数学建模
第八章 随机模型

机器学习研究室

时小虎 张禹

目录

1. 五步方法

① Ask the question.

② Select the modeling approach.

③ Formulate the model.

④ Solve the model.

⑤ Answer the question.

① 提出问题

② 选择建模方法

③ 推导模型的数学表达式

④ 求解模型

⑤ 回答问题

第1步，提出问题

① Make a list of all the variables in the problem, including appropriate units.

② Be careful not to confuse variables and constants.

③ State any assumptions you are making about these variables, including equations and inequalities.

④ Check units to make sure that your assumptions make sense.

⑤ State the objective of the problem in precise mathematical terms.

① 列出问题中涉及的变量，包括适当的单位.

② 注意不要混淆变量和常量.

③ 列出你对变量所做的全部假设，包括等式和不等式.

④ 检查单位从而保证你的假设有意义.

⑤ 用准确的数学术语给出问题的目标.

第2步，选择建模方法

① Choose a general solution procedure to be followed in solving this problem.

② Generally speaking, success in this step requires experience, skill, and familiarity with the relevant literature.

③ In this book we will usually specify the modeling approach to be used.

① 选择解决问题的一个一般的求解方法.

② 一般地，这一步的成功需要经验、技巧和熟悉相关文献.

③ 在授课中，我们通常会给定要用的建模方法.

第3步，推导模型的数学表达式

① Restate the question posed in step 1 in the terms of the modeling approach specified in step 2.

② You may need to relabel some of the variables specified in step 1 in order to agree with the notation used in step 2.

③ Note any additional assumptions made in order to fit the problem described.

① 将第一步中得到的问题重新表达成第二步选定的建模方法所需要的形式.

② 你可能需要将第一步中的一些变量名改成与第二步所用的记号一致.

③ 记下任何补充假设，这些假设是为了使第一步中描述的问题与第二步中选定的数学结构相适应而做出的.

第4步，求解模型

① Apply the general solution procedure specified in step 2 to the specific problem formulated in step 3.

② Be careful in your mathematics. Check your work for math errors. Does your answer make sense?

③ Use appropriate technology. Computer algebra systems, graphics, and numerical software will increase the range.

① 将第二步中所选的方法应用于第三步得到的表达式.

② 注意你的数学推导，检查是否有错误，你的答案是否有意义.

③ 采用适当的技术.计算机代数系统、图形工具、数值计算的软件等都能扩大你能解决问题的范围，并能减少计算错误.

第5步，回答问题

① Rephrase the results of step 4 in nontechnical terms.

② Avoid mathematical symbols and jargon.

③ Anyone who can understand the statement of the question as it was presented to you should be able to understand your answer.

① 用非技术性的语言将第四步的结果重新表述.

② 避免数学符号和术语.

③ 能理解最初提出的问题的人就应该能理解你给出的解答.

2. 马尔可夫链

例 8.1 一个宠物商店出售 20 加仑[⊖]的水族箱. 每个周末商店的老板要盘点存货，开出定单. 商店的策略是，如果当前所有的存货都被售出了就在这个周末进三个新的 20 加仑的水族箱. 如果只要在店内还保存有一个存货，就不再进新的水族箱. 这个策略是基于商店平均每周仅出售一个水族箱的事实提出的. 这个策略是不是能够保证防止当商店缺货时顾客需要水族箱而无货销售的损失?

2. 马尔可夫链

例 8.1 一个宠物商店出售 20 加仑[⊖]的水族箱. 每个周末商店的老板要盘点存货，开出定单. 商店的策略是，如果当前所有的存货都被售出了就在这个周末进三个新的 20 加仑的水族箱. 如果只要在店内还保存有一个存货，就不再进新的水族箱. 这个策略是基于商店平均每周仅出售一个水族箱的事实提出的. 这个策略是不是能够保证防止当商店缺货时顾客需要水族箱而无货销售的损失?

变量:

S_n = 第 n 周之初水族箱的供应

D_n = 第 n 周内水族箱的需求

假设:

如果 $D_{n-1} < S_{n-1}$，则 $S_n = S_{n-1} - D_{n-1}$

如果 $D_{n-1} \geqslant S_{n-1}$，则 $S_n = 3$

$\Pr\{D_n = k\} = e^{-1}/k!\quad k = 0, 1, 2, 3, \cdots$

目标: 计算 $\Pr\{D_n > S_n\}$

2. 马尔可夫链

例 8.1 一个宠物商店出售 20 加仑[⊖]的水族箱. 每个周末商店的老板要盘点存货，开出定单. 商店的策略是，如果当前所有的存货都被售出了就在这个周末进三个新的 20 加仑的水族箱. 如果只要在店内还保存有一个存货，就不再进新的水族箱. 这个策略是基于商店平均每周仅出售一个水族箱的事实提出的. 这个策略是不是能够保证防止当商店缺货时顾客需要水族箱而无货销售的

n 周之初水族箱的供应

n 周内水族箱的需求

$_{n-1}<S_{n-1}$，则 $S_n=S_{n-1}-D_{n-1}$

$_{n-1}\geqslant S_{n-1}$，则 $S_n=3$

$=k\}=e^{-1}/k!$ $k=0$，1，2，3，…

目标： 计算 $\Pr\{D_n>S_n\}$

6. 泊松分布

(1) 应用场景

某一区间内发生随机事件次数的概率分布，比如：每小时出生3个婴儿，某网站平均每分钟有2次访问。

(2) 描述

一个离散型随机变量X 满足：

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda},(k=0,1\ldots)$$

期望：$E(X)=\lambda$

方差：$D(X)=\lambda$

2.马尔可夫链

第二步选择模型：马尔可夫链

- 如果一个过程的“将来”仅依赖“现在”而不依赖“过去”，则此过程具有**马尔可夫性**,或称此过程为**马尔可夫过程**
- $X(t+1) = f(X(t))$
- **时间**和**状态**都离散的马尔科夫过程称为马尔科夫链
- 记作$\{X_n = X(n), n = 0,1,2,\ldots\}$：在时间集$T_1 = \{0,1,2,\ldots\}$上对离散状态的过程相继观察的结果
- 链的状态空间记做$I = \{a_1, a_2,\ldots\}, a_i \in R$.
- 条件概率$P_{ij}(m, m+n)=P\{X_{m+n} = a_j | X_m = a_i\}$为马氏链在时刻m处于状态$a_i$条件下，在时刻m+n转移到状态$a_j$的**转移概率**。

