马尔可夫模型及其应用

职晓阳 xyzhi@ynu.edu.cn

目录

0.1 马尔可夫链

设 $x_1, x_2, \cdots, x_n$ 为一个有序的随机变量，根据条件概率公式，x 的联合分布律为

$$\begin{aligned} P(x_1, x_2, \cdots, x_n) &= P(x_n|x_{n-1}, \cdots, x_1) \cdot P(x_{n-1}, \cdots, x_1) \\ &= P(x_n|x_{n-1}, \cdots, x_1) \cdot P(x_{n-1}|x_{n-2}, \cdots, x_1) \cdot P(x_{n-2}, \cdots, x_1) \\ &\cdots P(x_3|x_2, x_1) \cdot P(x_2|x_1) \cdot P(x_1) \\ &= \prod_{i=2}^{n} P(x_i|x_{i-1}, \cdots, x_1) \cdot P(x_1) \end{aligned}$$

而马尔科夫链假设，在序列中的任何一个随机变量仅与变量序列中前一个变量有关，即

$$P(x_n|x_{n-1}, \cdots, x_1) = P(x_n|x_{n-1})$$

所以，

$$\begin{aligned} P(x_1, x_2, \cdots, x_n) &= \prod_{i=2}^{n} P(x_i|x_{i-1}, \cdots, x_1) \cdot P(x_1) \\ &= \prod_{i=2}^{n} P(x_i|x_{i-1}) \cdot P(x_1) \end{aligned}$$

此即**一阶马尔科夫链**的概率模型。

同理，如当前变量与序列中前两个变量有关，即为**二阶马尔科夫链**。以此类推，m **阶马尔科夫链**的表达式为

$$P(x_1, x_2, \cdots, x_n) = \prod_{i=m+1}^{n} P(x_i|x_{i-1}, x_{i-2}, \cdots, x_{i-m}) \cdot \prod_{j=1}^{m} P(x_j)$$

马尔可夫概率模型中，条件概率 $P(x_i|x_{i-1})$，常称为**转移概率**，记为 $a_{k\to l}$(由状态 $x_i = k$ 转变为状态 $x_i = l$ 的概率)。除转移概率外，模型中还有一个概率地位特殊，即 $P(x_1)$，表示有序随机变量的初始值 (或初始状态) 的概率。

0.2 隐马尔可夫链

0.2.1 基本概念

在一个满足马尔可夫模型的有序随机变量 $z_1, z_2, \cdots, z_n$ 的基础上，引入一个随机变量序列 $x_1, x_2, \cdots, x_n$，使 x_i 的取值与随机变量 z_i 的取值 (又称**状态**) 有关，且由 z_i 导致 x_i 的概率用函数 e 表示，即

$$e_k(b) = P(x_i = b|z_i = k)$$

上式表示当随机变量 z_i 取状态 k 时，导致随机变量 x_i 取值 b 的概率等于 $e_k(b)$。

需要注意的是，随机变量序列 $x_1, x_2, \cdots, x_n$ 之间相互独立，但随机变量 $z_1, z_2, \cdots, z_n$ 满足马尔可夫模型，即当前变量的取值与前一变量有关，也就是存在变量间的**转移概率** ($a_{k\to l}$)。而由 z_i 导致 x_i 的概率称为**发射概率** (emission probability)。由于实际问题中，通常只有随机变量 x 是可观察的，而有序随机变量 z 是隐匿的，因此该模型中的马尔可夫链是隐藏在可观察变量之后的，因此得名**隐马尔可夫链**。

关于隐马尔可夫模型 (Hidden Markov Model) 的应用中，掷骰子问题是较为容易理解的实例之一。假设现有三种骰子，分别为 4 面、6 面、8 面，随机选择其中一个骰子开始，并掷出一个点数，然后继续下去可得一个点数序列，如 $\{1, 3, 4, 2, 7, 2, 1, 6, 3, 5\}$。如果骰子的选择遵循某种概率条件，那么骰子的顺序就构成了一个隐马尔可夫链，所得点数序列则为观测变量。

