隐马尔可夫模型

机器学习研究室

计算机科学与技术学院
吉林大学

目　　录

- HMM的由来
- 马尔可夫性和马尔可夫链
- HMM实例
- HMM的三个基本算法
- 主要参考文献

HMM的由来

- 俄国数学家A.A.马尔可夫于1907年第一次提出马尔科夫模型
  - 马尔可夫模型
  - 马尔可夫链
  - 隐马尔可夫模型

Andrei A Markov

- **Gifted Russian mathematician**
- **June 14, 1856 -July 20, 1922**
- **St Petersburg University**
- **Doctoral Advisor: Pafnuty Chebyshev**

http://en.wikipedia.org/wiki/Andrey_Markov

Andrei A Markov

- **Research field:**
  - **number theory**
  - **continuous fraction theory**
  - **differential equations**
  - **probability theory**
  - **statistics**

– **Markov is best known for his work in probability and for stochastic processes especially Markov chains.**

- **At the age of 30 Markov became a professor at St. Petersburg University and a member of St. Petersburg Academy of Sciences**
- **More than 120 scientific papers**
- **Classical textbook "Calculus of Probabilities"**

Pafnuty Chebyshev

- One of the founding fathers of Russian mathematics
- His famous students :
  - Dmitry Grave (格雷夫)
  - Aleksandr Korkin,
  - Aleksandr Lyapunov
  - Andrei Markov
  - ……

- According to the Mathematics Genealogy Project, Chebyshev has about 4 000 mathematical descendants.
- His famous work in the field of probability, statistics and number theory

独立事件概率

- 设想我们再作一连串的实验，而每次实验所可能发生的结果定为$E_1,E_2,\ldots E_n,\ldots$。（可能是有限，也可能是无限）。每一个结果$E_k$，我们若能给定一个出现的可能性$p_k$（即概率），则对某一特定样本之序列$E_{j1}\ E_{j2}\ldots E_{jn}$，我们知道它出现的概率是$p(E_{j1}\ E_{j2}\ldots E_{jn})=p_{j1}\ldots p_{jn}$。

非独立事件概率

- 常用的统计中，彼此相互「独立」是最有用的一个概念。简单直白地说，互相「独立」就是彼此毫不相干，一点牵涉都没有。
- 但是实际生活中很多事件是相互关联的。
- [非独立」也就是说互相关联的意思，但是要怎样相关呢？如何在相关中作一些简单的分类呢？
- 马可夫连就是要描述在「相关」这个概念中最简单的一种。但即使如此，有关马可夫链的理论已经相当丰富了。在概率理论中，它占了很大一部分内容。

非独立事件概率

- 在马可夫链中我们考虑最简单的「相关」性。在在这种情况下，我们不能给任一个事件E_j一个概率p_j。但我们给一对事件(E_j，E_k)一个概率p_{jk}，这个时候p_{jk}的解释是一种条件概率，就是假设在某次实验中E_j已经出现，而在下一次实验中E_k出现的概率。除了p_{jk}之外，我们还需要知道第一次实验中E_j出现的机率a_j。有了这些资料后，一个样本序列$E_{j0}\ E_{j1}\ldots E_{jn}$（也就是说第零次实验结果是$E_{j0}$，第一次是$E_{j1}\ldots\ldots$第$n$次实验是$E_{jn}$）的概率就很清楚的是$P(E_{j0},E_{j1},E_{jn})=a_j\ p_{j0j1}\ p_{j1j2}\cdots p_{jn\text{-}1jn}$。

马尔可夫性

- 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**

- $X(t+1) = f(X(t))$

马尔科夫链

- **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$T= \{0,1,2,\ldots\}$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$P_{ij}(m, m+n)=P\{X_{m+n}= a_j|X_m= a_i\}$称为马氏链在时刻m处于状态$a_i$条件下，而在时刻m+n转移到状态$a_j$的**转移概率**。

