自然语言处理

第4讲：隐马尔科夫模型

刘洋

内容提要

- 马尔科夫模型
- 计算观测概率
- Viterbi算法
- 前向后向算法

例子

- 假定天气可分为三种状态：晴天、阴天和雨天。下面两张表给出了状态初始化概率分布和状态转移概率分布，则连续四天下雨的概率是多少？

晴天	阴天	雨天
0.4	0.3	0.3

表1：状态初始化概率分布

	晴天	阴天	雨天
晴天	0.4	0.5	0.1
阴天	0.3	0.4	0.3
雨天	0.1	0.5	0.4

表2：状态转移概率分布

马尔科夫模型

我们使用$\mathcal{S} = \{s_1, \ldots, s_N\}$表示状态集合。在上面的例子里，$N = 3$，$s_1 = \text{sunny}$，$s_2 = \text{cloudy}$，$s_3 = \text{rainy}$。因此，观测状态序列可以表示为$\mathbf{x} = x_1, \ldots, x_t, \ldots, x_T$，其中$x_t \in \mathcal{S}$。

可以使用马尔科夫模型计算观测状态序列的概率：

$$\begin{aligned} P(\mathbf{x}; \boldsymbol{\theta}) &= \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1}) \\ &\approx p(x_1) \times \prod_{t=2}^{T} p(x_t | x_{t-1}) \end{aligned}$$

其中，$\boldsymbol{\theta} = \{p(x) | x \in \mathcal{S}\} \cup \{p(x' | x) | x, x' \in \mathcal{S}\}$是模型参数。为了简化计算，假设当前状态$x_t$的生成只依赖于前一个状态$x_{t-1}$，通常称为二阶马尔科夫模型。

状态初始化概率

为了简便，我们将状态初始化概率定义为

$$\pi_i = p(x_1 = s_i), 1 \leq i \leq N$$

状态初始化概率满足以下性质：

① 非负性：$\pi_i \geq 0$

② 归一性：$\sum_{i=1}^{N} \pi_i = 1$

例如，在表1中，$\pi_1 = 0.4$，$\pi_2 = 0.3$，$\pi_3 = 0.3$。

晴天	阴天	雨天
0.4	0.3	0.3

表1：状态初始化概率分布

状态转移概率

为了简便，我们将状态转移概率定义为

$$a_{ij} = p(x_t = s_j | x_{t-1} = s_i), 1 \le i, j \le N$$

状态初始化概率满足以下性质：

① 非负性：$a_{ij} \ge 0$

② 归一性：$\sum_{j=1}^{N} a_{ij} = 1$

例如，在表2中，$a_{12} = 0.5$。

	晴天	阴天	雨天
晴天	0.4	0.5	0.1
阴天	0.3	0.4	0.3
雨天	0.1	0.5	0.4

表2：状态转移概率分布

状态序列概率计算

因此，连续四天下雨的概率可以计算如下：

$$
\begin{aligned}
& P(\text{rainy}, \text{rainy}, \text{rainy}, \text{rainy}; \boldsymbol{\theta}) \\
= & P(x_1 = s_3, x_2 = s_3, x_3 = s_3, x_4 = s_3; \boldsymbol{\theta}) \\
= & p(x_1 = s_3) \times p(x_2 = s_3 \,|\, x_1 = s_3) \times \\
& p(x_3 = s_3 \,|\, x_2 = s_3) \times p(x_4 = s_3 \,|\, x_3 = s_3) \\
= & \pi_3 \times a_{33} \times a_{33} \times a_{33} \\
= & 0.3 \times 0.4 \times 0.4 \times 0.4 \\
= & 0.0192
\end{aligned}
$$

如何利用积累的历史观测数据自动估计模型参数（即状态初始化概率与状态转移概率）？

基于训练数据的参数估计

假定历史观测数据包含三个序列：

① 晴天，阴天，雨天

② 晴天，阴天，晴天

③ 雨天，阴天

从直觉上，我们感觉可以使用相对频度来计算状态初始化概率，即第一天是某个状态的序列个数除以总共的序列数。例如，第一天是晴天的序列共有2个，总共有3个序列。因此，晴天的初始化概率为$\pi_1 = 2/3$。

类似地，我们也可以用相对频度计算状态转移概率。由于雨天出现在阴天之后的次数为1，而阴天之后出现天气的总次数为2，因此，从阴天到雨天的状态概率为$a_{23} = 1/2$。

那么，使用相对频度是否具备数学意义上的合理性呢？

极大似然估计

给定包含D个样本的训练数据集$\mathscr{D} = \left\{\mathbf{x}^{(d)}\right\}_{d=1}^{D}$，我们使用极大似然估计来从训练数据中自动获取最优模型参数：

$$\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}}\{L(\boldsymbol{\theta})\}$$

似然函数$L(\boldsymbol{\theta})$通常以对数的形式进行定义：

$$L(\boldsymbol{\theta}) = \sum_{d=1}^{D} \log P(\mathbf{x}^{(d)}; \boldsymbol{\theta})$$

$$= \sum_{d=1}^{D} \left(\log p\left(x_1^{(d)}\right) + \sum_{t=2}^{T^{(d)}} \log p\left(x_t^{(d)} | x_{t-1}^{(d)}\right) \right)$$

其中，$T^{(d)}$表示第d个序列$\mathbf{x}^{(d)}$的长度，$x_t^{(d)}$表示$\mathbf{x}^{(d)}$中的第d个状态。

约束条件下求极值

需要注意的是，模型参数需要满足以下两个约束条件：

$$\sum_{x \in \mathcal{S}} p(x) = 1$$

$$\forall x: \sum_{x' \in \mathcal{S}} p(x'|x) = 1$$

因此，我们需要引入拉格朗日乘子法实现约束条件下的极值求解：

$$J(\boldsymbol{\theta}, \lambda, \boldsymbol{\gamma}) = L(\boldsymbol{\theta}) - \lambda\left(\sum_{x \in \mathcal{S}} p(x) - 1\right) - \sum_{x \in \mathcal{S}} \gamma_x \left(\sum_{x' \in \mathcal{S}} p(x'|x) - 1\right)$$

其中，λ是与状态初始化概率约束相关的拉格朗日乘子，$\boldsymbol{\gamma} = \{\gamma_x | x \in \mathcal{S}\}$是与状态转移概率约束相关的拉格朗日乘子的集合，集合中的元素γ_x与状态x相对应。

估计状态初始化概率

首先可以计算状态初始概率$p(x)$的偏导

$$\frac{\partial J(\boldsymbol{\theta}, \lambda, \boldsymbol{\gamma})}{\partial p(x)} = \frac{\partial}{\partial p(x)} \sum_{d=1}^{D} \log p\left(x_1^{(d)}\right) - \lambda$$

$$= \frac{1}{p(x)} \sum_{d=1}^{D} \delta\left(x_1^{(d)}, x\right) - \lambda \qquad \text{（公式1）}$$

其中，$\delta(a, b)$的取值当$a = b$时为1，否则为0。

为了简便起见，我们用$c(x, \mathscr{D})$表示训练数据中第一个状态是x的次数：

$$c(x, \mathscr{D}) \equiv \sum_{d=1}^{D} \delta\left(x_1^{(d)}, x\right) \qquad \text{（公式2）}$$

估计状态初始化概率

然后计算拉格朗日乘子λ的偏导

$$\frac{\partial J(\boldsymbol{\theta}, \lambda, \boldsymbol{\gamma})}{\partial \lambda} = \sum_{x \in \mathcal{S}} p(x) - 1 \qquad \text{（公式3）}$$

令公式1的取值为0，根据公式2，可以得到

$$p(x) = \frac{c(x, \mathcal{D})}{\lambda} \qquad \text{（公式4）}$$

令公式3的取值为0，根据公式4，可以得到

$$\lambda = \sum_{x \in \mathcal{S}} c(x, \mathcal{D}) \qquad \text{（公式5）}$$

因此，根据公式4和公式5，我们可以得到状态初始化概率的估计公式

$$p(x) = \frac{c(x, \mathcal{D})}{\sum_{x' \in \mathcal{S}} c(x', \mathcal{D})} \qquad \text{（公式6）}$$

估计状态转移概率

类似地，我们首先计算状态转移概率的偏导：

$$\frac{J(\boldsymbol{\theta},\lambda,\boldsymbol{\gamma})}{\partial p(x'|x)} = \frac{\partial}{\partial p(x'|x)} \sum_{d=1}^{D} \sum_{t=2}^{T^{(d)}} \log p\left(x_t^{(d)} | x_{t-1}^{(d)}\right) - \gamma_x$$

$$= \frac{1}{p(x'|x)} \sum_{d=1}^{D} \sum_{t=2}^{T^{(d)}} \delta(x_{t-1}^{(d)}, x)\delta(x_t^{(d)}, x') - \gamma_x \qquad \text{（公式7）}$$

为了简便起见，我们将训练数据中x'紧跟着出现在x后面的次数表示为

$$c(x, x', \mathscr{D}) \equiv \sum_{d=1}^{D} \sum_{t=2}^{T^{(d)}} \delta(x_{t-1}^{(d)}, x)\delta(x_t^{(d)}, x') \qquad \text{（公式8）}$$

由此可见，统计模型在估计参数时的一个基本操作是“统计次数”。

估计状态转移概率

然后，可以计算拉格朗日乘子γ_x的偏导：

$$\frac{\partial J(\boldsymbol{\theta}, \lambda, \boldsymbol{\gamma})}{\partial \gamma_x} = \sum_{x' \in \mathcal{S}} p(x' | x) - 1 \qquad \text{（公式9）}$$

令公式7的取值为0，根据公式8，可以得到

$$p(x' | x) = \frac{c(x, x', \mathcal{D})}{\gamma_x} \qquad \text{（公式10）}$$

令公式9的取值为0，根据公式10，可以得到

$$\gamma_x = \sum_{x'' \in \mathcal{S}} c(x, x'', \mathcal{D}) \qquad \text{（公式11）}$$

因此，根据公式10和公式11，我们可以得到状态转移概率的估计公式

$$p(x' | x) = \frac{c(x, x', \mathcal{D})}{\sum_{x'' \in \mathcal{S}} c(x, x'', \mathcal{D})} \qquad \text{（公式12）}$$

参数估计公式

综上所述，给定训练集$\mathscr{D} = \left\{\mathbf{x}^{(d)}\right\}_{d=1}^{D}$，状态初始化概率的估计公式如下：

$$\pi_i = \frac{\sum_{d=1}^{D} \delta(x_1^{(d)}, s_i)}{D} \qquad \text{（公式13）}$$

状态转换概率的估计公式为

$$a_{ij} = \frac{\sum_{d=1}^{D} \sum_{t=2}^{T^{(d)}} \delta(x_{t-1}^{(d)}, s_i)\delta(x_t^{(d)}, s_j)}{\sum_{d=1}^{D} \sum_{t=2}^{T^{(d)}} \delta(x_{t-1}^{(d)}, s_i)} \qquad \text{（公式14）}$$

参数估计示例

假定历史观测数据包含三个序列：

① 晴天，阴天，雨天

② 晴天，阴天，晴天

③ 雨天，阴天

从中可以估计得到如表3和表4所示的模型参数。

晴天	阴天	雨天
2/3	0/3	1/3

表3：状态初始化概率分布

	晴天	阴天	雨天
晴天	0/2	2/2	0/2
阴天	1/2	0/2	1/2
雨天	0/1	1/1	0/1

表4：状态转移概率分布

内容提要

马尔科夫模型

计算观测概率

Viterbi算法

前向后向算法

隐状态与观测状态

假定一个囚徒被关在暗无天日的牢房里，不知道外面的天气（如晴天、阴天、雨天），但是可以通过触摸感受到地面的潮湿程度（如干燥、潮湿），从而推测天气情况。在这个例子里，地面潮湿程度是观测状态，外面的天气是隐状态。

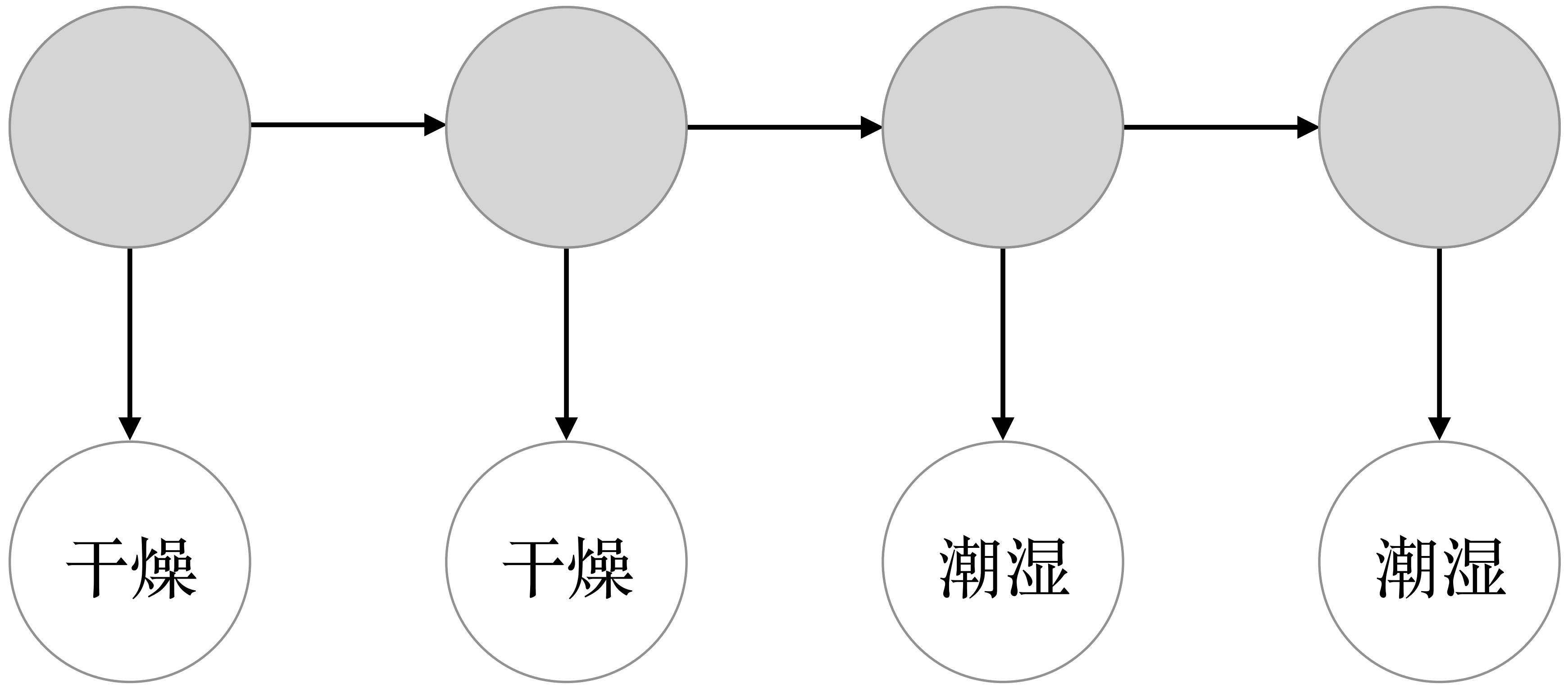

图1：隐状态与观测状态。

隐状态与观测状态

假定一个囚徒被关在暗无天日的牢房里，不知道外面的天气（如晴天、阴天、雨天），但是可以通过触摸感受到地面的潮湿程度（如干燥、潮湿），从而推测天气情况。在这个例子里，地面潮湿程度是观测状态，外面的天气是隐状态。