2.马尔可夫链

第二步选择模型：马尔可夫链

Weather: A Markov Model

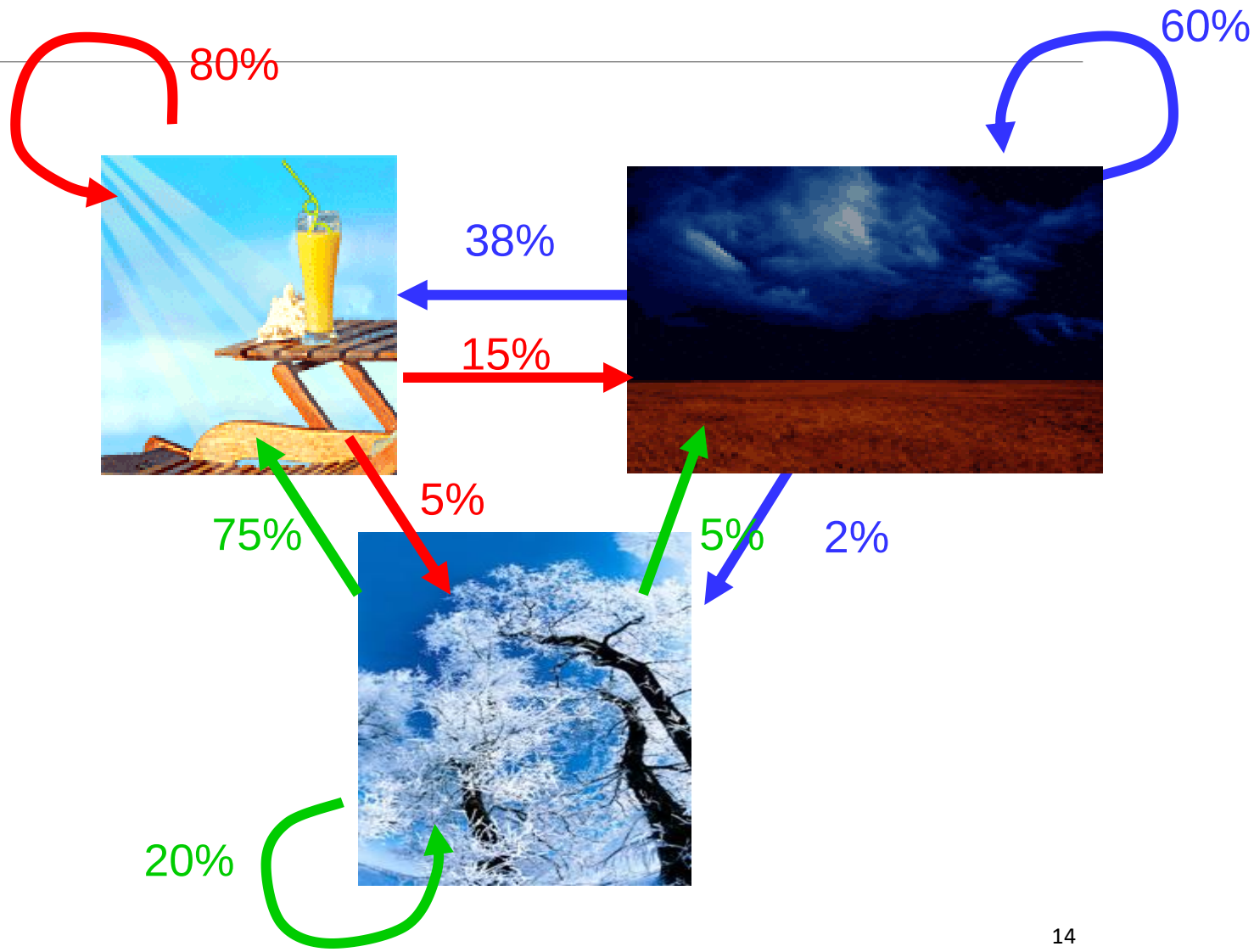
Sunny

Rainy

Snowy

2.马尔可夫链

第二步选择模型：马尔可夫链



2.马尔可夫链

第二步选择模型：马尔可夫链

States:

$$\{S_1, S_2, ..., S_N\}$$

State transition probabilities:

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

represents the probability of moving from state i to state j

$A = \{a_{ij}\}$: transition probability matrix

$$0 \le a_{ij} \le 1; \sum_{j=1}^{N} a_{ij} = 1, 1 \le i \le N$$

Initial state distribution:

$$\pi_i = P[q_1 = S_i]$$

$$0 \le \pi_i \le 1, 1 \le i \le N; \sum_{j=1}^{N} \pi_i = 1$$

2.马尔可夫链

第二步选择模型：马尔可夫链

Weather Markov Model

States: $\{S_{sunny}, S_{rainy}, S_{snowy}\}$

State transition probabilities:

$$A = \begin{pmatrix} .8 & .15 & .05 \\ .38 & .6 & .02 \\ .75 & .05 & .2 \end{pmatrix}$$

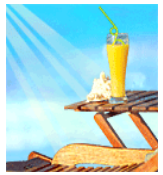
Initial state distribution:

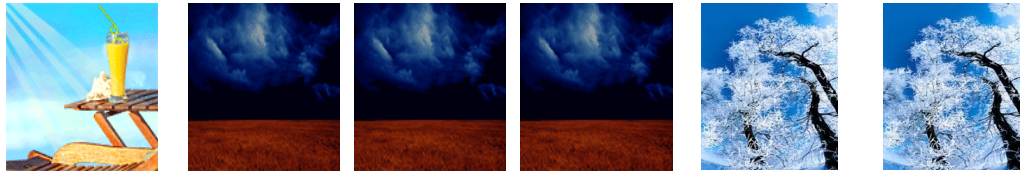
$$\pi = (.7 \quad .25 \quad .05)$$

2.马尔可夫链

第二步选择模型：马尔可夫链

Basic Calculations-1

Given the weather on the first day  , what is the probability that the weather for consecutive six days



t=1 t=2 t=3 t=4 t=5 t=6

Sunny Rainy Rainy Rainy Snowy Snowy

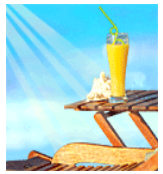
2.马尔可夫链

第二步选择模型：马尔可夫链

$$P(S_{sunny}, S_{rainy}, S_{rainy}, S_{rainy}, S_{snowy}, S_{snowy} \mid Model)$$
$$= P(S_{sunny}) \cdot P(S_{rainy} \mid S_{sunny}) \cdot P(S_{rainy} \mid S_{rainy}) \cdot P(S_{rainy} \mid S_{rainy})$$
$$\cdot P(S_{snowy} \mid S_{rainy}) \cdot P(S_{snowy} \mid S_{snowy})$$
$$= 0.7 \cdot 0.15 \cdot 0.6 \cdot 0.6 \cdot 0.02 \cdot 0.2 = 0.0001512$$

2.马尔可夫链

第二步选择模型：马尔可夫链

- **Basic Calculations-2**
  Given that the system is in a known weather, e.g.  ,what is the probability that it stays in the same weather for consecutive d days:

