times0.3=0.042$ $\quad \Psi_2(3)=3$

预测算法——维特比算法

- 例 已知模型参数 $A=\begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\end{bmatrix}$，$B=\begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\\ 0.7 & 0.3\end{bmatrix}$，$\pi=\begin{bmatrix}0.2\\ 0.4\\ 0.4\end{bmatrix}$，和观测序列 $O=\{红，白，红\}$，求最优路径$I^*=(i_1^*,i_2^*,i_3^*)$

- $t=3$:
  - $\delta_3(i)=\max\limits_{1\le j\le N}[\delta_2(j)a_{ji}]b_i(o_3)=\max\limits_{1\le j\le N}[\delta_2(j)a_{ji}]b_i(红)$　　$i=1,2,3$
  - $\delta_3(1)=\max\limits_{1\le j\le 3}[0.028\times 0.5,\ 0.0504\times 0.3,\ 0.042\times 0.2]\times 0.5=0.00756$　　$\Psi_3(1)=2$
  - $\delta_3(2)=\max\limits_{1\le j\le 3}[0.028\times 0.2,\ 0.0504\times 0.5,\ 0.042\times 0.3]\times 0.4=0.01008$　　$\Psi_3(2)=2$
  - $\delta_3(3)=\max\limits_{1\le j\le 3}[0.028\times 0.3,\ 0.0504\times 0.2,\ 0.042\times 0.5]\times 0.7=0.0147$　　$\Psi_3(3)=3$

预测算法——维特比算法

- 例 已知模型参数 $A=\begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\end{bmatrix}$，$B=\begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\\ 0.7 & 0.3\end{bmatrix}$，$\pi=\begin{bmatrix}0.2\\ 0.4\\ 0.4\end{bmatrix}$，

  和观测序列 $O=\{红，白，红\}$，求最优路径$I^*=(i_1^*,i_2^*,i_3^*)$

- 终止：

  - $P^*=\max\limits_{1\le i\le 3}\delta_3(i)=\max\limits_{1\le i\le 3}\{0.00756，\ 0.01008，\ 0.0147\}=0.0147$
  - $i_3^*=arg\max\limits_{1\le j\le 3}[\delta_3(j)]=arg\max\limits_{1\le j\le 3}[0.00756，\ 0.01008，\ 0.0147]=3$
  - $i_2^*=\Psi_3(i_3^*)=\Psi_3(3)=3$
  - $i_1^*=\Psi_2(i_2^*)=\Psi_2(3)=3$

最优路径为：$I^*=(3,3,3)$

预测算法——维特比算法

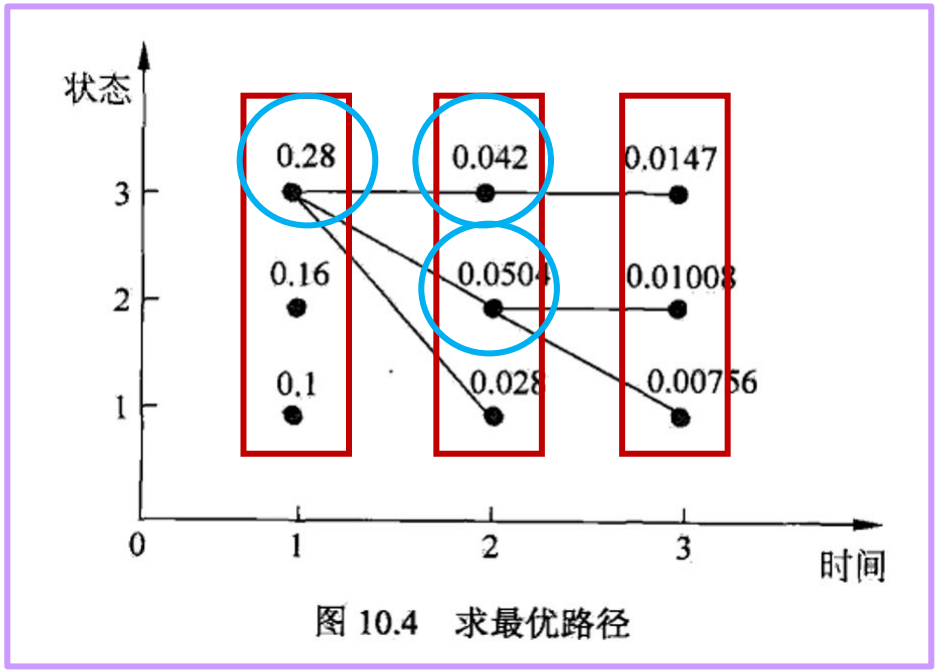


图 10.4　求最优路径

课堂练习

- 利用隐马尔科夫模型进行句子标注工作，假设词性共有：代词、名词、动词和介词，词汇表中共有7个词语[苏州大学，开创，坐落，教育，于，江苏省，苏州市]。经过统计，可得如下参数数据：每个词性状态出现在第一位的概率为(0.3,0.2,0.3,0.2), 词性转移概率表如下（横向看，即代词向其他词转移概率为[0.3,0.25,0.25,0.2]）：

	代词	动词	名词	介词
代词	0.3	0.25	0.25	0.2
动词	0.16	0.12	0.28	0.44
名词	0.14	0.43	0.27	0.16
介词	0.2	0.2	0.5	0.1

	苏州大学	开创	坐落	教育	于	江苏省	苏州市
代词	0.3	0.1	0.1	0.1	0.1	0.1	0.2
动词	0.1	0.2	0.3	0.1	0.1	0.1	0.1
名词	0.2	0.1	0.05	0.2	0.05	0.2	0.2
介词	0.05	0.05	0.05	0.05	0.7	0.05	0.05

- 请对“苏州大学/坐落/于/苏州市”进行词性标注