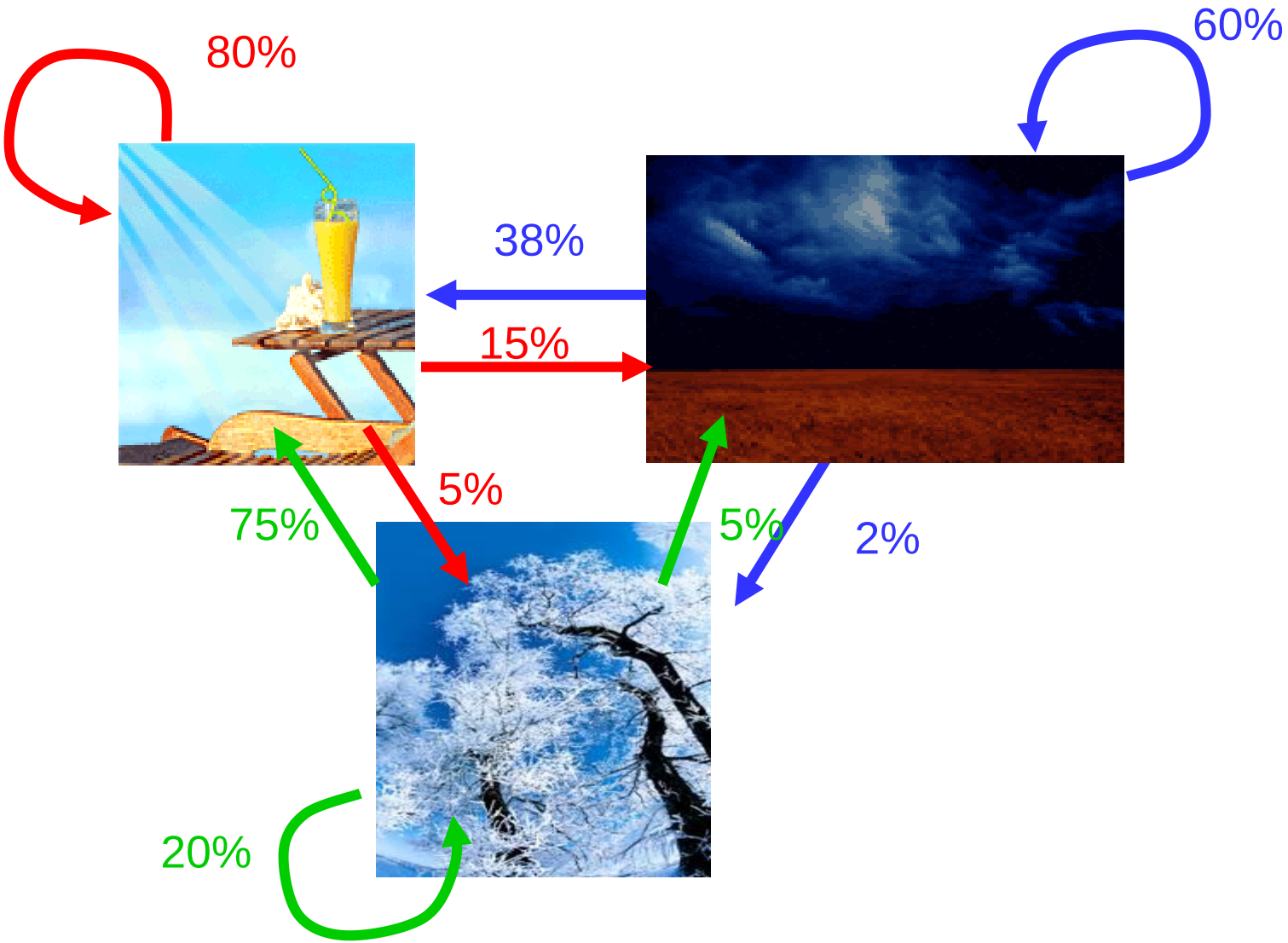
Example of Markov Chain

Weather: A Markov Model

Sunny

Rainy

Snowy


80%
60%
38%
15%
5%
75%
5%
2%
20%

Markov Model

- **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

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

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

Markov Model

- **Initial state distribution：**

$$\pi_i = P(q_1 = S_i)$$

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

Markov Model

- Markov Model=(π, A)
  - Begins (t=1) in some initial state (s).
  - At each time step (t=1, 2, …) the system moves from current to next state (possibly the same as the current state) according to transition probabilities associated with current state.