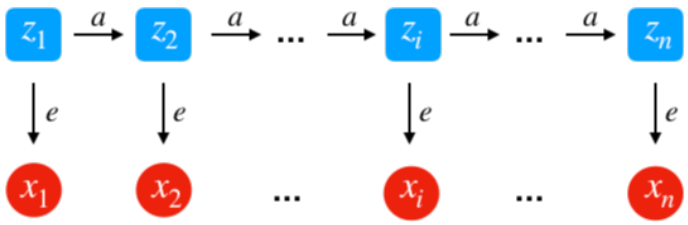

图 1: HMM 模型示意图

0.2.2 HMM 模型的要素

对于一个隐马尔可夫模型，如它的所有 N 个可能的状态由集合 $Q = \{q_1, \cdots, q_N\}$ 表示，所有 M 个可能的观测值由集合 $V = \{v_1, \cdots, v_M\}$ 表示。假设有一个长度为 T 的隐含状态序列 $I = \{i_1, \cdots, i_t, \cdots, i_T\}, i_t \in Q$，(举例说明：隐含状态集合 Q 中的某一隐含状态 q_i 作为隐含状态序列中第 t 个出现，则记为 i_t)，其中 i_t 表示在 t 时刻或位置上的隐含状态；由 I 产生的观测序列为 $O = \{o_1, \cdots, o_t, \cdots, o_T\}, o_t \in V$，(举例说明：观测值集合 V 中的某一观测值 v_i 作为观测序列中第 t 个出现，则记为 o_t)。那么构成隐马尔可夫模型的要素包括：

1. **状态转移概率矩阵 $\boldsymbol{A}$**，表示任意时刻 t 的状态 q_i, $(i = 1, \cdots, N)$ 转移为 $t+1$ 时刻的状态 q_j, $(j = 1, \cdots, N)$ 的概率。

$$\boldsymbol{A} = \begin{bmatrix} a_{q_1 \to q_1} & a_{q_1 \to q_2} & \cdots & a_{q_1 \to q_N} \\ a_{q_2 \to q_1} & a_{q_2 \to q_2} & \cdots & a_{q_2 \to q_N} \\ \vdots & \vdots & & \vdots \\ a_{q_N \to q_1} & a_{q_N \to q_2} & \cdots & a_{q_N \to q_N} \end{bmatrix} = [a_{q_i \to q_j}]_{N \times N}$$

2. **观测概率矩阵 $\boldsymbol{E}$**，表示任意时刻 t 的状态 q_i, $(i = 1, \cdots, N)$，产生观测值 v_i $(i = 1, \cdots, M)$ 的概率 (发射概率)。

$$\boldsymbol{E} = \begin{bmatrix} e_{q_1}(v_1) & e_{q_1}(v_2) & \cdots & e_{q_1}(v_M) \\ e_{q_2}(v_1) & e_{q_2}(v_2) & \cdots & e_{q_2}(v_M) \\ \vdots & \vdots & & \vdots \\ e_{q_N}(v_1) & e_{q_N}(v_2) & \cdots & e_{q_N}(v_M) \end{bmatrix} = [e_{q_i}(v_i)]_{N \times M}$$

3. **初始状态概率向量 $\boldsymbol{\pi}$**，表示在初始时刻 $t=1$ 时，状态为 $q_i,(i=1,\cdots,N)$ 的概率 $\pi_i,(i=1,\cdots,N)$ 构成的向量。

$$\boldsymbol{\Pi}=(\pi_1,\pi_2,\cdots,\pi_N)$$

因此，一个隐马尔可夫模型可由三元组 $\lambda=(\boldsymbol{A},\boldsymbol{E},\boldsymbol{\Pi})$ 来指代。有时也用五元组 $\lambda=(Q,V,\boldsymbol{A},\boldsymbol{E},\boldsymbol{\Pi})$ 表示。

0.2.3 HMM 模型的三个基本问题

0.2.3.1 解码问题 (Decoding)

问题描述：给定一个模型 $\lambda=(\boldsymbol{A},\boldsymbol{E},\boldsymbol{\Pi})$ 和一组长度为 T 的观测序列 $O=(o_1,o_2,\cdots,o_T)$，求导致观测序列的条件概率 $P(I|O)$ 最大的状态序列 $I=(i_1,i_2,\cdots,i_T)$。

Viterbi 算法

1. 第 1 个观测值的最大概率为

$$\delta_1=\max_{q_i}[\pi_i\cdot P(o_1|q_i)]=\max_{q_i}[\pi_i\cdot e_{q_i}(o_1)]$$