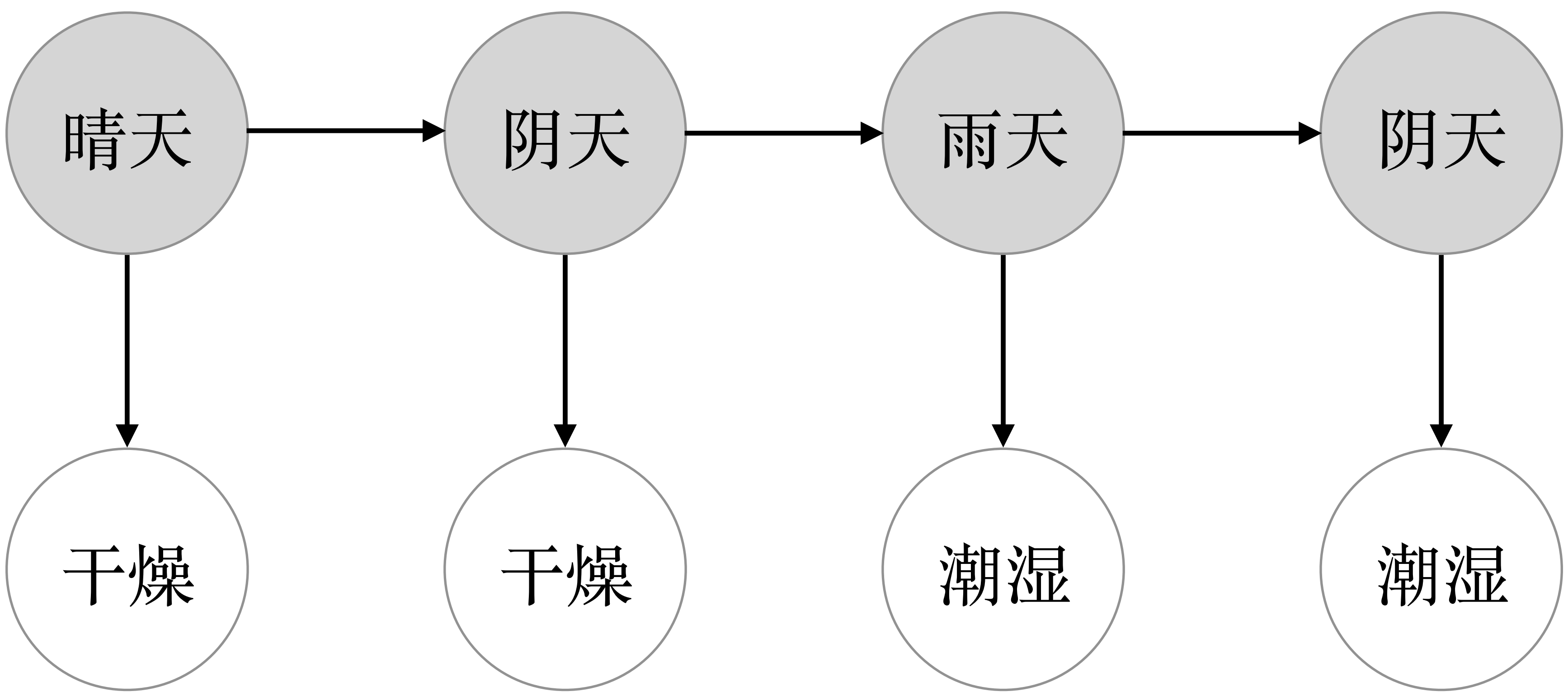

图1：隐状态与观测状态。

隐状态与观测状态

假定观测状态的生成取决于隐状态。例如，如果是晴天，那么地面很有可能是干燥的。在图1所示的例子中，箭头表示节点之间的依赖关系，即：必须首先生成第一个隐状态“晴天”，才能生成第一个观测状态“干燥”以及第二个隐状态“阴天”。

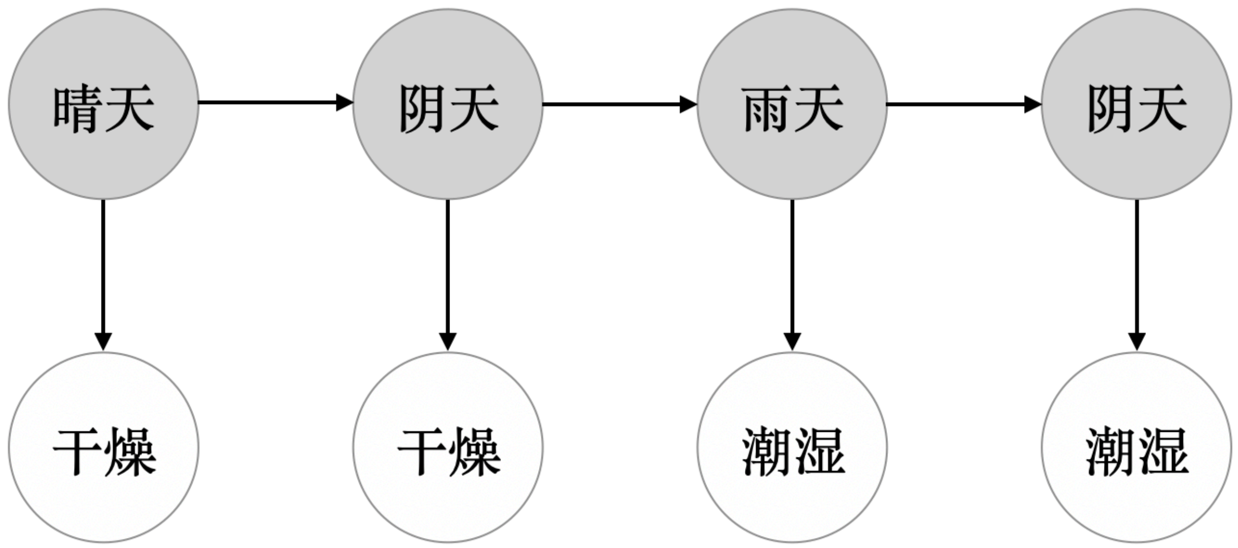

图1：隐状态与观测状态。

	干燥	潮湿
晴天	0.8	0.2
阴天	0.6	0.4
雨天	0.3	0.7

表5：观测状态生成概率分布

隐马尔科夫模型

我们使用$\mathcal{O} = \{o_1, \ldots, o_m\}$表示观测状态集合（例如，“干燥”和“潮湿”），$\mathcal{S} = \{s_1, \ldots, s_N\}$表示隐状态集合（例如，“晴天”、“阴天”和“雨天”），$\mathbf{x} = x_1, \ldots, x_t, \ldots, x_T$表示观测状态序列，$\mathbf{z} = z_1, \ldots, z_t, \ldots, z_T$表示隐状态序列。形式化地，一个二阶隐马尔科夫模型可以表示如下：

$$
\begin{aligned}
P(\mathbf{x}; \boldsymbol{\theta}) &= \sum_{\mathbf{z}} P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \\
&= \sum_{\mathbf{z}} p(z_1) \times p(x_1 \,|\, z_1) \times \prod_{t=2}^{T} p(z_t \,|\, z_{t-1}) \times p(x_t \,|\, z_t)
\end{aligned}
$$

其中，$\boldsymbol{\theta} = \{p(z) \,|\, z \in \mathcal{S}\} \cup \{p(z' \,|\, z) \,|\, z, z' \in \mathcal{S}\} \cup \{p(x \,|\, z) \,|\, x \in \mathcal{O} \wedge z \in \mathcal{S}\}$是模型参数，分别对应隐状态初始化概率、隐状态转移概率和观测状态生成概率。

隐状态初始化概率

为了简便，我们将隐状态初始化概率表示为

$$\pi_i = p(z_1 = s_i), 1 \le i \le N$$

满足以下性质：

① 非负性：$\pi_i \ge 0$。

② 归一性：$\sum_{i=1}^{N} \pi_i = 1$。

晴天	阴天	雨天
0.4	0.3	0.3

表6：隐状态初始化概率分布

隐状态转移概率

为了简便，我们将隐状态转移概率表示为

$$a_{ij} = p(z_t = s_j \,|\, z_{t-1} = s_i), 1 \le i, j \le N$$

满足以下性质：

① 非负性：$a_{ij} \ge 0$。

② 归一性：$\sum_{j=1}^{N} a_{ij} = 1$。

	晴天	阴天	雨天
晴天	0.4	0.5	0.1
阴天	0.3	0.4	0.3
雨天	0.1	0.5	0.4

表7：隐状态转移概率分布

观测状态生成概率

为了简便，我们将观测状态生成概率表示为

$$b_j(k) = p(x_t = o_k | z_t = s_j), 1 \leq j \leq N \wedge 1 \leq k \leq M$$

满足以下性质：

① 非负性：$b_j(k) \geq 0$。

② 归一性：$\sum_{k=1}^{M} b_j(k) = 1$。

	干燥	潮湿
晴天	0.8	0.2
阴天	0.6	0.4
雨天	0.3	0.7

表5：观测状态生成概率分布

概率图模型

隐马尔科夫模型是一个典型的概率图模型。作为有向图，边的箭头决定了节点生成的先后顺序：

$$P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = p(z_1) \times p(x_1 | z_1) \times \prod_{t=2}^{T} p(z_t | z_{t-1}) \times p(x_t | z_t)$$

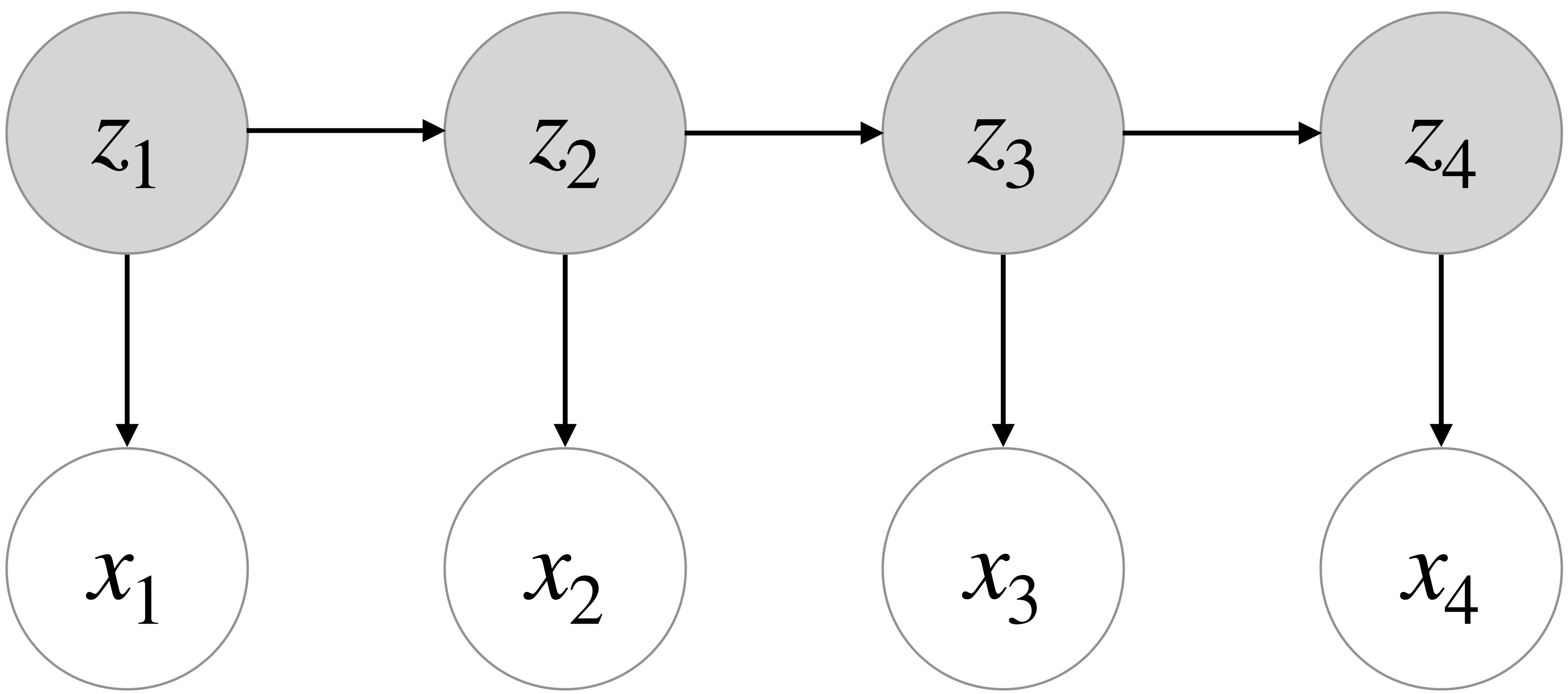

图2：隐马尔科夫模型是一个有向图。

三个基本问题

隐马尔科夫模型有三个基本问题：

观测状态序列概率计算
给定模型参数，如何计算一个观测状态序列的概率？

最优隐状态序列计算
给定模型参数和观测状态序列，如何计算最可能的隐状态序列？

模型参数估计
给定观测序列集合，如何从中估计模型参数？

观测状态序列概率计算

问题描述：给定模型参数，计算观测状态序列概率。

$$P(\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{z}} p(z_1) \times p(x_1 | z_1) \times \prod_{t=2}^{T} p(z_t | z_{t-1}) \times p(x_t | z_t)$$

主要的挑战在于如何处理指数级空间：隐状态序列$\mathbf{z}$的个数为N^T，当N和T非常大的时候，逐一枚举所有隐状态序列并计算求和是不现实的。

例如，假定观测序列为“干燥 潮湿”，则总共有$3^2 = 9$种隐状态序列：

$z_1 = (\text{sunny}, \text{sunny})$　　$z_4 = (\text{cloudy}, \text{sunny})$　　$z_7 = (\text{rainy}, \text{sunny})$

$z_2 = (\text{sunny}, \text{cloudy})$　　$z_5 = (\text{cloudy}, \text{cloudy})$　　$z_8 = (\text{rainy}, \text{cloudy})$

$z_3 = (\text{sunny}, \text{rainy})$　　$z_6 = (\text{cloudy}, \text{rainy})$　　$z_9 = (\text{rainy}, \text{rainy})$

搜索空间

隐马尔科夫模型的搜索空间可以组成一个有向图，其中节点表示隐状态，边表示隐状态之间的转移。为了简便，略去观测状态节点和相应的边，因为观测状态是固定不变的。例如，图3中的路径等价于

$$P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \pi_3 b_3(x_1) a_{32} b_2(x_2) a_{21} b_1(x_3) a_{13} b_3(x_4)$$

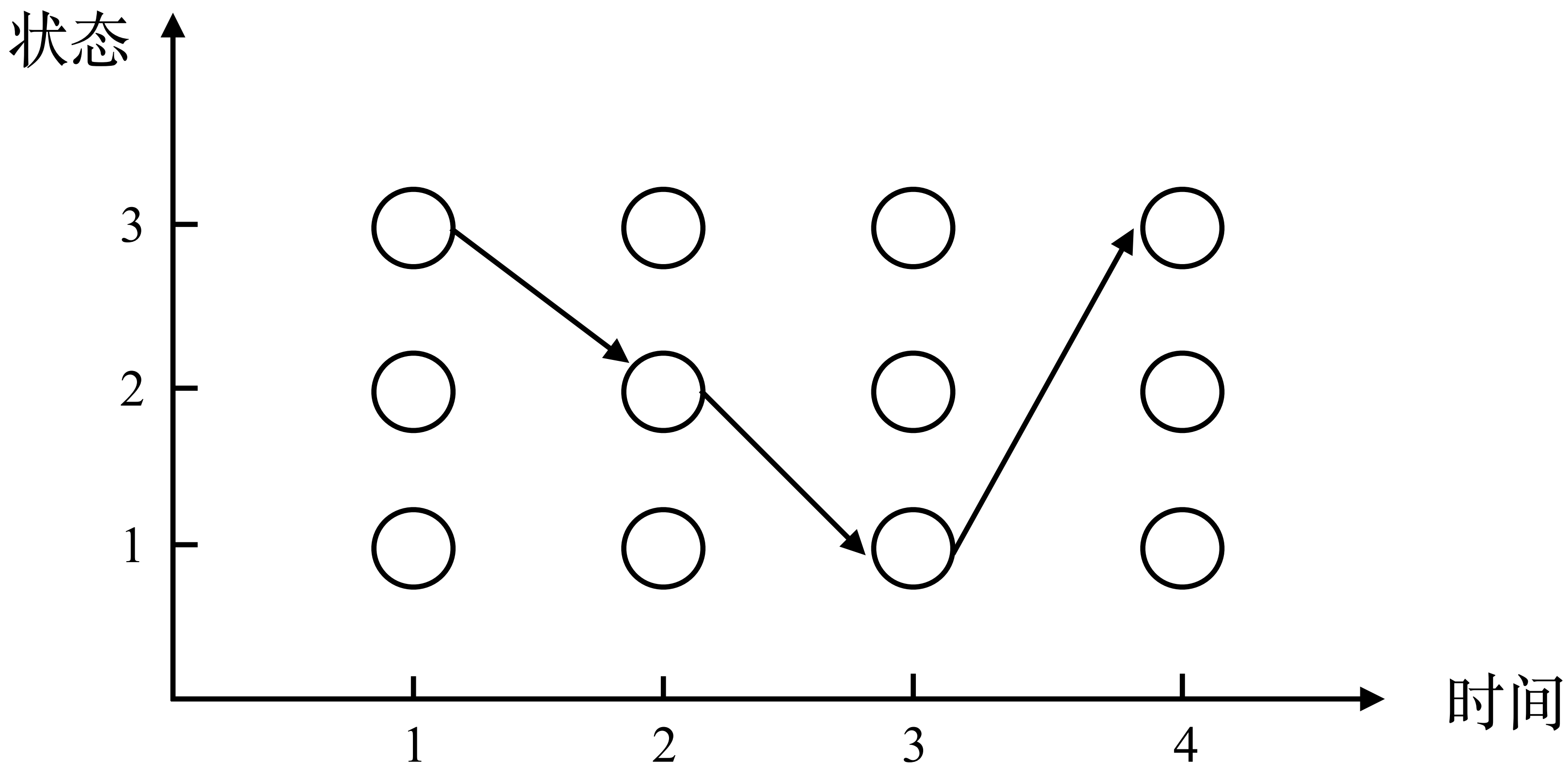

图3：隐马尔科夫模型的搜索空间。

路径求和

从图的角度来看，计算观测状态序列概率等价于计算所有路径之和。

$$P(\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{z}} P(\mathbf{x},\mathbf{z};\boldsymbol{\theta})$$