2.马尔可夫链

第二步选择模型：马尔可夫链

$$Q = \{\underbrace{s_i, s_i, s_i, ..., s_i}_{d}, s_j, i \neq j\}$$

$$p(Q \mid Model, q_1 = s_i) = p(q_1 = s_i, Q \mid Model) / p(q_1 = s_i)$$

$$= \sum_{j=1, j\neq i}^{N} p(q_1 = s_i, \{\underbrace{s_i, s_i, s_i, ..., s_i}_{d}, s_j\} \mid Model) / p(q_1 = s_i)$$

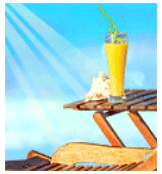
$$= p(q_1 = s_i)(p(s_i \mid s_i))^{d-1} \sum_{j=1, j\neq i}^{N} p(s_j \mid s_i) / p(q_1 = s_i)$$

$$= a_{ii}^{\ d-1}(1 - a_{ii}) = p_i(d)$$

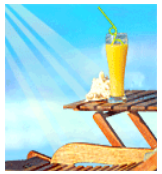
2.马尔可夫链

第二步选择模型：马尔可夫链

Basic Calculations-3

Conditioned on starting the weather , e.g.  **, compute the expected number of duration in weather**

$$\overline{d_i} = \sum_{d=1}^{\infty} d p_i(d) = \sum_{d=1}^{\infty} d a_{ii}^{\,d-1}(1-a_{ii}) = \frac{1}{1-a_{ii}}$$

Expected number of consecutive  **days is 5.**

2.马尔可夫链

第二步选择模型：马尔可夫链

若令 $\pi_n(i) = Pr\{X_n=i\}$， $\pi_n = (\pi_n(1), \pi_n(2), \ldots, \pi_n(m))'$, 则有 $\pi_{n+1} = P' \pi_n$

$\sum \pi_0(i) = 1, \sum_j P_{ij} = 1, \rightarrow \sum \pi_n(i) = 1$

1是P的最大特征值，$\pi_n = P^n \pi_0$, 乘幂法，收敛于主特征向量。称之为稳定状态分布。

2.马尔可夫链

第三步 推导模型：

我们现在回到例 8.1 的存货问题. 我们将使用马尔可夫链来给这个问题建模. 步骤三是组建模型. 我们从研究状态空间开始. 这里状态的概念与确定性的动态系统是一样的. 状态包含有为预报这个过程(概率上)的将来所必须的全部信息. 我们将取 $X_n = S_n$ 作为状态变量，它表明在我们这个销售周一开始库存水族箱的数量. 需求量 D_n 与模型的动态有关，将被用来构成状态转移矩阵 P. 状态空间是

$$X_n \in \{1,2,3\}.$$

我们不知道初始状态，但是似乎有理由假设 $X_0 = 3$. 为了确定 P，我们将从画状态转移图开始. 参见图 8-3. 需求量的分布为

2.马尔可夫链

第三步 推导模型：

$$\Pr\{D_n=k\}=e^{-1}/k!\quad k=0,\ 1,\ 2,\ 3,\ \cdots$$

$$\Pr\{D_n=0\}=0.368$$
$$\Pr\{D_n=1\}=0.368$$
$$\Pr\{D_n=2\}=0.184$$
$$\Pr\{D_n=3\}=0.061$$
$$\Pr\{D_n>3\}=0.019,$$

于是，如果 $X_n=3$，则

$$\Pr\{X_{n+1}=1\}=\Pr\{D_n=2\}=0.184$$
$$\Pr\{X_{n+1}=2\}=\Pr\{D_n=1\}=0.368$$
$$\Pr\{X_{n+1}=3\}=1-(0.184+0.368)=0.448.$$

$$P=\begin{bmatrix}0.368 & 0 & 0.632\\ 0.368 & 0.368 & 0.264\\ 0.184 & 0.368 & 0.448\end{bmatrix}$$

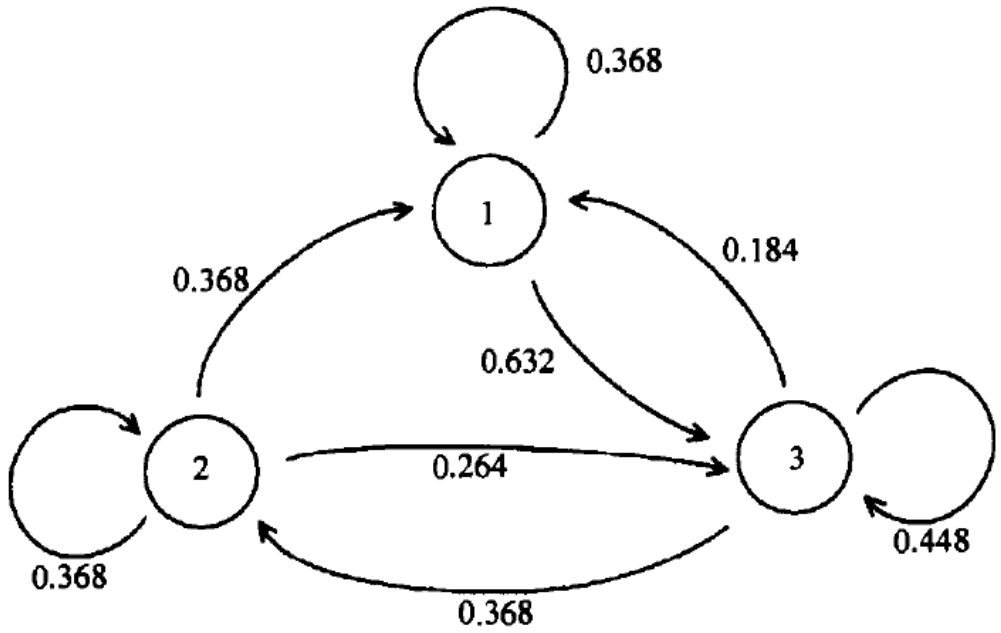


图 8-3　存货问题的状态转移图

2.马尔可夫链

第四步 求解模型：

现在我们进行步骤四. 分析的目标是要计算需求超过供给的概率

$$\Pr\{D_n > S_n\}$$

一般来说，这个概率依赖于 n. 更具体地说，它依赖于 X_n，如果 $X_n=3$，则

$$\Pr\{D_n > S_n\} = \Pr\{D_n > 3\} = 0.019,$$

等等. 为了得到关于需求多么经常超过供给的更好的想法，我们需要更多关于 X_n 的信息.

因为$\{X_n\}$是一个遍历的马尔可夫链，我们知道一定存在惟一的渐近稳定的概率向量 π，它可以通过求解稳定状态方程计算出来. 将方程(9)代回(7)式，我们得到

$$\begin{aligned}\pi_1 &= 0.368\pi_1 + 0.368\pi_2 + 0.184\pi_3\\ \pi_2 &= 0.368\pi_2 + 0.368\pi_3\\ \pi_3 &= 0.632\pi_1 + 0.264\pi_2 + 0.448\pi_3,\end{aligned} \tag{10}$$

我们需要在条件

$$\pi_1 + \pi_2 + \pi_3 = 1,$$

下求解得到 X_n 的稳定状态分布.