HMM的应用领域

- 语音识别
- 机器视觉
  - 人脸检测
  - 机器人足球
- 图像处理
  - 图像去噪
  - 图像识别
- 生物医学分析
  - DNA/蛋白质序列分析

Weather Markov Model

- **States:**

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

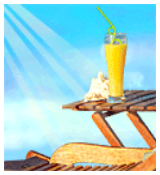
- **State transition probabilities:**

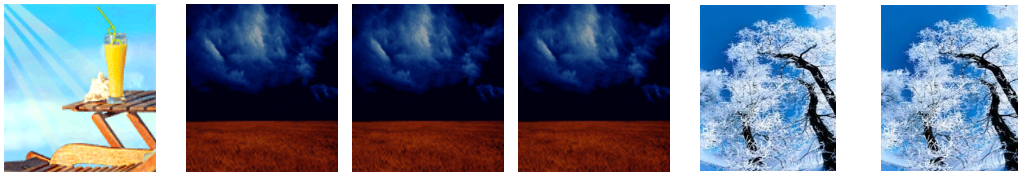
$$A = \begin{pmatrix} .8 & .15 & .05 \\ .38 & .6 & .02 \\ .75 & .05 & .2 \end{pmatrix}$$

- **Initial state distribution:**

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

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

Basic Calculations-1

$$P(S_{sunny}, S_{rainy}, S_{rainy}, S_{rainy}, S_{snowy}, S_{snowy} \mid Model)$$

$$= P(S_{sunny}) \cdot P(S_{rainy} \mid S_{sunny}) \cdot P(S_{rainy} \mid S_{rainy}) \cdot P(S_{rainy} \mid S_{rainy})$$

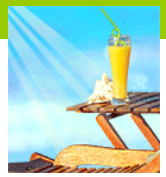
$$\cdot P(S_{snowy} \mid S_{rainy}) \cdot P(S_{snowy} \mid S_{snowy})$$

$$= 0.7 \cdot 0.15 \cdot 0.6 \cdot 0.6 \cdot 0.02 \cdot 0.2 = 0.0001512$$

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

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

Basic Calculations-2



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

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

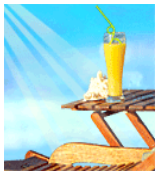
$$p(Q \mid Model, q_1 = s_i) = p(q_1 = s_i, Q \mid Model) / \; p(q_1 = s_i)$$

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

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

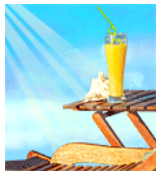
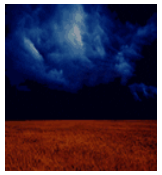
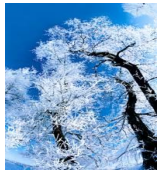
$$= a_{ii}^{\;d-1}(1 - a_{ii}) = p_i(d)$$

Basic Calculations-3

- **Conditioned on starting the weather**  **, compute the expected number of duration in weather**

无穷级数的求解

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

- **Expected number of consecutive**  **days is 5.**
- **Expected number of consecutive**  **days is 2.5.**
- **Expected number of consecutive**  **days is 1.25.**

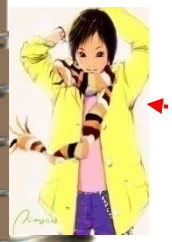
注：等比数列通项 An=A1*q^（n－1），前n项和 Sn=[A1(1-q^n)]/(1-q)

The World is not fully observable!!!