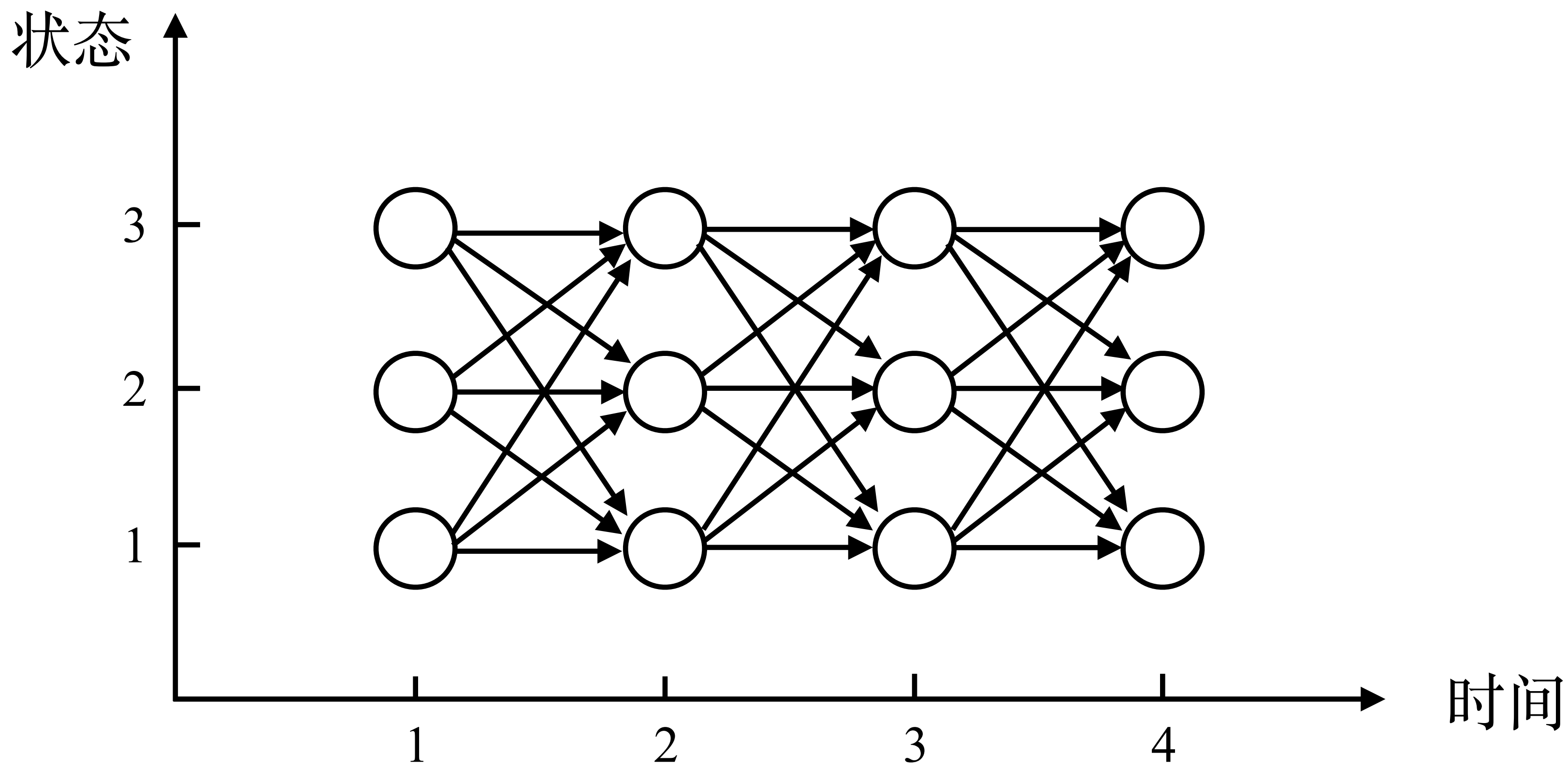

图4：观测状态概率等价于所有路径之和。

计算路径之和：全枚举

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的所有路径之和。

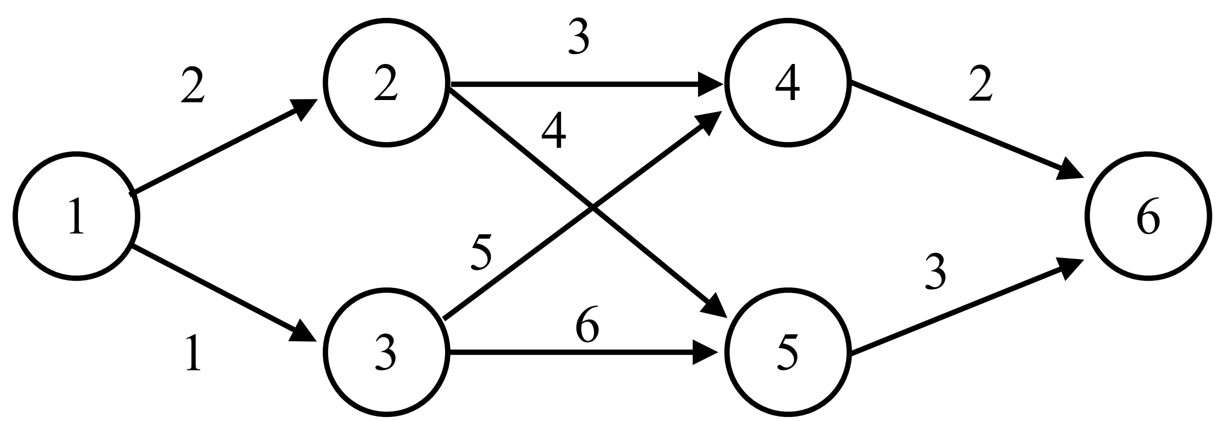

图5：计算路径之和的例子。

最直观的是枚举法，即枚举出所有的路径，计算其取值，再求和。

$$\left.\begin{array}{l}
① \text{ 路径} 1 \rightarrow 2 \rightarrow 4 \rightarrow 6\text{：} 2 \times 3 \times 2 = 12 \\
② \text{ 路径} 1 \rightarrow 2 \rightarrow 5 \rightarrow 6\text{：} 2 \times 4 \times 3 = 24 \\
③ \text{ 路径} 1 \rightarrow 3 \rightarrow 4 \rightarrow 6\text{：} 1 \times 5 \times 2 = 10 \\
④ \text{ 路径} 1 \rightarrow 3 \rightarrow 5 \rightarrow 6\text{：} 1 \times 6 \times 3 = 18
\end{array}\right\} \quad 64$$