2.马尔可夫链

第四步 求解模型：

$$\pi = (\pi_1, \pi_2, \pi_3) = (0.285, 0.263, 0.452).$$

对于充分大的 n，近似有

$$\Pr\{X_n = 1\} = 0.285$$
$$\Pr\{X_n = 2\} = 0.263$$
$$\Pr\{X_n = 3\} = 0.452.$$

将它与我们关于 D_n 的信息放在一起，我们得到，

$$\Pr\{D_n > S_n\} = \sum_{i=1}^{3} \Pr\{D_n > S_n \mid X_n = i\} \Pr\{X_n = i\}$$
$$= (0.264)(0.285) + (0.080)(0.263) + (0.019)(0.452)$$
$$= 0.105$$

在长时间的运行中，需求将有 10%的时间超过供给.

2.马尔可夫链

第五步 回答问题：

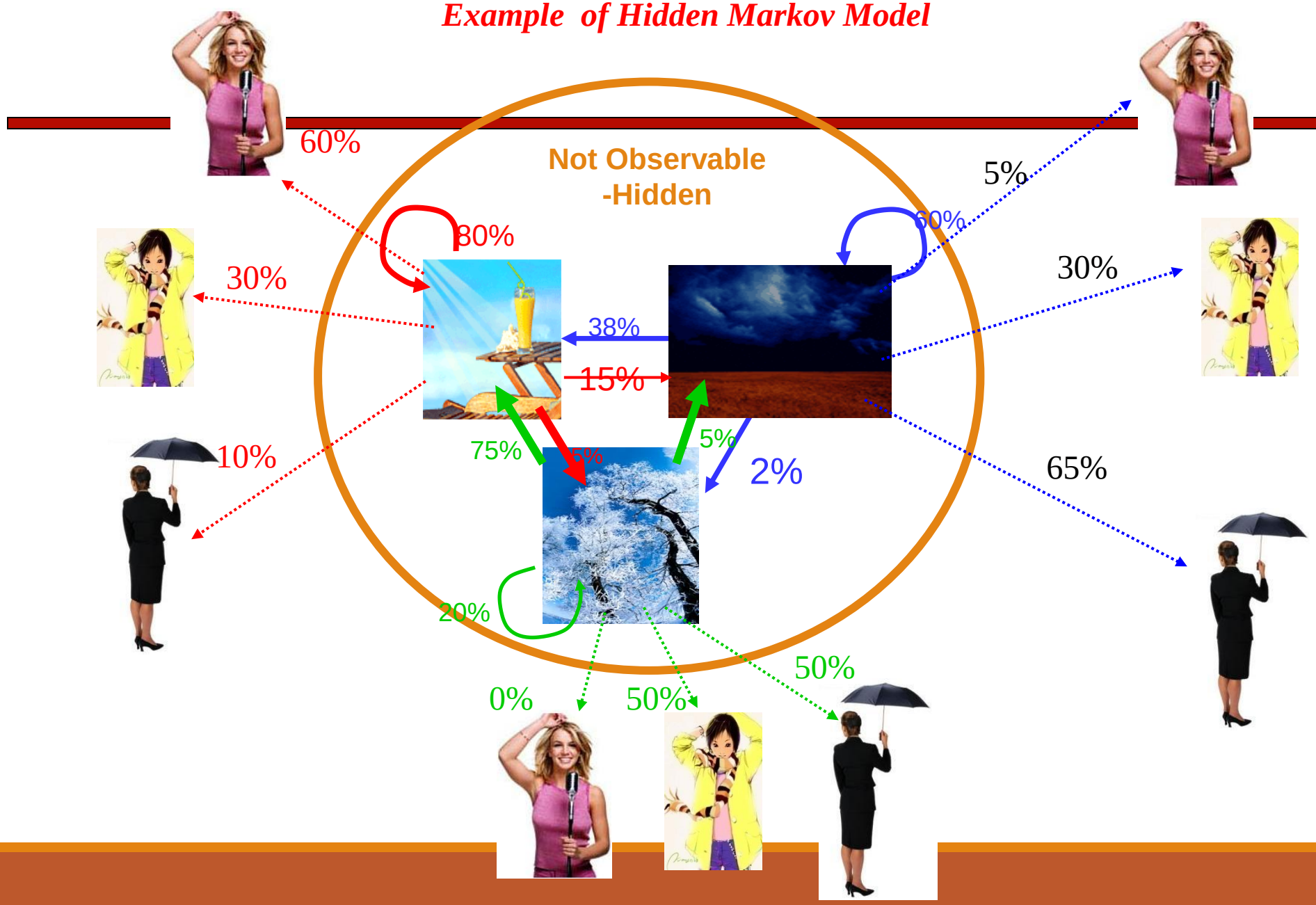
最后，我们进行步骤五. 当前的存货策略导致有大约 10%的时间的无货销售的损失，或者说每年至少有 5 次缺货. 这主要是由于当仅有一个水族箱的库存时我们没有更多地进货. 虽然我们每周平均仅仅出售一个，每一周潜在销售的实际个数(需求)从一周到下一周是波动的. 因此当我们开始仅有一个水族箱库存的这一周的销售时，我们冒着很大的(大约是四分之一)由于不充足的库存而失去潜在的销售机会的损失. 如果不存在其他的因素，例如对预定三个或更多的水族箱时打折扣等，似乎有理由尝试新的策略使得不会在开始一周的销售时仅有一个水族箱.

3.隐马尔可夫模型

The World is not fully observable!!!