Hidden Markov Model

Example of Hidden Markov Model

60%



30%

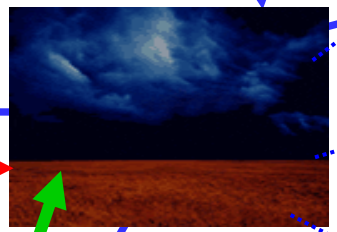
10%

Not Observable
-Hidden

80%

38%

15%



60%

5%

30%

65%

75%

5%

5%

2%

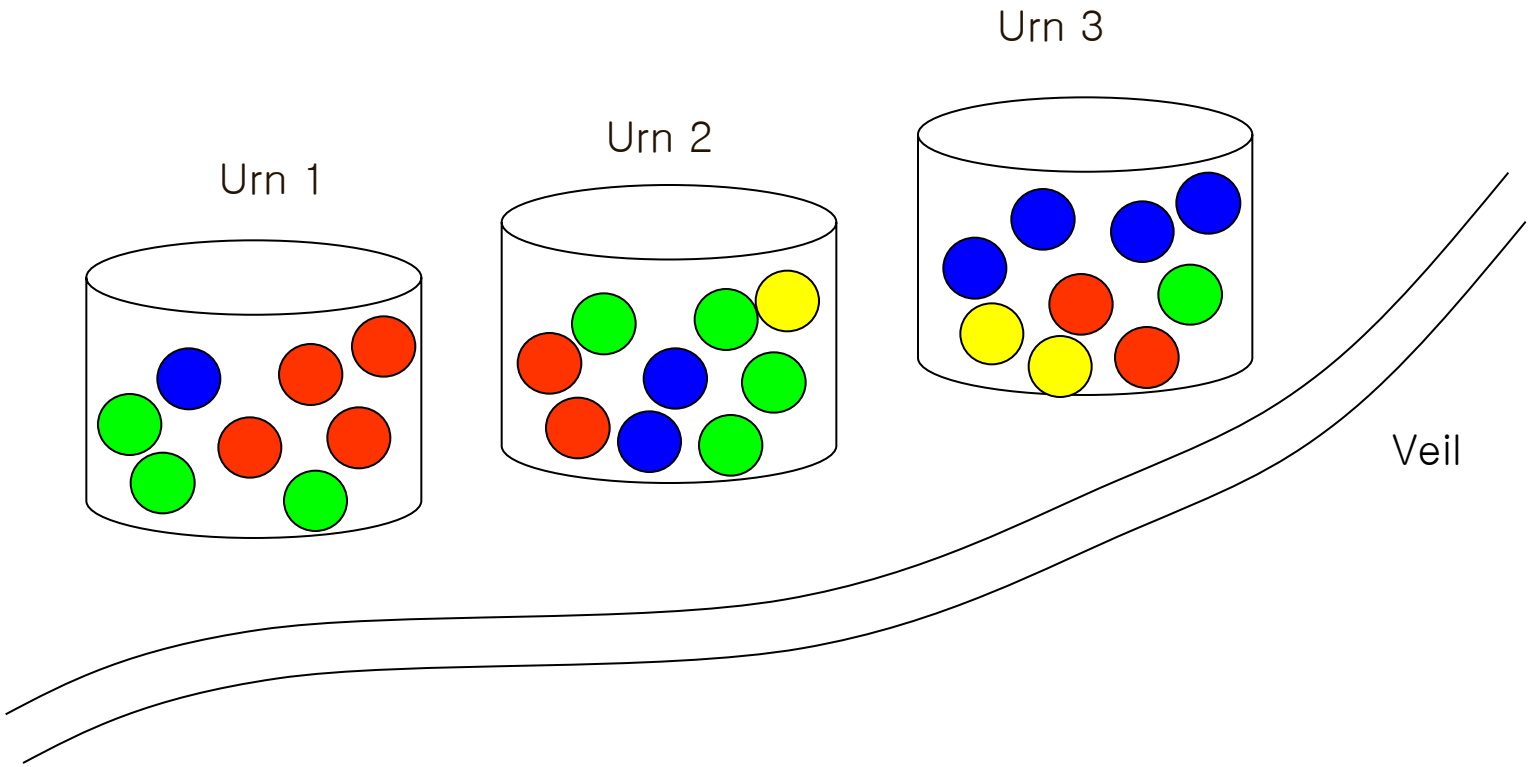
20%

0%

50%

50%

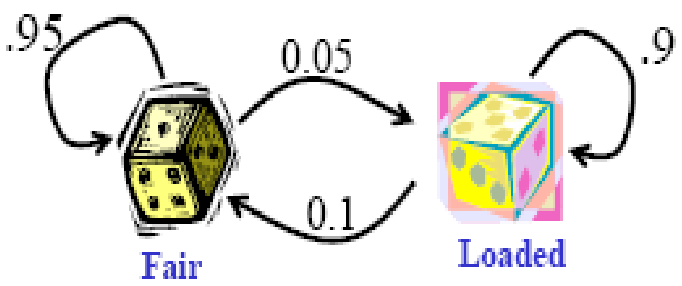
HMM实例



又一个HMM实例

- 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 1 1 1 1 1 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?