计算路径之和：动态规划

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的所有路径之和。

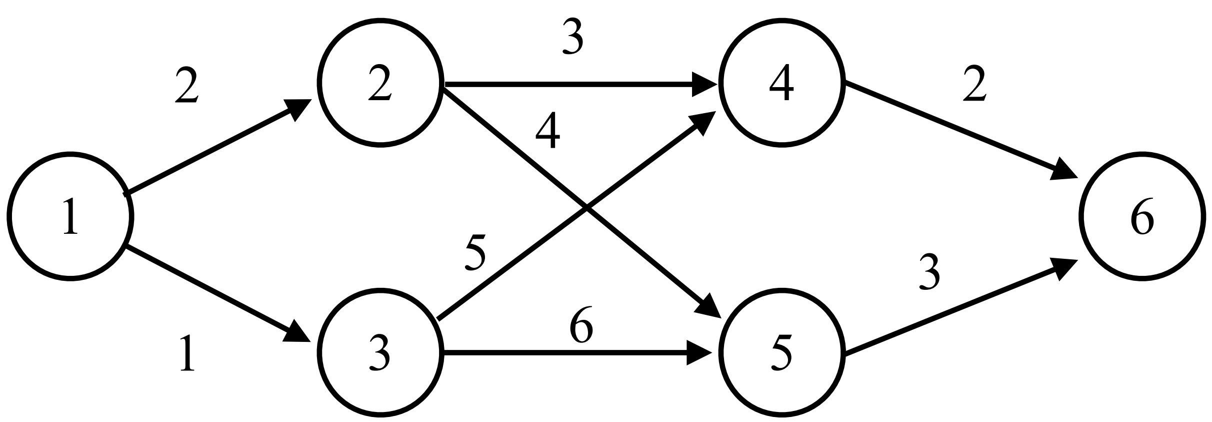

图5：计算路径之和的例子。

令α_i是从节点1到节点i（$i \in [1, 6]$）的路径之和，且规定$\alpha_1 = 1$，则节点i的路径之和可以递归计算：

$$\alpha_i = \sum_{j \in \mathrm{heads}(i)} w_{ji} \alpha_j$$

其中，w_{ji}是从节点j到节点i的连线上的权重，heads(i)表示节点i的所有前继。

计算路径之和：动态规划

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的所有路径之和。

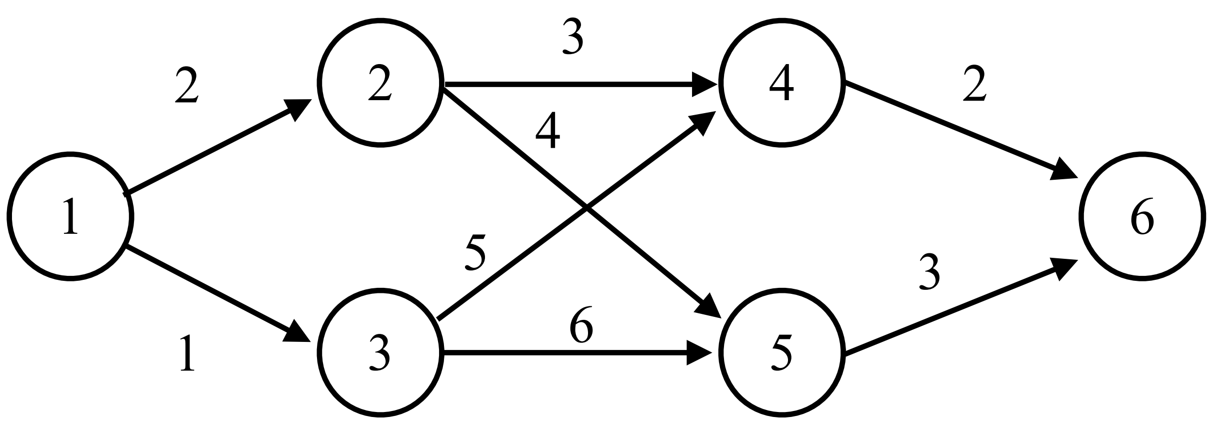

图5：计算路径之和的例子。

$$\alpha_i = \sum_{j \in \mathrm{heads}(i)} w_{ji}\alpha_j$$

i	1	2	3	4	5	6
α_i						

计算路径之和：动态规划

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的所有路径之和。

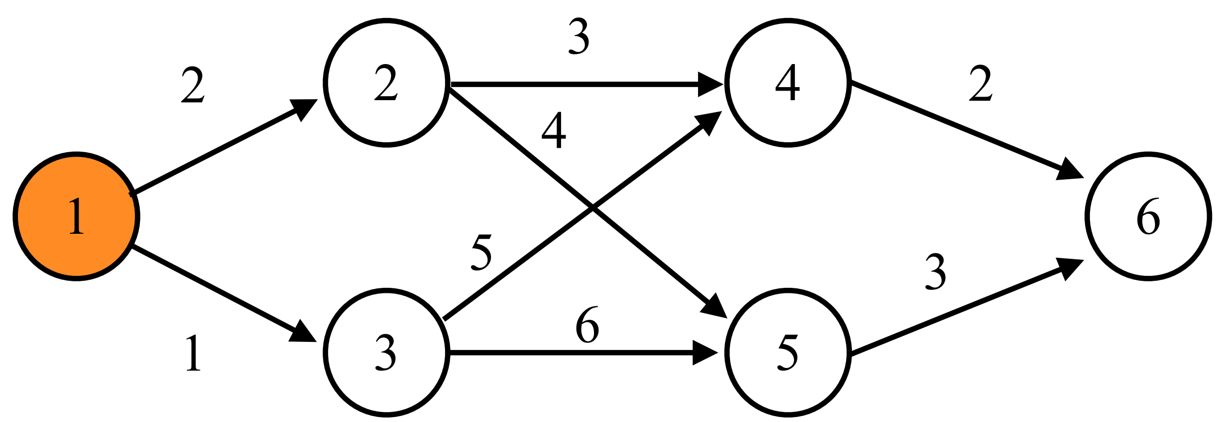

图5：计算路径之和的例子。

$$\alpha_i = \sum_{j \in \text{heads}(i)} w_{ji} \alpha_j$$

i	1	2	3	4	5	6
α_i	1					

$$\alpha_1 = 1$$

计算路径之和：动态规划

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的所有路径之和。

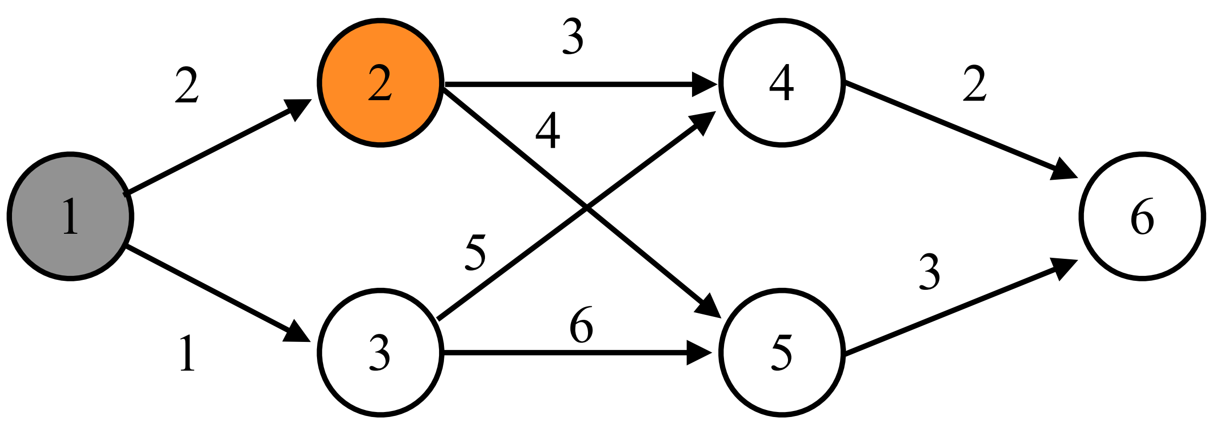

图5：计算路径之和的例子。

$$\alpha_i = \sum_{j \in \text{heads}(i)} w_{ji}\alpha_j$$

i	1	2	3	4	5	6
α_i	1	2				

$$\alpha_2 = w_{12}\alpha_1 = 2 \times 1 = 2$$

计算路径之和：动态规划

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的所有路径之和。

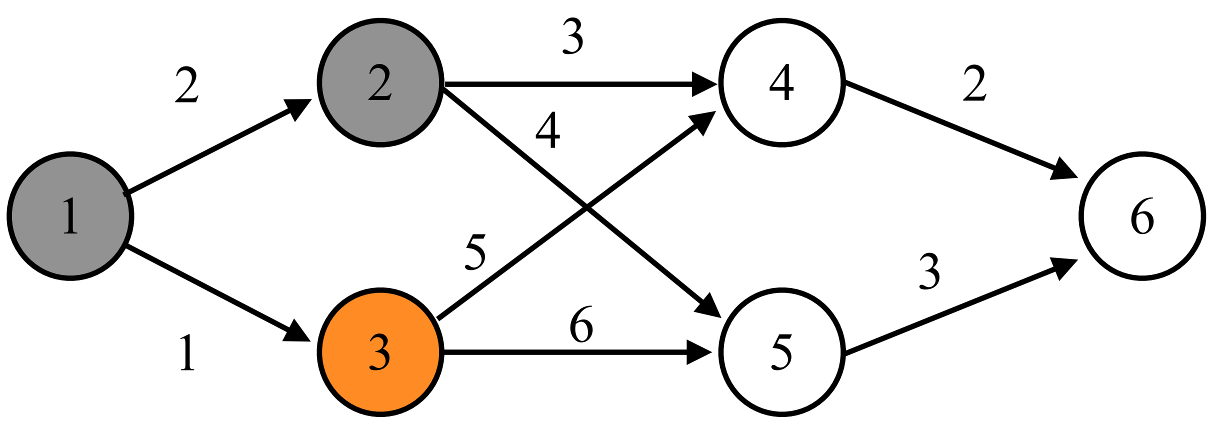

图5：计算路径之和的例子。

$$\alpha_i = \sum_{j \in \text{heads}(i)} w_{ji}\alpha_j$$

i	1	2	3	4	5	6
α_i	1	2	1			

$$\alpha_3 = w_{13}\alpha_1 = 1 \times 1 = 1$$

计算路径之和：动态规划

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的所有路径之和。

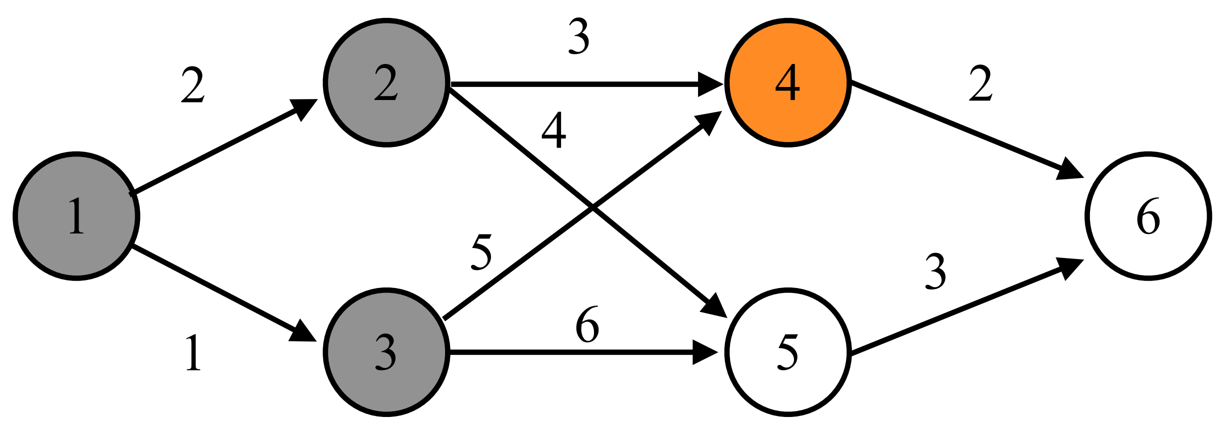

图5：计算路径之和的例子。

$$\alpha_i = \sum_{j \in \text{heads}(i)} w_{ji}\alpha_j$$

i	1	2	3	4	5	6
α_i	1	2	1	11		

$$\alpha_4 = w_{24}\alpha_2 + w_{34}\alpha_3 = 3 \times 2 + 5 \times 1 = 11$$

计算路径之和：动态规划

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的所有路径之和。

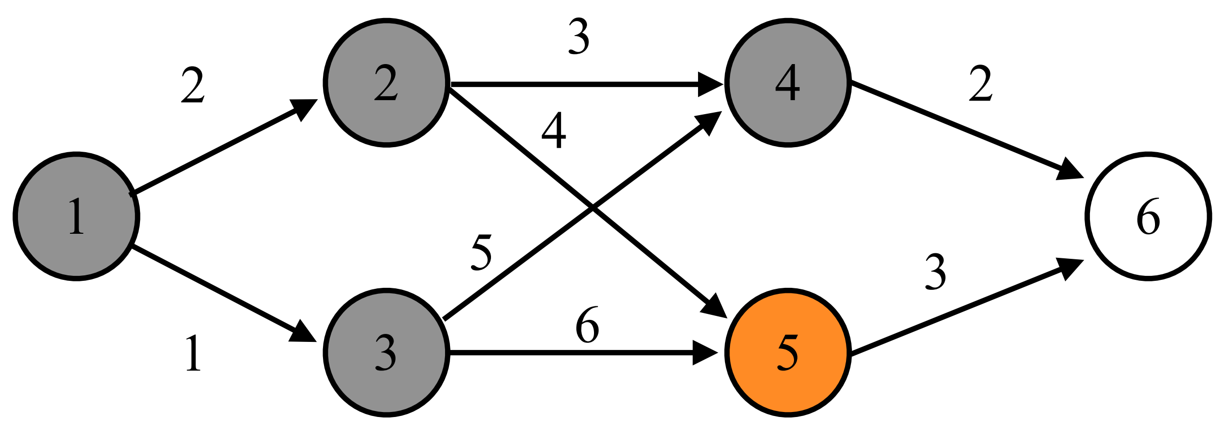

图5：计算路径之和的例子。

$$\alpha_i = \sum_{j \in \text{heads}(i)} w_{ji}\alpha_j$$

i	1	2	3	4	5	6
α_i	1	2	1	11	14	

$$\alpha_5 = w_{25}\alpha_2 + w_{35}\alpha_3 = 4 \times 2 + 6 \times 1 = 14$$

计算路径之和：动态规划

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的所有路径之和。

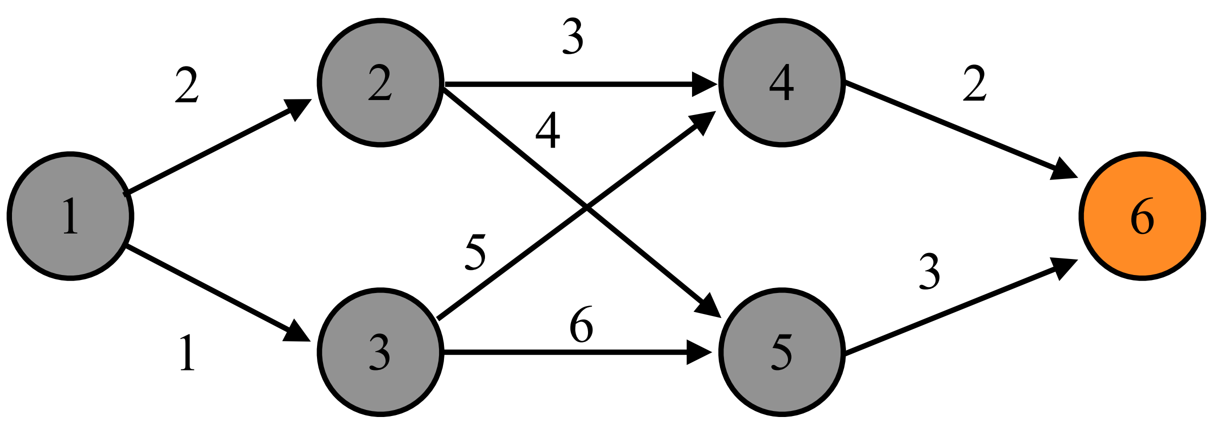

图5：计算路径之和的例子。

$$\alpha_i = \sum_{j \in \text{heads}(i)} w_{ji}\alpha_j$$

i	1	2	3	4	5	6
α_i	1	2	1	11	14	64

$$\alpha_6 = w_{46}\alpha_4 + w_{56}\alpha_5 = 2 \times 11 + 3 \times 14 = 64$$

前向概率

部分观测状态序列$x_1,\ldots,x_t$与第t个隐状态为s_i的联合概率称为前向概率，定义如下：

$$\alpha_t(i) = P(x_1,\ldots,x_t, z_t = s_i; \boldsymbol{\theta})$$

可使用动态规划算法递归计算如下：

(1) 初始化：$t = 1$

$$\alpha_1(i) = \pi_i b_i(x_1), 1 \leq i \leq N$$

(2) 递归：$t = 2,\ldots,T$

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

(3) 终止：

$$P(\mathbf{x}; \boldsymbol{\theta}) = \sum_{i=1}^{N} \alpha_T(i)$$

前向概率的图形化解释

一个节点的前向概率是从$t=1$时刻所有节点出发到达该节点的所有路径取值之和。

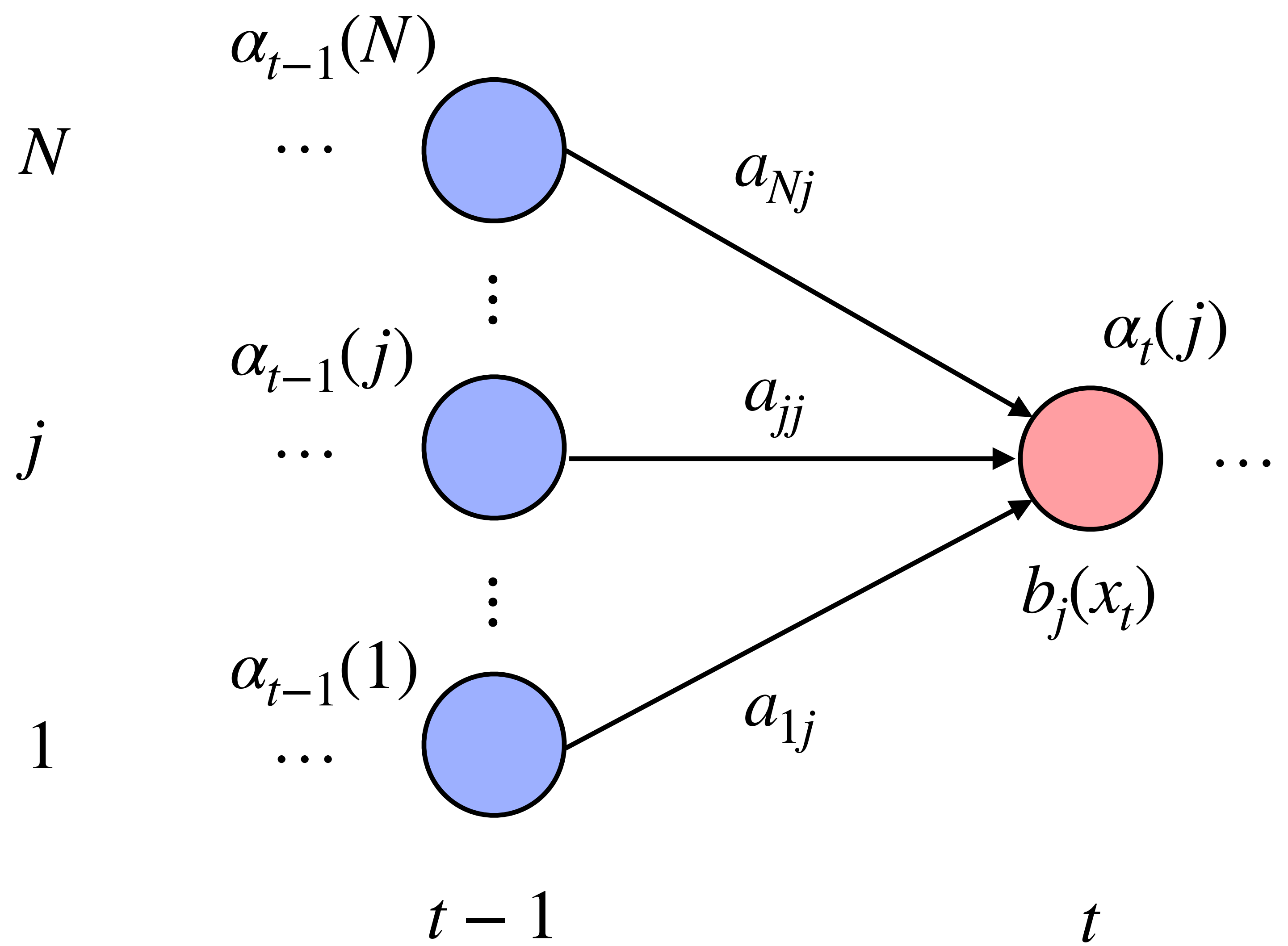

图6：前向概率的图形化解释。

计算前向概率

晴天	阴天	雨天
0.4	0.3	0.3

表6：隐状态初始化概率分布

	晴天	阴天	雨天
晴天	0.4	0.5	0.1
阴天	0.3	0.4	0.3
雨天	0.1	0.5	0.4

表7：隐状态转移概率分布

	干燥	潮湿
晴天	0.8	0.2
阴天	0.6	0.4
雨天	0.3	0.7

表5：观测状态生成概率分布

计算前向概率

给定模型参数，计算观测状态序列“干燥、潮湿、干燥”的前向概率。

	$t=1$	$t=2$	$t=3$
晴天	0.3200	0.0382	0.0456
阴天	0.1800	0.1108	0.0637
雨天	0.0900	0.0854	0.0214

表8：前向概率计算。

根据表8，可知观测状态序列“干燥、潮湿、干燥”的概率为

$$P(\mathbf{x};\boldsymbol{\theta}) = 0.0456 + 0.0637 + 0.0214$$
$$= 0.1307$$

所有观测状态序列概率

x	$P(\mathbf{x};\boldsymbol{\theta})$
干燥、干燥、干燥	0.2124
干燥、干燥、潮湿	0.1432
潮湿、干燥、干燥	0.1315
干燥、潮湿、干燥	0.1307
干燥、潮湿、潮湿	0.1037
潮湿、潮湿、干燥	0.1004
潮湿、干燥、潮湿	0.0939
潮湿、潮湿、潮湿	0.0842