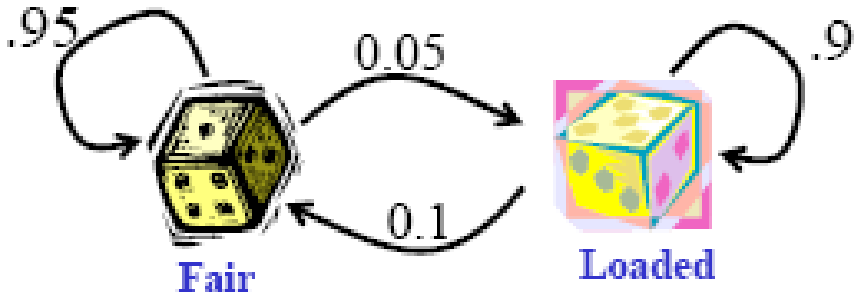
Hidden Markov Model

Example of Hidden Markov Model



- Occasionally dishonest casino

1:	1/6
2:	1/6
3:	1/6
4:	1/6
5:	1/6
6:	1/6



1:	1/10
2:	1/10
3:	1/10
4:	1/10
5:	1/10
6:	5/10

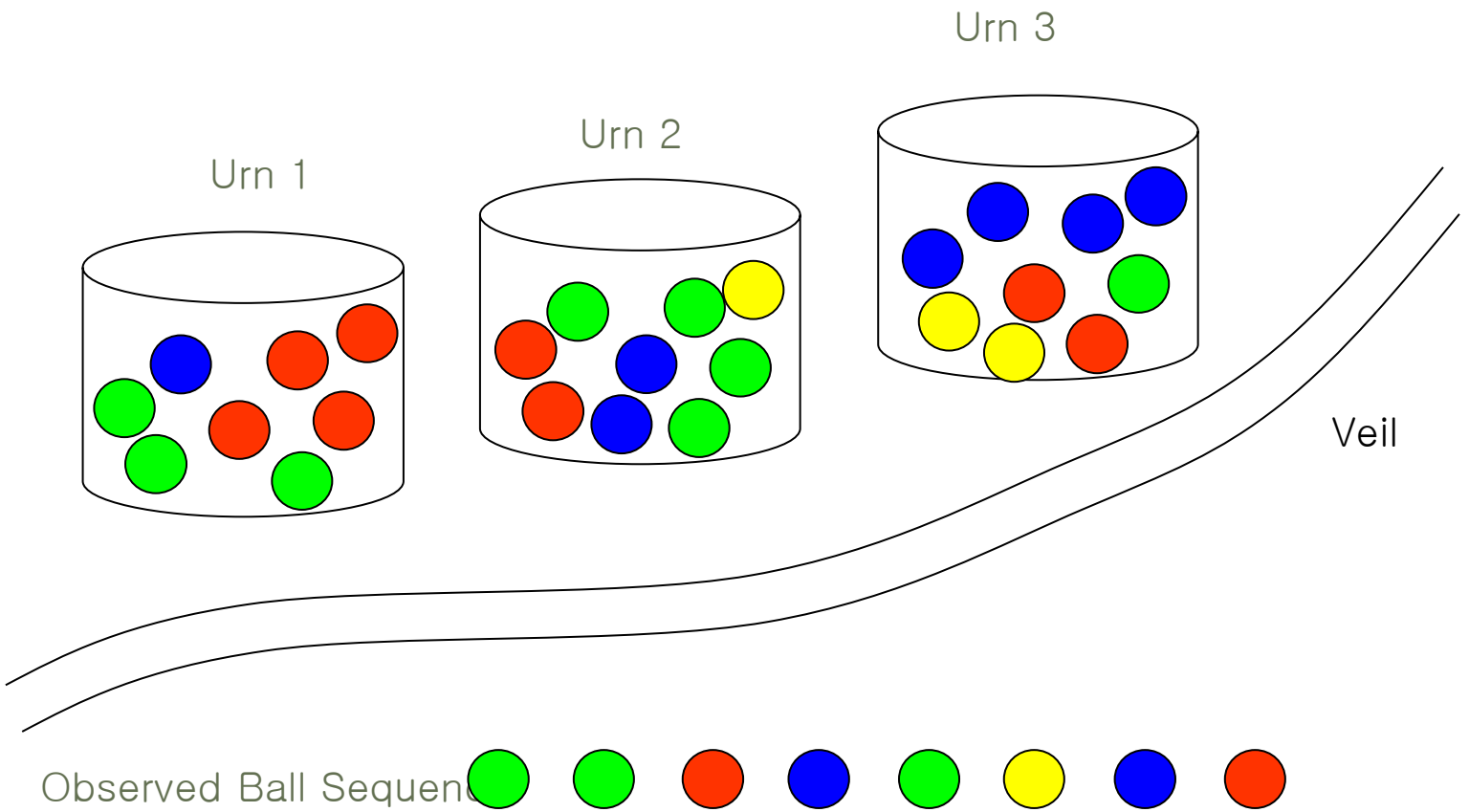
Simulted a sequence of 100 "die usages"

1 1 1 1 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 1 6 2 1 6 6 5 6 6 6 3 5 2 3 2 1 2 6 4 6 2 2 5 3 3 3 1 4 3 1 5 1 3 6 1 6 3 5 1 6 3 1 2 3 1 4 6 3 6

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2
5 1 3 3 5 6 1 3 5 5 4 6 3 2 4 1 6 2 5 4 2 4 4 2 1 2 3 2 6 3 6 6 6 4 5 6 2 2 4 6 6 1 4 6 3 4 2 6 4 6

Question: what's the long-run fraction of each type?

However, we only observe the result of the rolls of a die: can we tell?


Urn 3
Urn 2
Urn 1
Veil
Observed Ball Sequen

3.隐马尔可夫模型

在上述实验中，有几个要点需要注意：

缸间的转移不能被直接观察

从缸中所选取的球的颜色和缸并不是

一一对应的

每次选取哪个缸由一组转移概率决定

3.隐马尔可夫模型

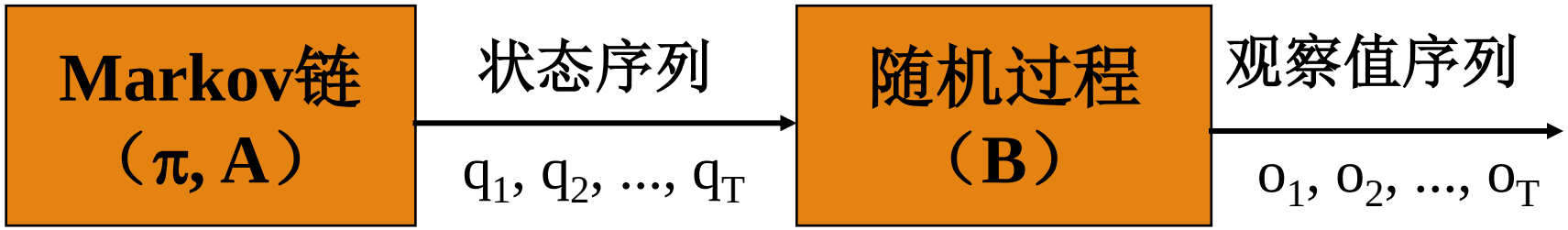
HMM的状态是不确定或不可见的，只有通过观测序列的随机过程才能表现出来

观察到的事件与状态并不是一一对应，而是通过一组概率分布相联系

HMM是一个双重随机过程，两个组成部分：

- **马尔可夫链**：描述状态的转移，用**转移概率**描述。
- **一般随机过程**：描述状态与观察序列间的关系， 用**观察值概率**描述。

3.隐马尔可夫模型



HMM的组成示意图

3.隐马尔可夫模型

用模型五元组 λ =（ N, M, π， A， B）用来描述HMM，或简写为λ =(π， A， B)

参数	含义	实例
N	状态数目	缸的数目
M	每个状态可能的观察值数目	彩球颜色数目
A	与时间无关的状态转移概率矩阵	在选定某个缸的情况下，选择另一个缸的概率
B	给定状态下，观察值概率分布	每个缸中的颜色分布
π	初始状态空间的概率分布	初始时选择某口缸的概率