HMM实例—描述

- 设有N个缸，每个缸中装有很多彩球，球的颜色由一组概率分布描述。实验进行方式如下
  - 根据初始概率分布，随机选择N个缸中的一个开始实验
  - 根据缸中球颜色的概率分布，随机选择一个球，记球的颜色为O_1，并把球放回缸中
  - 根据描述缸的转移的概率分布，随机选择下一口缸，重复以上步骤。
- 最后得到一个描述球的颜色的序列$O_1,O_2,\ldots$，称为观察值序列O。

HMM实例——约束

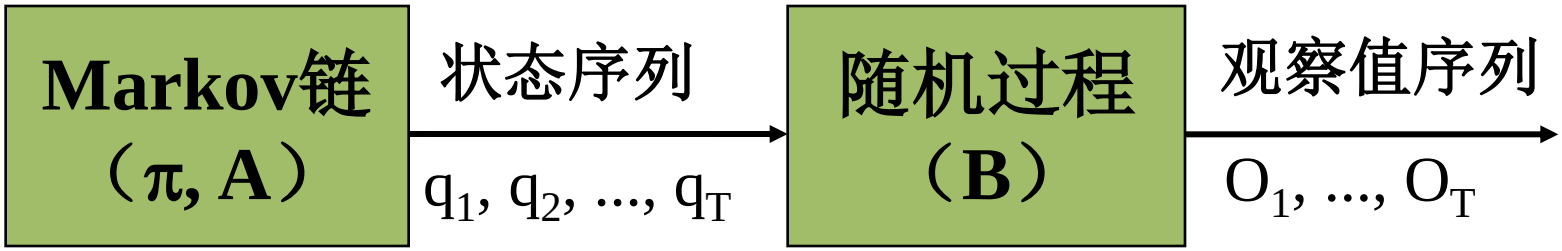
在上述实验中，有几个要点需要注意：

- 不能被直接观察缸间的转移
- 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

HMM概念

- 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。

HMM组成



HMM的组成示意图

HMM的基本要素

- 用五元组λ= (N, M, π，A，B) 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

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$P(O|\lambda)$？ 计算问题
- 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？ 解码问题
- 问题3：如何调整模型参数$\lambda=(\pi, A, B)$,使得$P(O|\lambda)$最大？ 学习问题

解决问题1 基础方法

- 给定一个固定的状态序列$S=(q_1, q_2, q_3 \ldots)$

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

其中，$b_{q_t}(O_t)$表示在q_t状态下观测到O_t的概率

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

- N=5, M=100, => 计算量10^72

计算量过大

解决问题1 前向法

- 动态规划
- 定义前向变量

第t个状态为i

$$\alpha_t(i) = P(O_1, O_2, \cdots O_t, q_t = \theta_i \mid \lambda) \qquad 1 \le t \le T$$