表9：所有观测状态序列概率。

后向计算路径之和

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的所有路径之和。

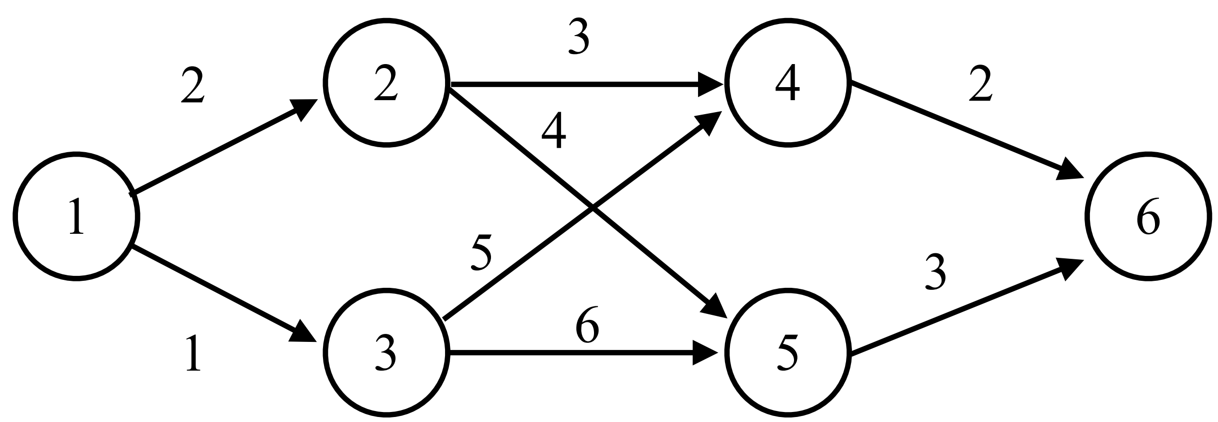

图7：后向计算路径之和的例子。

令β_i是从节点6到节点i（$i \in [1, 6]$）的后向路径之和，且规定$\beta_6 = 1$，则节点i的后向路径之和可以递归计算：

$$\beta_i = \sum_{j \in \mathrm{tails}(i)} w_{ij}\beta_j$$

其中，w_{ij}是从节点i到节点j的连线上的权重，tails(i)表示节点i的所有后继。

后向计算路径之和

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的所有路径之和。

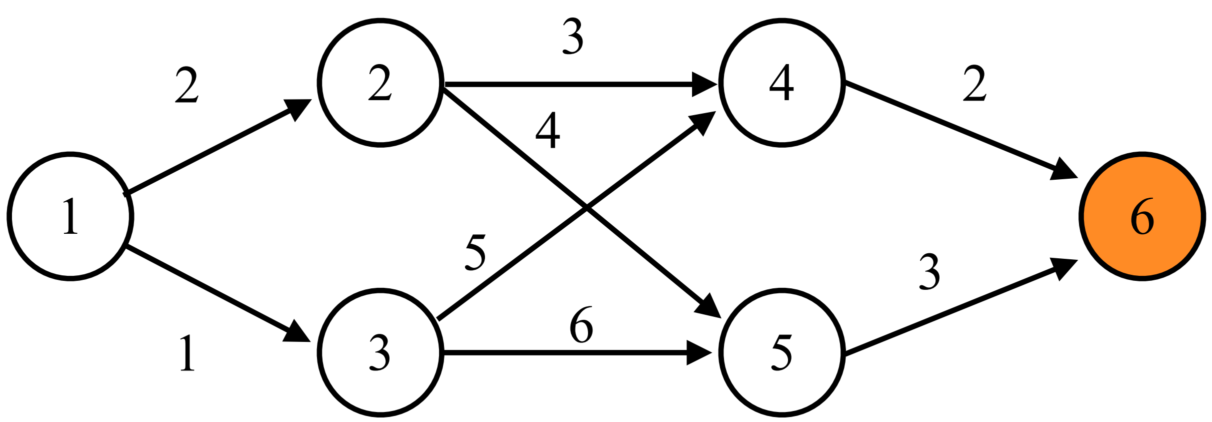

图7：后向计算路径之和的例子。

$$\beta_i = \sum_{j \in \text{tails}(i)} w_{ij}\beta_j$$

i	6	5	4	3	2	1
β_i	1					

$\beta_6 = 1$

后向计算路径之和

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的所有路径之和。

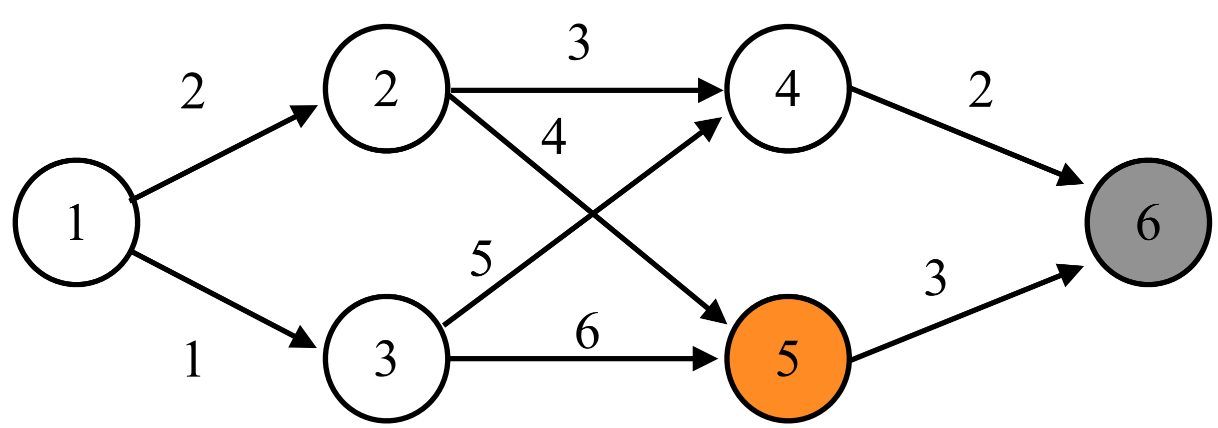

图7：后向计算路径之和的例子。

$$\beta_i = \sum_{j \in \mathrm{tails}(i)} w_{ij}\beta_j$$

i	6	5	4	3	2	1
β_i	1	3				

$$\beta_5 = w_{56}\beta_6 = 3 \times 1 = 3$$

后向计算路径之和

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的所有路径之和。

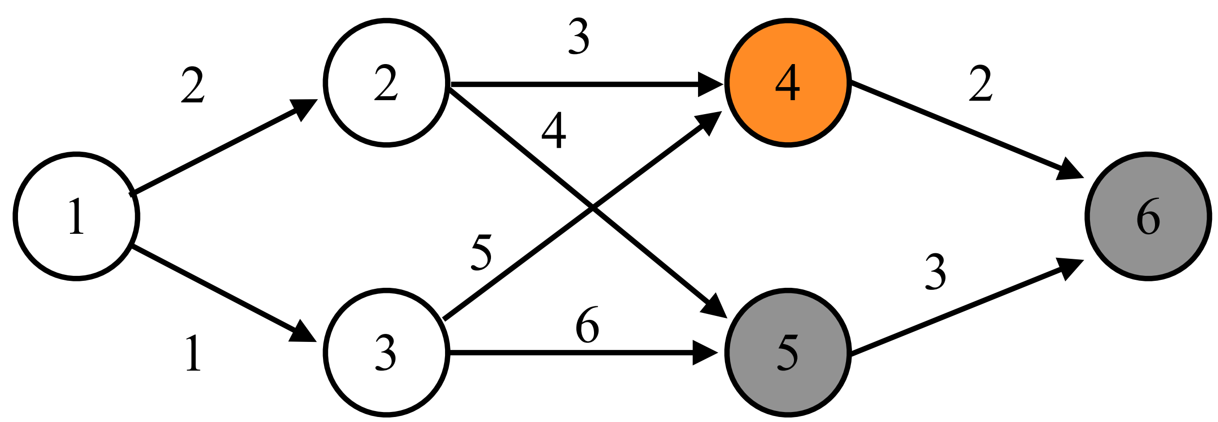

图7：后向计算路径之和的例子。

$$\beta_i = \sum_{j \in \mathrm{tails}(i)} w_{ij}\beta_j$$

i	6	5	4	3	2	1
β_i	1	3	2			

$$\beta_4 = w_{46}\beta_6 = 2 \times 1 = 2$$

后向计算路径之和

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的所有路径之和。

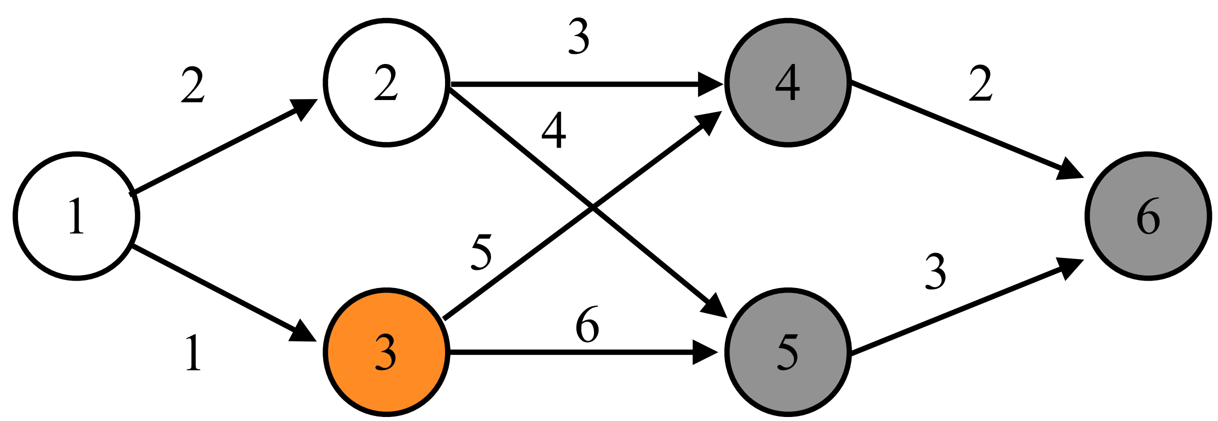

图7：后向计算路径之和的例子。

$$\beta_i = \sum_{j \in \text{tails}(i)} w_{ij}\beta_j$$

i	6	5	4	3	2	1
β_i	1	3	2	28		

$$\beta_3 = w_{34}\beta_4 + w_{35}\beta_5 = 5 \times 2 + 6 \times 3 = 28$$

后向计算路径之和

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的所有路径之和。

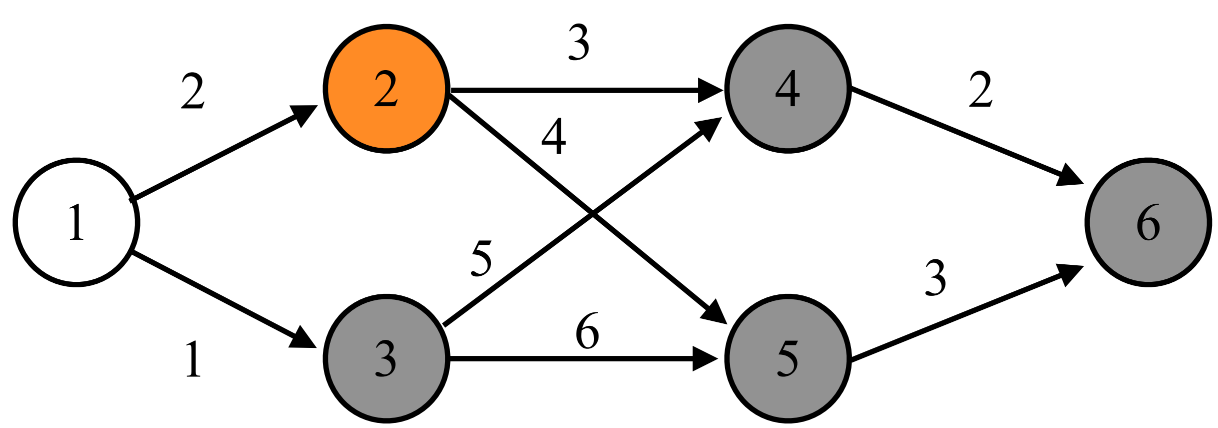

图7：后向计算路径之和的例子。

$$\beta_i = \sum_{j \in \text{tails}(i)} w_{ij}\beta_j$$

i	6	5	4	3	2	1
β_i	1	3	2	28	18	

$$\beta_2 = w_{24}\beta_4 + w_{25}\beta_5 = 3 \times 2 + 4 \times 3 = 18$$

后向计算路径之和

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的所有路径之和。

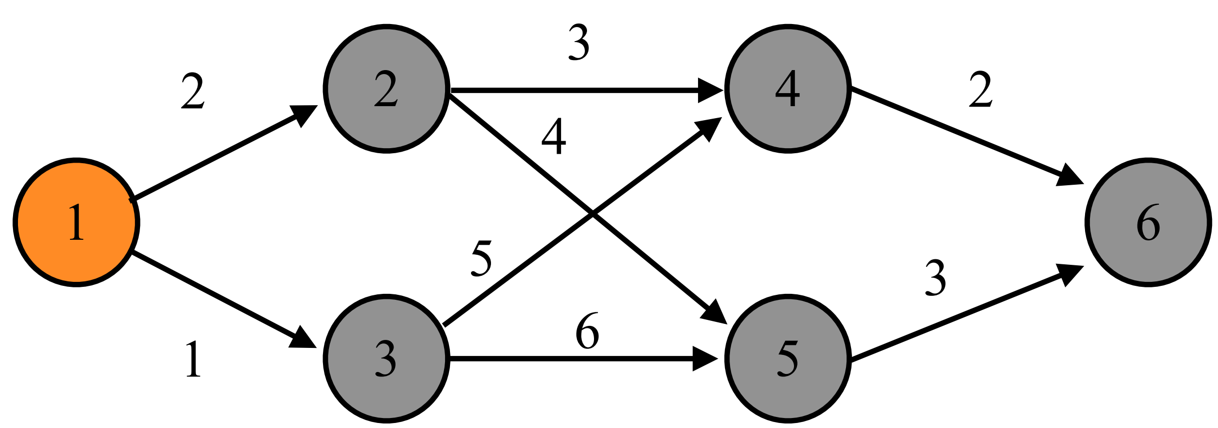

图7：后向计算路径之和的例子。

$$\beta_i = \sum_{j \in \text{tails}(i)} w_{ij}\beta_j$$

i	6	5	4	3	2	1
β_i	1	3	2	28	18	64

$$\beta_1 = w_{12}\beta_2 + w_{13}\beta_3 = 2 \times 18 + 1 \times 28 = 64$$

后向概率

第t个隐状态为s_i生成部分观测状态序列$x_{t+1}, \ldots, x_T$的条件概率称为后向概率，定义为：

$$\beta_t(i) = P(x_{t+1}, \ldots, x_T \,|\, z_t = s_i; \boldsymbol{\theta})$$

可使用动态规划算法递归计算如下：

(1) 初始化：$t = T$

$$\beta_T(i) = 1,\ 1 \leq i \leq N$$

(2) 递归：$t = T - 1, \ldots, 1$

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

(3) 终止：

$$P(\mathbf{x}; \boldsymbol{\theta}) = \sum_{i=1}^{N} \pi_i b_i(x_1) \beta_1(i)$$

后向概率的图形化解释

一个节点的后向概率是从$t = T$时刻所有节点反向出发到达该节点的所有路径取值之和（但不包含该节点的观测状态生成概率）。

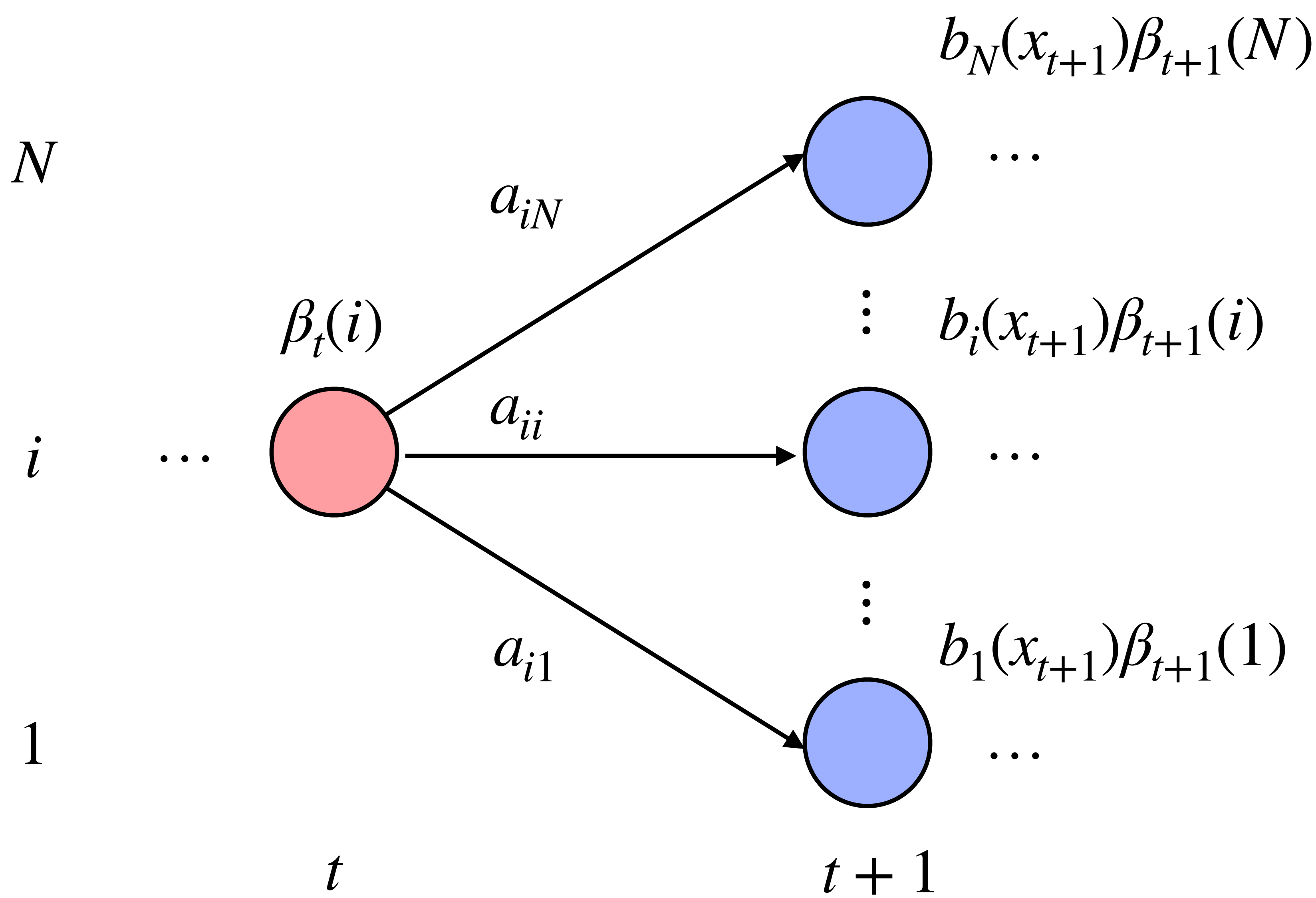

图8：后向概率的图形化解释。

计算后向概率

给定模型参数，计算观测状态序列“干燥、潮湿、干燥”的后向概率。

	$t=1$	$t=2$	$t=3$
晴天	0.2010	0.6500	1.0000
阴天	0.2352	0.5700	1.0000
雨天	0.2670	0.5000	1.0000