3.隐马尔可夫模型

HMM可解决的问题

问题1：给定观察序列$O=O_1,O_2,\ldots O_T$,以及模型$\lambda=(\pi, A, B)$，如何计算$P(O|\lambda)$？

问题2：给定观察序列$O=O_1,O_2,\ldots O_T$以及模型λ,如何选择一个对应的状态序列 $S=q_1,q_2,\ldots q_T$，使得S能够最为合理的解释观察序列O？

问题3：如何调整模型参数$\lambda=(\pi, A, B)$，使得$P(O|\lambda)$最大？

3.隐马尔可夫模型

问题1：给定观察序列$O=O_1,O_2,\ldots O_T$,以及模型$\lambda=(\pi, A, B)$，如何计算$P(O|\lambda)$？

给定一个固定的状态序列$S=(q_1, q_2, q_3\ldots)$

$$P(O/S,\lambda)=\prod_{t=1}^{T}P(O_t/q_t,\lambda)=b_{q_1}(O_1)b_{q_2}(O_2)\cdots b_{q_T}(O_T)$$

$b_{q_t}(O_t)$ 表示在q_t状态下观测到O_t的概率

$$P(O/\lambda)=\sum_{\text{所有S}}P(O/S,\lambda)P(S/\lambda)$$

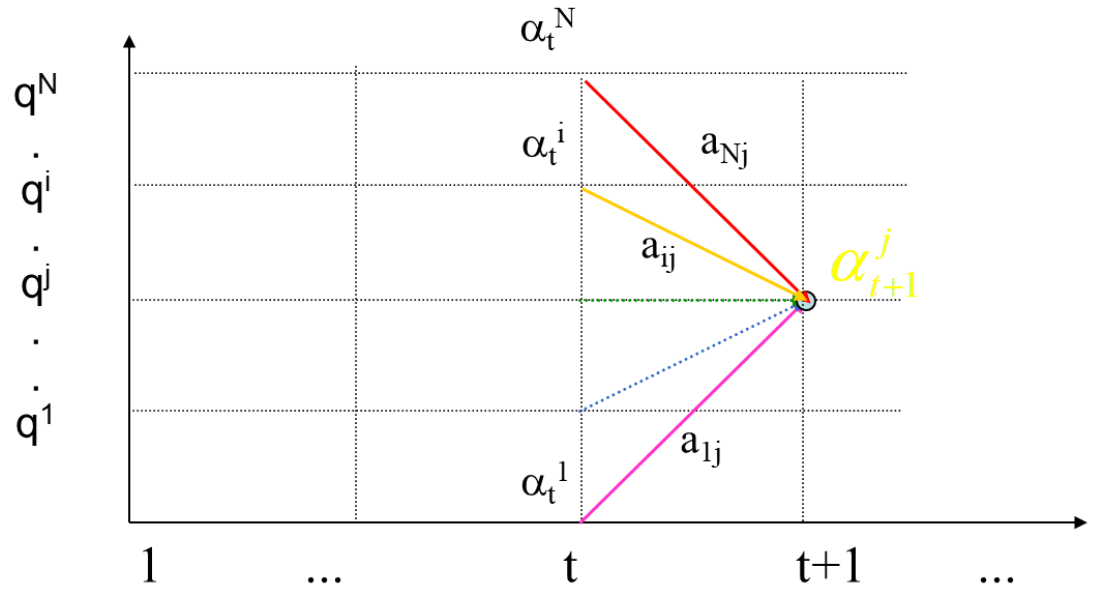
N=5, M=100, => 计算量10^72

3.隐马尔可夫模型

动态规划

定义前向变量 $\alpha_t(i) = P(O_1, O_2, \cdots O_t, q_t = \theta_i / \lambda)$ $1 \le t \le T$

- 初始化：
- 递归：
- 终结：



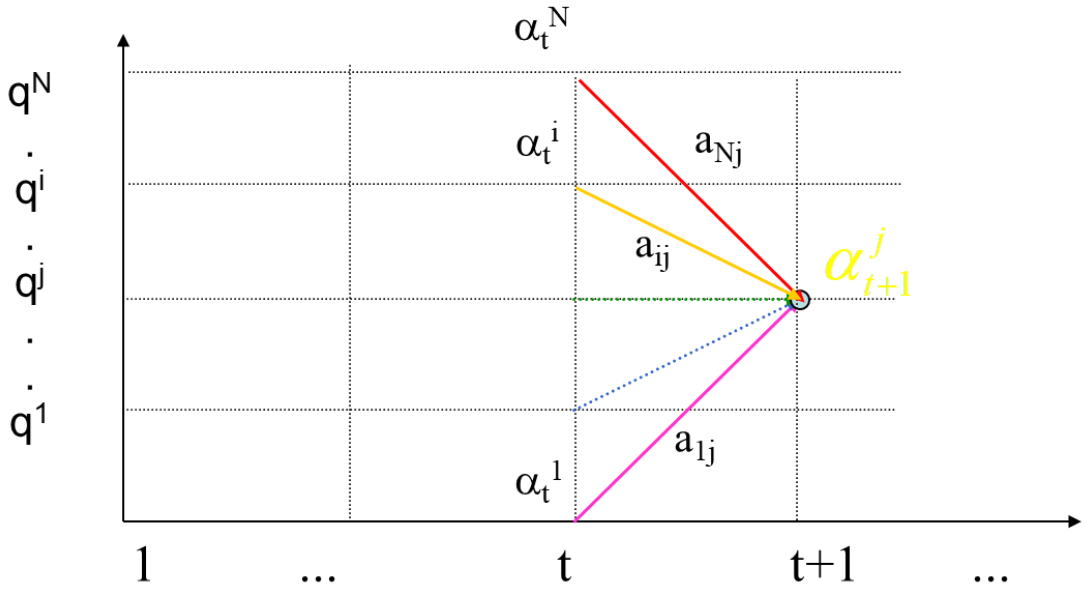
3.隐马尔可夫模型

动态规划

定义前向变量 $\alpha_t(i) = P(O_1, O_2, \cdots O_t, q_t = \theta_i / \lambda) \qquad 1 \le t \le T$

- 初始化： $\alpha_1(i) = \pi_i b_i(O_1) \qquad 1 \le i \le N$
- 递归： $\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_i(i) a_{ij}] b_j(O_{t+1})$

  $1 \le t \le T-1, 1 \le j \le N$
- 终结： $P(O / \lambda) = \sum_{i=1}^{N} \alpha_T(i)$