– 初始化：

$$\alpha_1(i) = \pi_i b_i(O_1) \qquad 1 \le i \le N$$

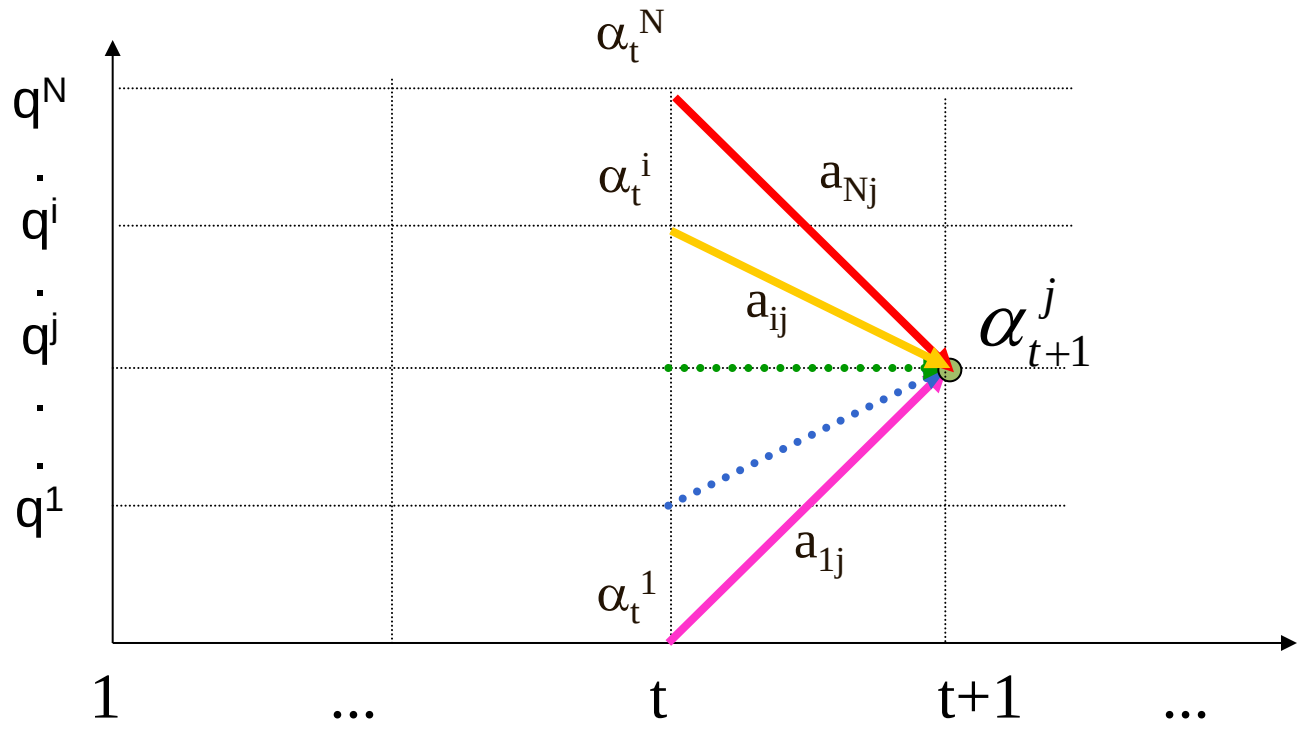
– 递归：

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_i(i) a_{ij}] b_j(O_{t+1}) \qquad 1 \le t \le T-1, 1 \le j \le N$$

– 终结：

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

前向法示意图



N=5, M=100, => 计算量3000

解决问题1 后向法

观测规律:

$$p(o_1 \ldots o_T \mid \lambda) = \sum_{i=1}^{N} \sum_{q_2 \ldots q_T} p(o_1 \ldots o_T, q_1 = s_i, q_2 \ldots q_T \mid \lambda)$$

$$= \sum_{i=1}^{N} \pi_i b_i(o_1) \sum_{q_2 \ldots q_T} p(o_2 \ldots o_T, q_2 \ldots q_T \mid q_1 = s_i, \lambda)$$

(o1…ot already generated)
Starting from at time t state s_i
Generating $o_{t+1} \ldots o_T$

算法设计:

$$\beta_t(i) = \sum_{q_{t+1} \ldots q_T} p(o_{t+1} \ldots o_T, q_{t+1} \ldots q_T \mid q_t = s_i, \lambda)$$

$$= \sum_{q_{t+1} \ldots q_T} p(o_{t+2} \ldots o_T, q_{t+2} \ldots q_T \mid q_{t+1}, \lambda) p(q_{t+1} \mid q_t = s_i, \lambda) p(o_{t+1} \mid q_{t+1}, \lambda)$$

$$= \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \sum_{q_{t+2} \ldots q_T} p(o_{t+2} \ldots o_T, q_{t+2} \ldots q_T \mid q_{t+1} = s_j, \lambda)$$

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

解决问题1 后向法

- **The data likelihood is**

$$p(o_1 \ldots o_T \mid \lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i)$$

$$= \sum_{i=1}^{N} \alpha_1(i) \beta_1(i) = \sum_{i=1}^{N} \alpha_t(i) \beta_t(i) \quad for\ any\ t$$