表10：后向概率计算。

根据表10，可知观测状态序列“干燥、潮湿、干燥”的概率为

$$
\begin{aligned}
P(\mathbf{x};\boldsymbol{\theta}) &= 0.4\times0.8\times0.2010+0.3\times0.6\times0.2352+0.3\times0.3\times0.2670 \\
&= 0.1307
\end{aligned}
$$

内容提要

马尔科夫模型

计算观测概率

Viterbi算法

前向后向算法

计算最优隐状态序列

给定一个观测状态序列$\mathbf{x} = x_1, \ldots, x_t, \ldots, x_T$和模型参数$\boldsymbol{\theta}$，最优隐状态序列可计算如下：

$$
\begin{aligned}
\hat{\mathbf{z}} &= \mathrm{argmax}_{\mathbf{z}} \left\{ P(\mathbf{z} \,|\, \mathbf{x}; \boldsymbol{\theta}) \right\} \\
&= \mathrm{argmax}_{\mathbf{z}} \left\{ \frac{P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})}{P(\mathbf{x}; \boldsymbol{\theta})} \right\} \\
&= \mathrm{argmax}_{\mathbf{z}} \left\{ P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \right\} \\
&= \mathrm{argmax}_{\mathbf{z}} \left\{ p(z_1) \times p(x_1 \,|\, z_1) \times \prod_{t=2}^{T} p(z_t \,|\, z_{t-1}) \times p(x_t \,|\, z_t) \right\}
\end{aligned}
$$

计算最大路径

从图的角度来看，计算最优隐状态序列概率等价于计算最大路径。

$$P(\mathbf{x};\boldsymbol{\theta}) = \mathrm{argmax}_{\mathbf{z}}\left\{ P(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) \right\}$$

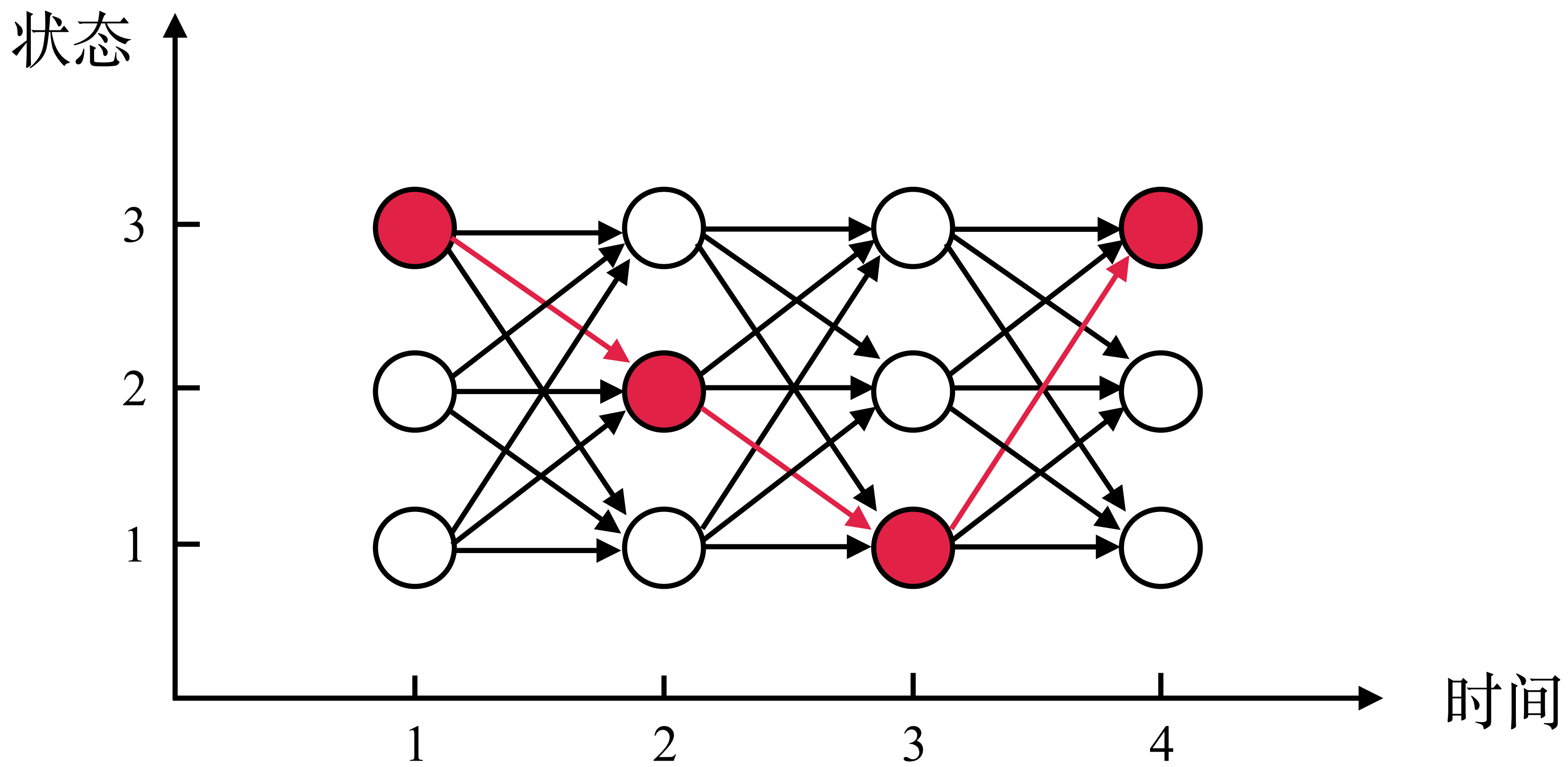

图9：计算最优状态状态概率等价于计算最大路径。

计算最大路径：全枚举

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的最大路径。

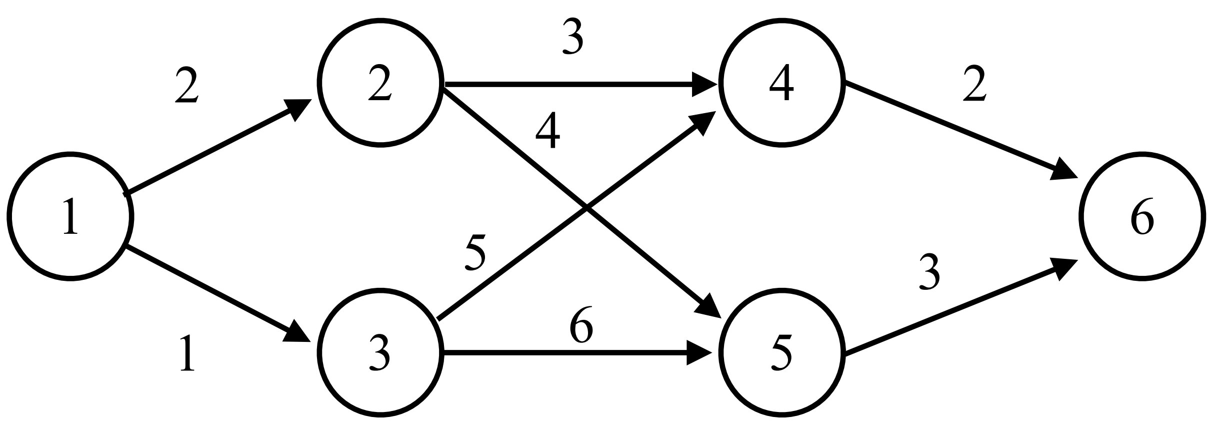

图10：计算最大路径的例子。

最直观的是枚举法，即枚举出所有的路径，计算其取值，再求最大路径。

① 路径$1 \rightarrow 2 \rightarrow 4 \rightarrow 6$：$2 \times 3 \times 2 = 12$

② 路径$1 \rightarrow 2 \rightarrow 5 \rightarrow 6$：$2 \times 4 \times 3 = 24$　　最大路径

③ 路径$1 \rightarrow 3 \rightarrow 4 \rightarrow 6$：$1 \times 5 \times 2 = 10$

④ 路径$1 \rightarrow 3 \rightarrow 5 \rightarrow 6$：$1 \times 6 \times 3 = 18$

计算最大路径：动态规划

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的最大路径。

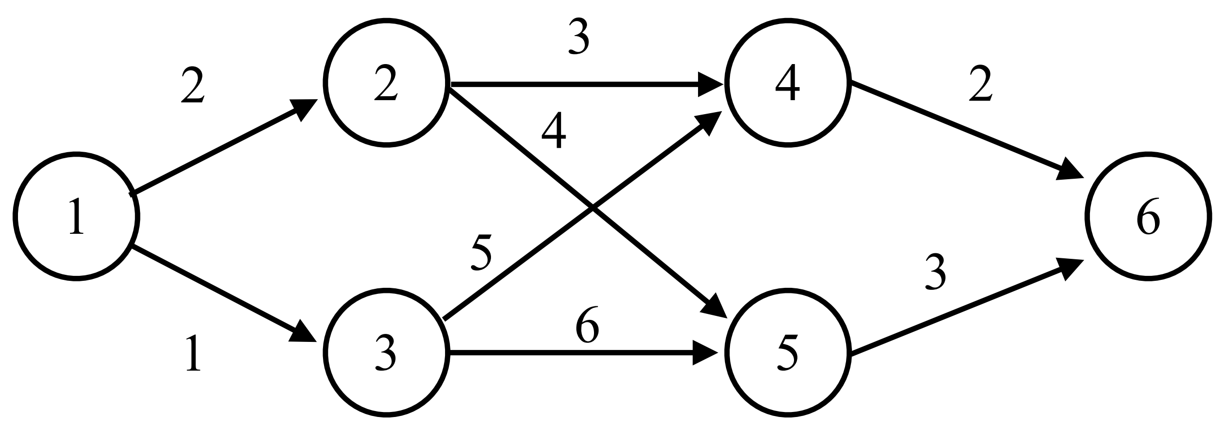

图10：计算最大路径的例子。

令δ_i是从节点1到节点i（$i \in [1, 6]$）的最大路径取值，ψ_i是节点i的最优前继，规定$\delta_1 = 1$且$\psi_1 = 0$，则节点i的最大路径取值及其最优前继可以递归计算：

$$\delta_i = \max_{j \in \mathrm{heads}(i)} w_{ji}\delta_j \qquad \psi_i = \mathrm{argmax}_{j \in \mathrm{heads}(i)} \{w_{ji}\delta_j\}$$

其中，w_{ji}是从节点j到节点i的连线上的权重，$\mathrm{heads}(i)$表示节点i的所有前继。

计算最大路径：动态规划

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的最大路径。

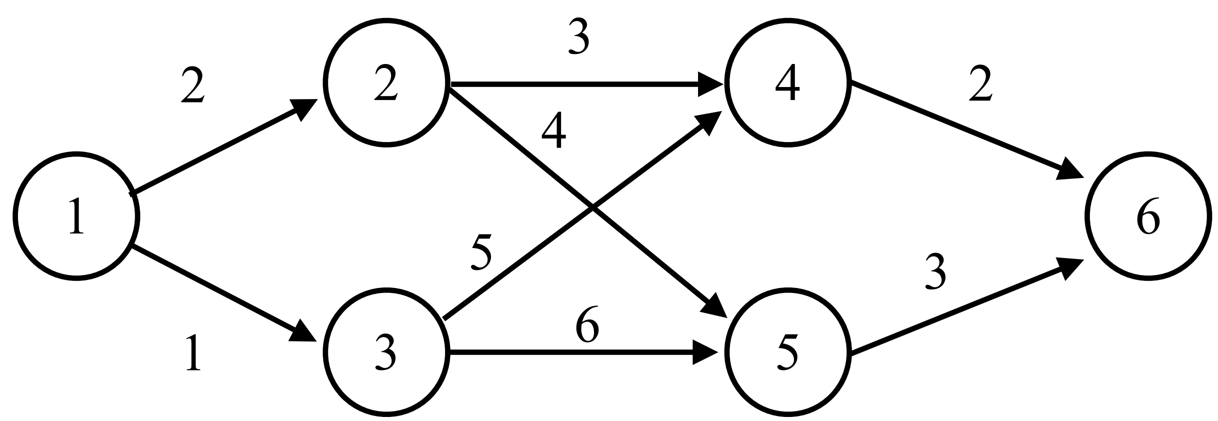

图10：计算最大路径的例子。

$$\delta_i = \max_{j \in \text{heads}(i)} w_{ji}\delta_j$$

$$\psi_i = \text{argmax}_{j \in \text{heads}(i)} \left\{ w_{ji}\delta_j \right\}$$