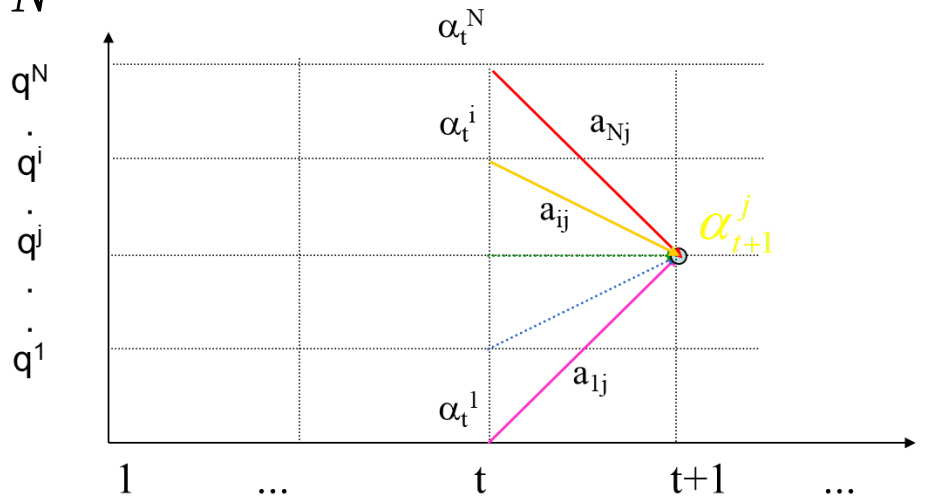
N=5, M=100, => 计算量3000

3.隐马尔可夫模型

问题2：给定观察序列O=$O_1,O_2,\ldots O_T$以及模型λ,如何选择一个对应的状态序列 S = $q_1,q_2,\ldots q_T$，使得S能够最为合理的解释观察序列O？

定义前向变量 $\delta_t(i) = \max_{q_1,q_2,\cdots,q_{t-1}} P(q_1 q_2 \cdots q_{t-1}, q_t = i, O_1, O_2, \cdots O_t / \lambda)$

- 初始化：
$$\delta_1(i) = \pi_i b_i(O_1), \quad 1 \le i \le N$$
$$\varphi_1(i) = 0, \quad 1 \le i \le N$$
- 递归：
$$\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_j(O_i), \quad 2 \le t \le T, 1 \le j \le N$$
$$\varphi_t(j) = \arg\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}], \quad 2 \le t \le T, 1 \le j \le N$$
- 终结：
$$P^* = \max_{1\le i\le N}[\delta_T(i)])$$
$$q_T^* = \arg\max_{1\le i\le N}[\delta_T(i)]$$
- 求S:
$$q_t^* = \varphi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, \ldots, 1$$



3.隐马尔可夫模型

问题3：如何调整模型参数 λ =(π，A，B)，使得P(O|λ)最大？

算法步骤：

1. 初始模型（待训练模型） λ_0,
2. 基于λ_0 以及观察值序列O，训练新模型 λ；
3. 如果 log P(X|λ) - log(P(X|λ_0) < Delta，说明训练已经达到预期效果， 算法结束。
4. 否则，令λ_0 ＝ λ，继续第2步工作

3.隐马尔可夫模型

问题3：如何调整模型参数 λ =(π，A，B)，使得P(O|λ)最大？

给定模型λ和观察序列条件下，从i到j的转移概率定义为$\xi_t(i,j)$

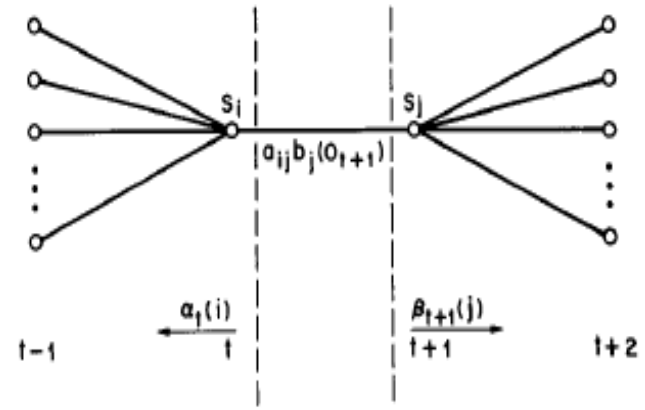
$$\xi_t(i,j)=P(s_t=i,s_{t+1}=j\mid X,\lambda)$$

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

$\gamma_t(i)=\sum_{j=1}^{N}\xi_t(i,j)$　t时刻处于状态S_i的概率

$\sum_{t=1}^{T-1}\gamma_t(i)=$整个过程中从状态$S_i$转出的次数(number of time)的预期

$\sum_{t=1}^{T-1}\xi_t(i,j)=$从$S_i$跳转到$S_j$次数的预期



3.隐马尔可夫模型

问题3：如何调整模型参数 λ =(π，A，B)，使得P(O|λ)最大？

参数估计

Reestimate :

$$\hat{a}_{ij} = \frac{\text{expected count of transitions from i to j}}{\text{expected count of stays at i}} = \frac{\sum_t \xi_t(i,j)}{\sum_t \sum_j \xi_t(i,j)}$$

$$\hat{b}_j(k) = \frac{\text{expected number of times in state } j \text{ and observing symbol } k}{\text{expected number of times in state } j} = \frac{\sum_{t,O_t=k} \gamma_t(j)}{\sum_t \gamma_t(j)}$$

π_i ＝当t＝1时处于S_i的概率 ＝ $\gamma_1(i)$

3.隐马尔可夫模型

CpG岛识别

- 指DNA上一个区域，此区域含有大量相联的胞嘧啶（C）、鸟嘌呤（G），以及使两者相连的磷酸酯键（p）。哺乳类基因中的启动子上，含有约40%的CpG岛（人类约70%）。一般CpG岛的长度约300到3000个碱基对（bp）。
- 在许多基因的启动子（promotor）或“起始”区域周围，甲基化经常被抑制。这些区域包含浓度相对较高的CpG对，与此段区域对应的染色体区段一起被称作CpG岛
- CpG岛常位于管家基因和其他在细胞中被频繁表达基因的启动子区域

3.隐马尔可夫模型

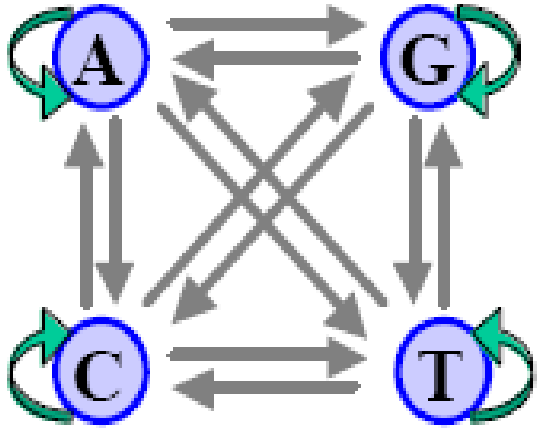
CpG岛识别

- Problem:

  Given a short stretch of genomic data, does it come from a CpG island or not?