Complexity:

$$O(TN^2)$$

解决问题1 后向法

- 与前向法类似
- 定义后向变量

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \cdots O_T, q_t = \theta_i \mid \lambda) \qquad 1 \le t \le T-1$$

– 初始化：

$$\beta_T(i) = 1 \qquad 1 \le i \le N$$

– 递归：

$$\beta_t(i) = \sum_{i=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \quad t = T-1, T-2, \ldots, 1, 1 \le i \le N$$

– 终结：

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

解决问题2 Viterbi算法

- 目的：给定观察序列O以及模型λ,如何选择一个对应的状态序列S，使得S能够最为合理的解释观察序列O？
- N和T分别为状态个数和序列长度

定义：

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

我们所要找的，就是T时刻最大的 $\delta_T(i)$ 所代表的那个状态序列

Viterbi算法(续)

- 初始化：

$$\delta_1(i) = \pi_i b_i(O_1), \quad 1 \le \mathrm{i} \le \mathrm{N}$$

$$\varphi_1(i) = 0, \quad 1 \le \mathrm{i} \le \mathrm{N}$$

- 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序列：

$$q_t^* = \varphi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, \ldots, 1$$

Baum-Welch算法(模型训练算法)

- 目的：给定观察值序列O，通过计算确定一个模型λ， 使得$P(O|\lambda)$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 $\log P(X|\lambda) - \log(P(X|\lambda_0) < Delta$，说明训练已经达到预期效果， 算法结束。

4. 否则，令$\lambda_0 = \lambda$，继续第2步工作

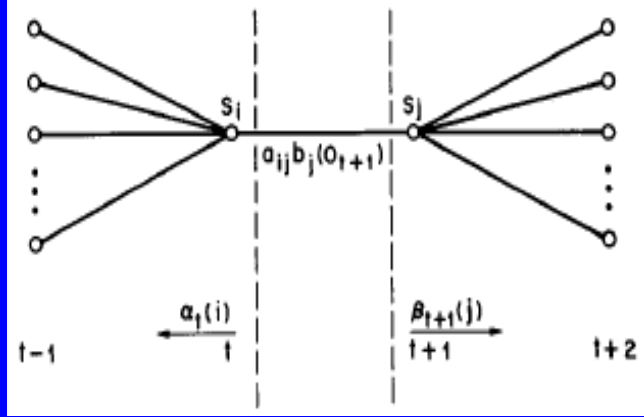
Baum-Welch算法(续)

- 定义：

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

Baum-Welch算法(续2)

- 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＝$\gamma_1(i)$

基于HMM的CpG岛识别

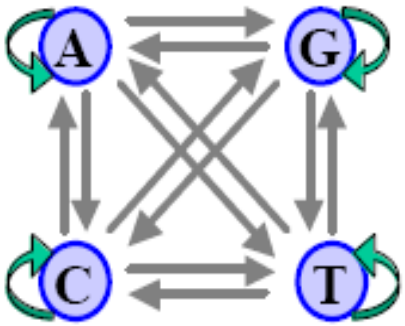
- 指DNA上一个区域，此区域含有大量相联的胞嘧啶（C）、鸟嘌呤（G），以及使两者相连的磷酸酯键（p）。哺乳类基因中的启动子上，含有约40%的CpG岛（人类约70%）。一般CpG岛的长度约300到3000个碱基对（bp）。
- 在许多基因的启动子（promotor）或“起始”区域周围，甲基化经常被抑制。这些区域包含浓度相对较高的CpG对，与此段区域对应的染色体区段一起被称作CpG岛
- CpG岛常位于管家基因和其他在细胞中被频繁表达基因的启动子区域