i	1	2	3	4	5	6
δ_i						
ψ_i						

计算最大路径：动态规划

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的最大路径。

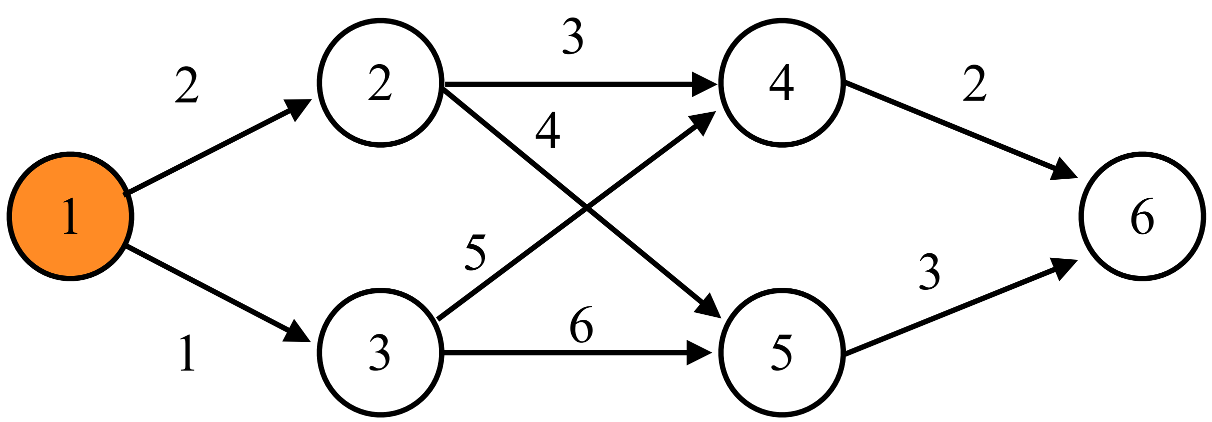

图10：计算最大路径的例子。

$$\delta_i = \max_{j \in \text{heads}(i)} w_{ji}\delta_j$$

$$\psi_i = \text{argmax}_{j \in \text{heads}(i)}\left\{w_{ji}\delta_j\right\}$$

i	1	2	3	4	5	6
δ_i	1					
ψ_i	0					

$\delta_1 = 1$

计算最大路径：动态规划

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的最大路径。

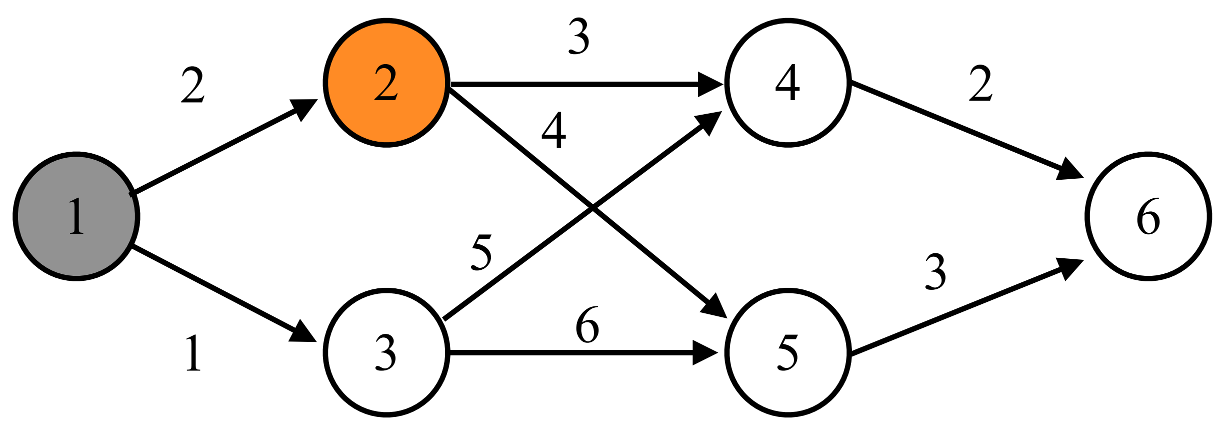

图10：计算最大路径的例子。

$$\delta_i = \max_{j \in \text{heads}(i)} w_{ji}\delta_j$$

$$\psi_i = \text{argmax}_{j \in \text{heads}(i)}\left\{w_{ji}\delta_j\right\}$$

i	1	2	3	4	5	6
δ_i	1	2				
ψ_i	0	1				

$$\delta_2 = \max\{2 \times 1\}$$

计算最大路径：动态规划

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的最大路径。

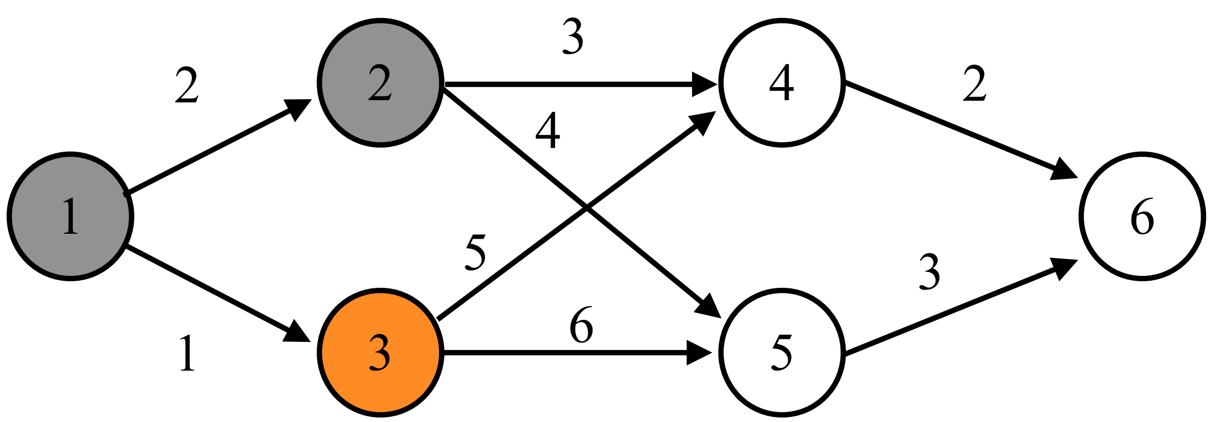

图10：计算最大路径的例子。

$$\delta_i = \max_{j \in \text{heads}(i)} w_{ji}\delta_j$$

$$\psi_i = \text{argmax}_{j \in \text{heads}(i)}\left\{w_{ji}\delta_j\right\}$$

i	1	2	3	4	5	6
δ_i	1	2	1			
ψ_i	0	1	1			

$$\delta_3 = \max\{1 \times 1\}$$

计算最大路径：动态规划

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的最大路径。

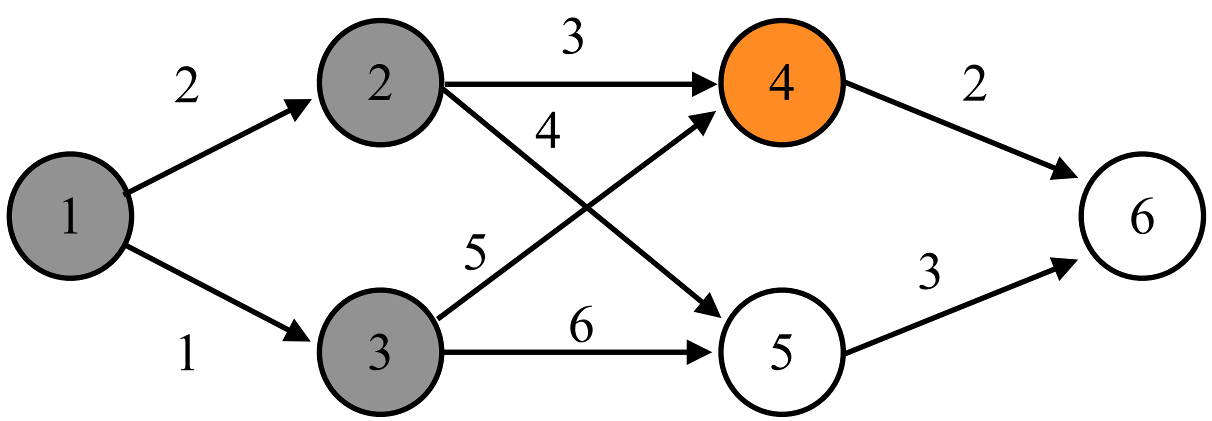

图10：计算最大路径的例子。

$$\delta_i = \max_{j \in \text{heads}(i)} w_{ji}\delta_j$$

$$\psi_i = \text{argmax}_{j \in \text{heads}(i)} \left\{ w_{ji}\delta_j \right\}$$

i	1	2	3	4	5	6
δ_i	1	2	1	6		
ψ_i	0	1	1	2		

$$\delta_4 = \max\{3 \times 2,\ 5 \times 1\}$$

计算最大路径：动态规划

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的最大路径。

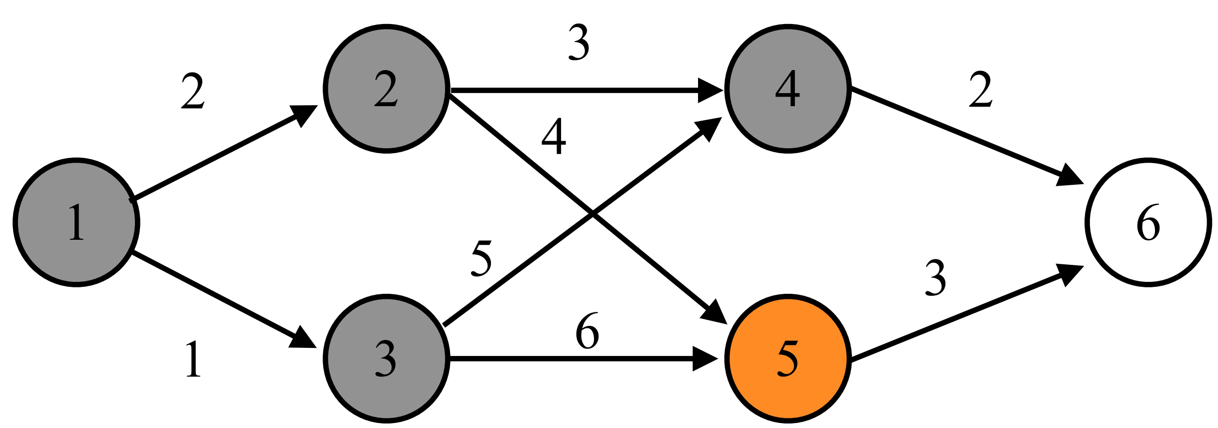

图10：计算最大路径的例子。

$$\delta_i = \max_{j \in \text{heads}(i)} w_{ji}\delta_j$$

$$\psi_i = \text{argmax}_{j \in \text{heads}(i)} \left\{ w_{ji}\delta_j \right\}$$

i	1	2	3	4	5	6
δ_i	1	2	1	6	8	
ψ_i	0	1	1	2	2	

$$\delta_5 = \max\{4 \times 2,\ 6 \times 1\}$$

计算最大路径：动态规划

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的最大路径。

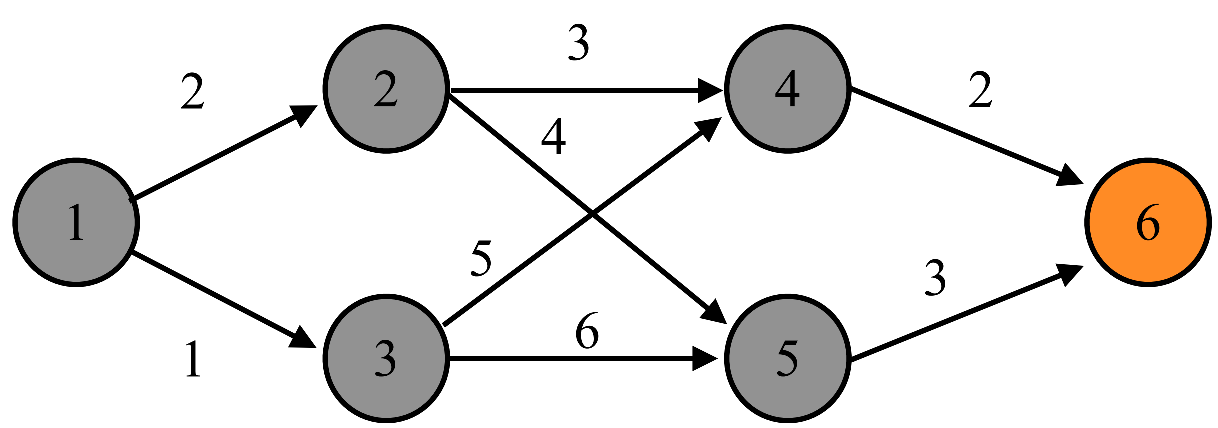

图10：计算最大路径的例子。

$$\delta_i = \max_{j \in \text{heads}(i)} w_{ji}\delta_j$$

$$\psi_i = \text{argmax}_{j \in \text{heads}(i)}\left\{w_{ji}\delta_j\right\}$$

i	1	2	3	4	5	6
δ_i	1	2	1	6	8	24
ψ_i	0	1	1	2	2	5

$$\delta_6 = \max\{2 \times 6,\ 3 \times 8\}$$

回溯

问题描述：设定路径的取值等于边上权重的乘积，请计算从节点1到节点6的最大路径。

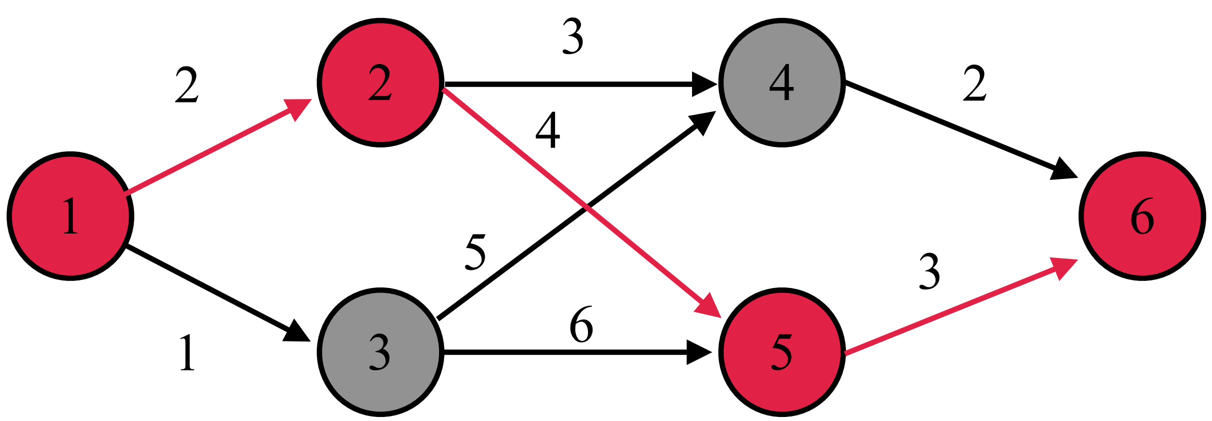

图10：计算最大路径的例子。

$$\delta_i = \max_{j \in \mathrm{heads}(i)} w_{ji}\delta_j$$

$$\psi_i = \mathrm{argmax}_{j \in \mathrm{heads}(i)}\left\{w_{ji}\delta_j\right\}$$

i	1	2	3	4	5	6
δ_i	1	2	1	6	8	24
ψ_i	0	1	1	2	2	5

$$\delta_6 = \max\{2 \times 6,\ 3 \times 8\}$$

Viterbi算法

(1) 初始化:

$$\delta_1(i) = \pi_i b_1(x_1) \qquad \psi_1(i) = 0$$

(2) 递归: $t = 2,\ldots,T$

$$\delta_t(j) = \max_{1 \le i \le N} \{\delta_{t-1}(i) a_{ij}\} b_j(x_t) \qquad \psi_t(j) = \mathrm{argmax}_i \left\{\delta_{t-1}(i) a_{ij}\right\}$$

(3) 结束:

$$\hat{P} = \max_{1 \le i \le N} \{\delta_T(i)\} \qquad \hat{\mathbf{z}}_T = \mathrm{argmax}_i \{\delta_T(i)\}$$

(4) 回溯: $t = T-1,\ldots,1$

$$\hat{\mathbf{z}}_t = \psi_{t+1}(\hat{\mathbf{z}}_{t+1})$$

注意: Viterbi算法与前向概率计算思想一致，只是将加法替换为求最大。

Viterbi算法计算示例

给定模型参数，计算观测状态序列“干燥、潮湿、潮湿”对应的的最优隐状态序列。

	$t = 1$	$t = 2$	$t = 3$
晴天	0.3200 | 0	0.0256 | 1	0.0038 | 2
阴天	0.1800 | 0	0.0640 | 1	0.0102 | 2
雨天	0.0900 | 0	0.0378 | 2	0.0134 | 2

表11： Viterbi算法计算示例。

根据表11，可知观测状态序列“干燥、潮湿、潮湿”对应的最优隐状态序列为“晴天、阴天、雨天”，其概率为0.0134。

Viterbi算法计算示例

$\mathbf{x}$	$\hat{\mathbf{z}}$	$\hat{P}$
干燥、干燥、干燥	晴天、晴天、晴天	0.0328
干燥、干燥、潮湿	晴天、晴天、阴天	0.0205
干燥、潮湿、干燥	晴天、阴天、晴天	0.0154
干燥、潮湿、潮湿	晴天、阴天、雨天	0.0134
潮湿、干燥、干燥	雨天、阴天、晴天	0.0151
潮湿、干燥、潮湿	雨天、阴天、雨天	0.0132
潮湿、潮湿、干燥	雨天、雨天、阴天	0.0176
潮湿、潮湿、潮湿	雨天、雨天、雨天	0.0165

表12：观测状态序列与相应的最优隐状态序列。

内容提要

马尔科夫模型

计算观测概率

Viterbi算法

前向后向算法

估计模型参数

给定训练集$\mathscr{D} = \{\mathbf{x}^{(d)}\}_{d=1}^{D}$，与马尔科夫模型一样，我们使用极大似然估计来获得模型的最优参数：

$$\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}}\Big\{L(\boldsymbol{\theta})\Big\}$$

其中，对数似然函数定义为

$$\begin{aligned} L(\boldsymbol{\theta}) &= \sum_{d=1}^{D} \log P(\mathbf{x}^{(d)};\boldsymbol{\theta}) \\ &= \sum_{d=1}^{D} \log \sum_{\mathbf{z}} P(\mathbf{x}^{(d)},\mathbf{z};\boldsymbol{\theta}) \end{aligned}$$

参数估计的主要挑战在于需要对指数级的隐状态序列进行求和。

EM算法

Expectation-Maximization（简称EM）算法被广泛用于估计隐状态模型的参数。令$\mathbf{X}$表示一组观测数据，$\mathbf{Z}$表示未观测数据，$P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$表示观测数据与未观测数据的联合概率分布，$\boldsymbol{\theta}$表示模型参数。

EM算法在以下两个步骤中迭代运行：

(1) E步：确定对数似然的期望值

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{old}) = \mathbb{E}_{\mathbf{Z}|\mathbf{X};\boldsymbol{\theta}^{old}}\left[\log P(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})\right]$$

(2) M步：计算最大化该期望值的参数

$$\boldsymbol{\theta}^{new} = \mathrm{argmax}_{\boldsymbol{\theta}}\left\{ Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{old}) \right\}$$

利用EM算法训练隐马尔科夫模型

给定训练集$\mathscr{D} = \{\mathbf{x}^{(d)}\}_{d=1}^{D}$，可以使用EM算法训练隐马尔科夫模型。在E步实际使用的目标函数定义如下

$$
\begin{aligned}
J(\boldsymbol{\theta}, \lambda, \boldsymbol{\gamma}, \boldsymbol{\phi}) = & \sum_{d=1}^{D} \mathbb{E}_{\mathbf{z}|\mathbf{x}^{(d)};\boldsymbol{\theta}^{old}}\Big[\log P(\mathbf{x}^{(d)}, \mathbf{z}; \boldsymbol{\theta})\Big] - \\
& \lambda\left(\sum_{z \in \mathscr{S}} p(z) - 1\right) - \\
& \sum_{z \in \mathscr{S}} \gamma_z \left(\sum_{z' \in \mathscr{S}} p(z'|z) - 1\right) - \\
& \sum_{z \in \mathscr{S}} \phi_z \left(\sum_{x \in \mathscr{O}} p(x|z) - 1\right)
\end{aligned}
$$