- States: {A,T,G,C}

- Transition probabilities matrix P



3.隐马尔可夫模型

CpG岛识别

- Training the Markov Models
  - Transition probabilities matrix P- based on known non-CpG Island sequences
  - Transition probabilities matrix P+ based on known CpG Island sequences
- "+"model:
  - Use transition matrix P+
- "-"model:
  - Use transition matrix P-

P+

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

P-

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

3.隐马尔可夫模型

CpG岛识别

- For a new sequence x:

$x=x_1 \ldots x_L$

- Probability of a sequence

$$P(x) = P(x_L \mid x_{L-1}) \ldots P(x_2 \mid x_1) P(x_1)$$

$$P(x_1) = P(x_1 \mid x_0)$$

$$P(x) = \prod_{i=0}^{L-1} P(x_{i+1} \mid x_i) = \prod_{i=0}^{L-1} a_{x_i, x_{i+1}}$$

3.隐马尔可夫模型

- New sequence: x=TGCAGCG, x from CpG island regions?

$$
\begin{aligned}
P(x|+model) &= P_+(G|C) * P_+(C|G) * P_+(G|A) * P_+(A|C) * P_+(C|G) * P_+(G|T) \\
&= 0.274 * 0.339 * 0.426 * 0.171 * 0.339 * 0.384 \\
&= 0.000880819444 \\
P(x|-model) &= P_-(G|C) * P_-(C|G) * P_-(G|A) * P_-(A|C) * P_-(C|G) * P_-(G|T) \\
&= 0.078 * 0.246 * 0.285 * 0.322 * 0.246 * 0.292 \\
&= 0.00012648773
\end{aligned}
$$

P+

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

P-

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

- Ratio=P(x | +model)/P(x | -model)=6.96>1
- X more likely comes from CpG island regions
- In fact， take ‘log’: * to +

3.隐马尔可夫模型

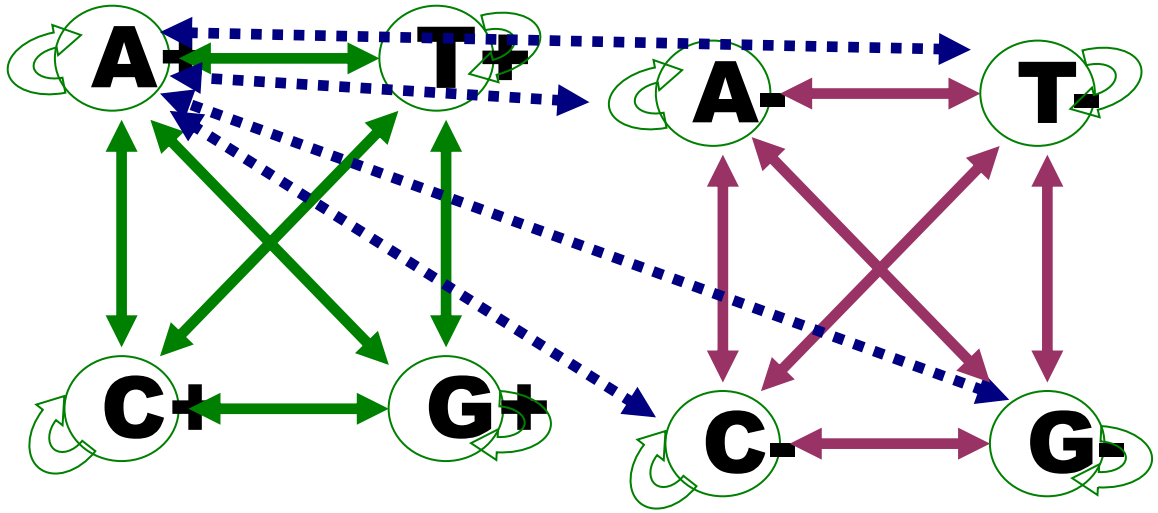
CpG岛识别

- Further Problem:
  For a (stretch of a) genomic sequence, where are the CpG islands?
  TAAAAAATAAATATGTTTAATTTGTGAACTGATTACCATCAGAAT
- States: {N, C}
- Observations: {A,T,G,C}
- Transition probabilities matrix P: 2*2

3.隐马尔可夫模型

CpG岛识别

- Further Problem:
  For a (stretch of a) genomic sequence, where are the CpG islands?
  TAAAAAATAAATATGTTTAATTTGTGAACTGATTACCATCAGAAT
- States: 8 States: {A+, C+, G+, T+ , A-, C-, G-, T- }
- Observations: {A,T,G,C}
- Transition probabilities matrix P: 8*8



3.隐马尔可夫模型

中文分词

小明硕士毕业于吉林大学计算机学院

小明/硕士/毕业于/吉林大学/计算机/学院

假设我们有一个标注好的数据集S={s1, s2, …, sn}，格式如上。目标是训练一个模型，对于任意一句话，能够给出分词。

3.隐马尔可夫模型

中文分词，就是给一个汉语句子作为输入，以“BEMS”组成的序列串作为输出，然后再进行切词，进而得到输入句子的划分。其中，B代表该字是词语中的起始字，M代表是词语中的中间字，E代表是词语中的结束字，S则代表是单字成词。

小明硕士毕业于中国科学院计算所

B E B E B M E B E B M E B E S

BE/BE/BME/BE/BME/BE/S

小明/硕士/毕业于/中国/科学院/计算/所

3.隐马尔可夫模型

HMM 有以下5个要素:

观测序列O: 小明硕士毕业于中国科学院计算所

状态序列S: B E B E B M E B E B M E B E S

初始概率分布π:

	P
B	-0.263
E	-3.14e+100
M	-3.14e+100
S	-1.465

状态转移概率矩阵A:

	B	E	M	S
B	-3.14e+100	-0.511	-0.916	-3.14e+100
E	-0.590	-3.14e+100	-3.14e+100	-0.809
M	-3.14e+100	-0.333	-1.260	-3.14e+100
S	-0.721	-3.14e+100	-3.14e+100	-0.666

观测概率矩阵B:

	耀	涉	谈	伊	洞	...
B	-10.460	-8.766	-8.039	-7.683	-8.669	...
E	-9.267	-9.096	-8.436	-10.224	-8.366	...
M	-8.476	-10.560	-8.345	-8.022	-9.548	...
S	-10.006	-10.523	-15.269	-17.215	-8.370	...

3.隐马尔可夫模型

中文分词这个例子属于第二个问题，即解码问题。

我们希望找到 s_1,s_2,s_3,… 使 P (s_1,s_2,s_3,…|o_1,o_2,o_3….) 达到最大。
意思是，当我们观测到语音信号 o_1,o_2,o_3,… 时，我们要根据这组信号推测出发送的句子 s_1,s_2,s_3,….，显然，我们应该在所有可能的句子中找最有可能性的一个。