基于HMM的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



基于HMM的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

基于HMM的CpG岛识别

- **For a new sequence x:**

$$x = x_1 \ldots x_L$$

- **Probability of a sequence**

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

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

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

基于HMM的CpG岛识别

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

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

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

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

基于HMM的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

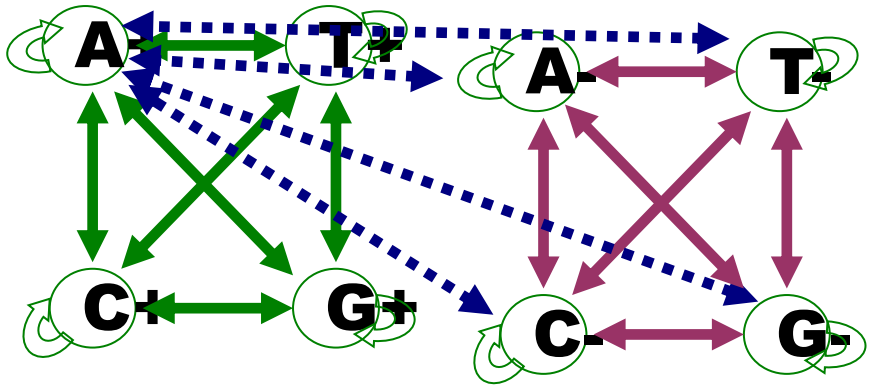
基于HMM的CpG岛识别

- 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



基于HMM的中文分词

- 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。

基于HMM的中文分词

- 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
- B E B E B E S BM M E B M E B E
- B表示分词后词语的首字，M表示分词后词语的非首非尾自（如果词语字数>=3），E表示词语的尾字，S表示单字成词
- λ＝（ N, M, π ， A，B） N=4, M=|D|（汉字总数）

基于HMM的中文分词

- 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
- B E B E B E S BM M E B M E B E
- B表示分词后词语的首字，M表示分词后词语的非首非尾自（如果词语字数>=3），E表示词语的尾字，S表示单字成词
- λ＝（ N, M, π ， A，B） N=4, M=|D|（汉字总数）
- 统计初始概率矩阵π
- 就是统计初始状态B、M、E、S的出现频率

基于HMM的中文分词

- 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
- B E B E B E S BM M E B M E B E
- B表示分词后词语的首字，M表示分词后词语的非首非尾自（如果词语字数>=3），E表示词语的尾字，S表示单字成词
- λ＝（N, M, π，A，B） N=4, M=|D|（汉字总数）
- 统计状态转移概率矩阵A

	B	M	E	S
B	b_b	b_m	b_e	b_s
M	m_b	m_m	m_e	m_s
E	e_b	e_m	e_e	e_s
S	s_b	s_m	s_e	s_s

基于HMM的中文分词

- 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
- B E B E B E S BM M E B M E B E
- B表示分词后词语的首字，M表示分词后词语的非首非尾自（如果词语字数>=3），E表示词语的尾字，S表示单字成词
- λ＝（N, M, π，A，B） N=4, M=|D|（汉字总数）
- 统计发射概率矩阵B

	B	M	E	S
字1	b_c1	m_c1	e_c1	s_c1
字2	b_c2	m_c2	e_c2	s_c2
…				
字m	b_cm	m_cm	e_cm	s_cm

基于HMM的中文词性标注

- 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
- 名词 名词 动词介词 名词 名词 名词
- $\lambda=(N, M, \pi, A, B)$ N=词性类别数, M=|C|（词的总数）

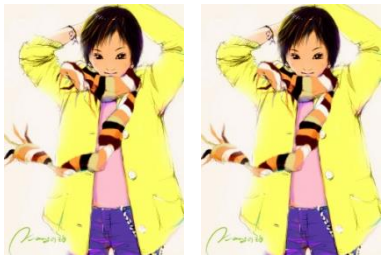
作业

- **States:** $\{S_{sunny}, S_{rainy}, S_{snowy}\}$
- **Observations:** $\{O_{skirt}, O_{coat}, O_{umbrella}\}$
- **State transition probabilities:** $A = \begin{pmatrix} .8 & .15 & .05 \\ .38 & .6 & .02 \\ .75 & .05 & .2 \end{pmatrix}$
- **emission probability :** $B = \begin{pmatrix} .6 & .3 & .1 \\ .05 & .3 & .65 \\ 0 & .5 & .5 \end{pmatrix}$
- **Initial state distribution:**

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

作业

- 给定下列观察序列O:



- **1. 出现此序列的概率为多少(前向算法)?**
- **2. 此序列对应的天气状况序列是什么?**

Pafnuty Chebyshev (契比雪夫1821-1894)