其中，$\boldsymbol{\gamma} = \{\gamma_z | z \in \mathscr{S}\}$和$\boldsymbol{\phi} = \{\phi_z | z \in \mathscr{S}\}$是分别与隐状态转移概率和观测状态生成概率相关的拉格朗日乘子集合。

估计模型参数

类似于马尔科夫模型，我们可以通过计算偏导，得到以下公式：

$$p(z) = \frac{c(z, \mathcal{D})}{\sum_{z' \in \mathcal{S}} c(z', \mathcal{D})}$$

$$p(z' | z) = \frac{c(z, z', \mathcal{D})}{\sum_{z'' \in \mathcal{S}} c(z, z'', \mathcal{D})}$$

$$p(x | z) = \frac{c(z, x, \mathcal{D})}{\sum_{x' \in \mathcal{O}} c(z, x', \mathcal{D})}$$

其中，$c(\cdot)$是计数函数：$c(z, \mathcal{D})$是在训练集$\mathcal{D}$上第一个隐状态是z的次数的期望值，$c(z, z', \mathcal{D})$是在训练集$\mathcal{D}$上隐状态z'出现在隐状态z的次数的期望值，$c(z, x, \mathcal{D})$是训练集上隐状态z生成观测状态x的次数的期望值。

基于后验概率的期望

上页所述的期望值定义如下：

$$c(z, \mathscr{D}) \equiv \sum_{d=1}^{D} \mathbb{E}_{\mathbf{z}|\mathbf{x}^{(d)};\boldsymbol{\theta}^{old}} \left[\delta(z_1, z) \right]$$

$$c(z, z', \mathscr{D}) \equiv \sum_{d=1}^{D} \mathbb{E}_{\mathbf{z}|\mathbf{x}^{(d)};\boldsymbol{\theta}^{old}} \left[\sum_{t=2}^{T^{(d)}} \delta(z_{t-1}, z)\delta(z_t, z') \right]$$

$$c(z, x, \mathscr{D}) \equiv \sum_{d=1}^{D} \mathbb{E}_{\mathbf{z}|\mathbf{x}^{(d)};\boldsymbol{\theta}^{old}} \left[\sum_{t=1}^{T^{(d)}} \delta(z_t, z)\delta(x_t^{(d)}, x) \right]$$

需要注意的是，这些期望基于隐状态的后验概率$P(\mathbf{z} \mid \mathbf{x}^{(d)}; \boldsymbol{\theta}^{old})$，面临着指数级数量的隐状态求和问题。如何高效计算指数级空间内的求和？

计算期望

下面以隐状态转换次数的期望为例

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{z}|\mathbf{x}^{(d)};\boldsymbol{\theta}^{old}}\Big[\delta(z_{t-1}, z)\delta(z_t, z')\Big] \\
&= \sum_{\mathbf{z}} P(\mathbf{z} \,|\, \mathbf{x}^{(d)}; \boldsymbol{\theta}^{old})\delta(z_{t-1}, z)\delta(z_t, z') \\
&= \sum_{\mathbf{z}} \frac{P(\mathbf{x}^{(d)}, \mathbf{z}; \boldsymbol{\theta}^{old})}{P(\mathbf{x}^{(d)}; \boldsymbol{\theta}^{old})}\delta(z_{t-1}, z)\delta(z_t, z') \\
&= \frac{1}{P(\mathbf{x}^{(d)}; \boldsymbol{\theta}^{old})} \sum_{\mathbf{z}} P(\mathbf{x}^{(d)}, \mathbf{z}; \boldsymbol{\theta}^{old})\delta(z_{t-1}, z)\delta(z_t, z') \\
&= \frac{P(\mathbf{x}^{(d)}, z_{t-1} = z, z_t = z'; \boldsymbol{\theta}^{old})}{P(\mathbf{x}^{(d)}; \boldsymbol{\theta}^{old})}
\end{aligned}
$$

上式的分母可以使用前向概率进行计算，而分子该怎么计算?

利用前向概率和后向概率

巧妙之处在于，我们可以使用前向概率和后向概率计算看似复杂的概率。一般的形式如下：

$$
\begin{aligned}
& P(\mathbf{x}, z_{t-1} = s_i, z_t = s_j; \boldsymbol{\theta}) \\
= & P(x_1, \ldots, x_{t-1}, z_{t-1} = s_i; \boldsymbol{\theta}) \times \\
& P(z_t = s_j \,|\, z_{t-1} = s_i; \boldsymbol{\theta}) \times \\
& P(x_t \,|\, z_t = s_j; \boldsymbol{\theta}) \times \\
& P(x_{t+1}, \ldots, x_T \,|\, z_t = s_j; \boldsymbol{\theta}) \\
= & \alpha_{t-1}(i) a_{ij} b_j(x_t) \beta_t(j)
\end{aligned}
$$

因此，我们可以借助多项式时间复杂度的动态规划算法精准计算指数级空间的求和问题。

图形化解释

从图的角度，$P(\mathbf{x}, z_{t-1} = s_i, z_t = s_j; \boldsymbol{\theta})$的计算过程可以解释如下：

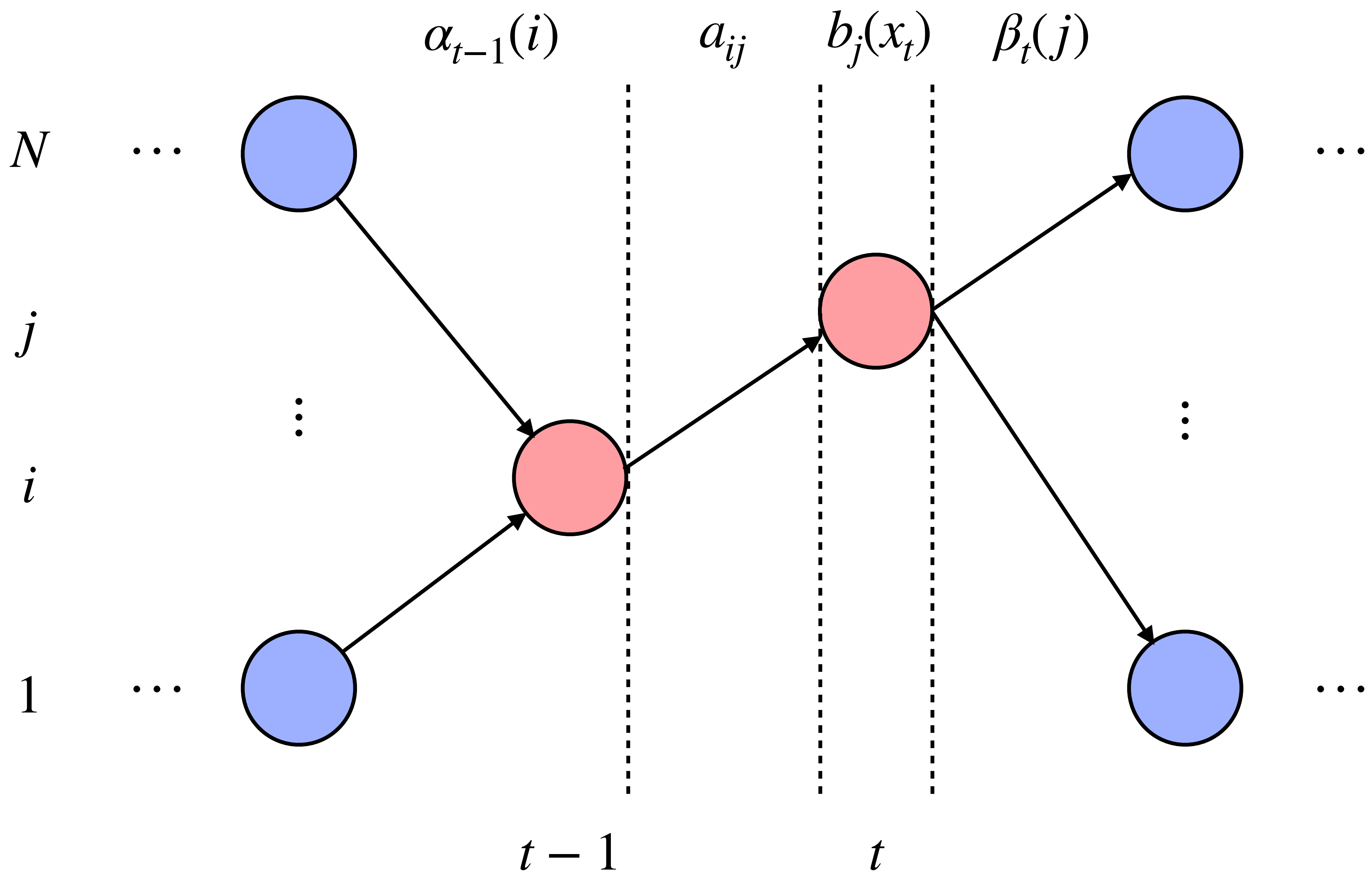

图11：图形化解释。

图形化解释

类似地，$P(\mathbf{x}, z_{t-1} = s_i; \boldsymbol{\theta})$的计算过程可以解释如下：

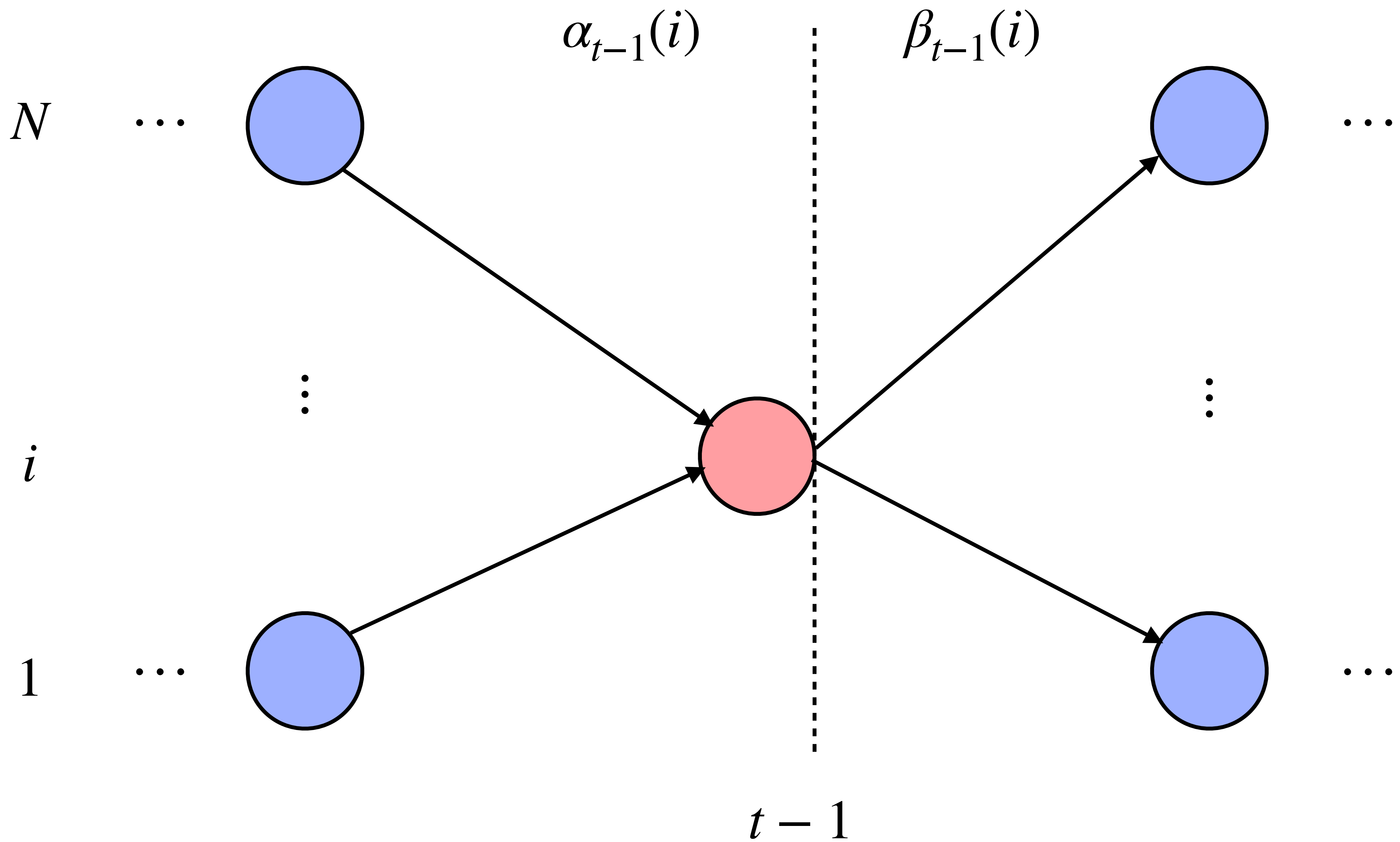

图12：图形化解释。

估计隐状态转换概率

因此，隐状态转换概率可以使用下列公式估计：

$$
\begin{aligned}
p(z'|z) &= \frac{c(z, z', \mathscr{D})}{\sum_{z'' \in \mathcal{S}} c(z, z'', \mathscr{D})} \\
&= \frac{\sum_{d=1}^{D} \sum_{t=2}^{T^{(d)}} P(\mathbf{x}^{(d)}, z_{t-1} = z, z_t = z'; \boldsymbol{\theta}^{old})}{\sum_{d=1}^{D} \sum_{t=2}^{T^{(d)}} P(\mathbf{x}^{(d)}, z_{t-1} = z; \boldsymbol{\theta}^{old})}
\end{aligned}
$$

因此，新的隐状态转换概率可以表示为

$$
\bar{a}_{ij} = \frac{\sum_{d=1}^{D} \sum_{t=2}^{T^{(d)}} \alpha_{t-1}(i) a_{ij} b_j(x_t^{(d)}) \beta_t(j)}{\sum_{d=1}^{D} \sum_{t=2}^{T^{(d)}} \alpha_{t-1}(i) \beta_{t-1}(i)}
$$

我们使用$\bar{a}_{ij}$表示新的概率，a_{ij}表示旧的概率。

估计隐状态初始化概率

类似地，隐状态初始化概率可以使用下列公式估计：

$$p(z) = \frac{c(z, \mathscr{D})}{\sum_{z' \in \mathcal{S}} c(z', \mathscr{D})}$$

$$= \frac{\sum_{d=1}^{D} P(\mathbf{x}^{(d)}, z_1 = z; \boldsymbol{\theta}^{old})}{\sum_{d=1}^{D} P(\mathbf{x}^{(d)}; \boldsymbol{\theta}^{old})}$$

因此，新的隐状态初始化概率可以表示为

$$\bar{\pi}_i = \frac{\sum_{d=1}^{D} \alpha_1(i)\beta_1(i)}{\sum_{d=1}^{D} \sum_{i=1}^{N} \alpha_{T^{(d)}}(i)}$$

估计观测状态生成概率

类似地，观测状态生成概率可以使用下列公式估计：

$$p(x \mid z) = \frac{c(z, x, \mathscr{D})}{\sum_{x' \in \mathscr{O}} c(z, x', \mathscr{D})}$$

$$= \frac{\sum_{d=1}^{D} \sum_{t=1}^{T^{(d)}} \delta(x_t^{(d)}, x) P(\mathbf{x}^{(d)}, z_t = z; \boldsymbol{\theta}^{old})}{\sum_{d=1}^{D} \sum_{t=1}^{T^{(d)}} P(\mathbf{x}^{(d)}, z_t = z; \boldsymbol{\theta}^{old})}$$

因此，新的观测状态生成概率可以表示为

$$\bar{b}_i(k) = \frac{\sum_{d=1}^{D} \sum_{t=1}^{T^{(d)}} \delta(x^{(d)}, o_k) \alpha_t(i) \beta_t(i)}{\sum_{d=1}^{D} \sum_{t=1}^{T^{(d)}} \alpha_t(i) \beta_t(i)}$$

总结

- 隐马尔科夫模型在自然语言处理中具有广泛的应用，是生成式模型的主要代表。
- 要深刻理解隐马尔科夫模型所蕴含的思想，并熟练掌握所涉及的数学工具：
  - 极大似然估计：参数估计的标准准则。
  - 拉格朗日乘子法：带约束条件下的函数极值求解。
  - 动态规划算法：在多项式时间实现指数级空间的精准求和。
  - EM算法：隐状态模型的参数估计标准方法。

谢谢