其中 π_i 为状态 q_i 作为初始状态出现的概率；$e_{q_i}(o_1)$ 是状态 q_i 作为初始状态而导致第 1 个观测值 o_1 的发射概率。所以第 1 个观测值的概率为 $\pi_i\cdot P(o_1|q_i)$。需要说明的是，$o_1\in V$，即观测值集合 V 中的某 1 个观测值 (如 v_i) 作为第 1 个观测值出现。所以上述表达式的意义是：**寻找能够使 $\pi_i\cdot P(o_1|q_i)$ 达到最大值的隐含状态 q_i，在隐含状态序列 I 中将 (在求第 2 个观测值的概率时) 被记为 i_1**。

2. 第 2 个观测值的最大概率为

$$\begin{aligned}\delta_2&=\max_{i_1,i_2}[\delta_1\cdot P(i_2|i_1)\cdot P(o_2|i_2)]\\&=\max_{i_1,i_2}[\pi_i\cdot e_{i_1}(o_1)\cdot a_{i_1\to i_2}\cdot e_{i_2}(o_2)]\end{aligned}$$

显然，第 2 个观测值的概率由第 1 个观测值的概率 δ_1、以及第 1 个隐含状态到第 2 个隐含状态的转移概率 $a_{i_1\to i_2}$、和第 2 个隐含状态到第 2 个观测值的发射概率 $e_{i_2}(o_2)$ 有关。所以上式的意义是：**通过取上式的最大值，确定第 1 个隐含状态 q_i，并记为 i_1；而 i_2 将在求第 3 个观测值的概率时被确定**。

3. 第 3 个观测值的最大概率为

$$\begin{aligned}\delta_3 &= \max_{i_2,i_3}[\delta_2 \cdot P(i_3|i_2) \cdot P(o_3|i_3)] \\ &= \max_{i_2,i_3}[\pi_i \cdot e_{i_1}(o_1) \cdot a_{i_1 \to i_2} \cdot e_{i_2}(o_2) \cdot a_{i_2 \to i_3} \cdot e_{i_3}(o_3)]\end{aligned}$$

上式的意义是：**通过取上式的最大值，确定第 2 个隐含状态 q_i，并记为 i_2；而 i_3 将在求第 4 个观测值的概率时被确定**。

4. 依次递归，第 T 个观测值的概率为

$$\begin{aligned}\delta_T &= \max[\delta_{T-1} \cdot P(i_T|i_{T-1}) \cdot P(o_T|i_T)] \\ &= \max[\pi_i \cdot e_{i_1}(o_1) \cdot \prod_{j=1}^{T-3} (a_{i_j \to i_{j+1}} \cdot e_{i_{j+1}}(o_{j+1})) \\ &\quad \cdot a_{i_{T-2} \to i_{T-1}} \cdot e_{i_{T-1}}(o_{T-1}) \cdot a_{i_{T-1} \to i_T} \cdot e_{i_T}(o_T)]\end{aligned}$$

上式的意义是：**通过取上式的最大值，确定第 $T-1$ 个隐含状态 q_i，并记为 i_{T-1}**。并最终在 i_{T-1} 的基础上，确定使 δ_T 最大的最后一个隐含状态 i_T。

通过上述递归运算，每次通过使当前观测序列的概率最大，确定上一个最优的隐藏状态，最终可确定概率最大的一个隐含状态序列 $i_1, i_2, \cdots, i_T$。

在实现维特比算法时需注意许多编程语言使用浮点数计算，当概率 p 很小时可能会导致结果下溢。避免这一问题的常用技巧是在整个计算过程中使用对数概率。

Viterbi 算法是一种动态规划算法。它用于寻找最有可能产生观测序列的维特比路径——隐含状态序列。由安德鲁·维特比 (Andrew Viterbi) 于 1967 年提出，用于在数字通信链路中解卷积以消除噪音。此算法被广泛应用于 CDMA 和 GSM 数字蜂窝网络、拨号调制解调器、卫星、深空通信和 802.11 无线网络中解卷积码。现今也被常常用于语音识别、关键字识别、计算语言学和生物信息学中。

0.2.3.2 评估问题 (Likelihood)

问题描述：给定一个模型 $\lambda = (\boldsymbol{A}, \boldsymbol{E}, \boldsymbol{\Pi})$ 和一组观测序列 $O = (o_1, o_2, \cdots, o_T)$，求在模型 λ 条件下观测序列 O 的概率 $P(O|\lambda)$。

直接求法

在某一个长度为 T 的隐含状态序列 $I = i_1, i_2, \cdots, i_T$ 的作用下得到观测序列 O 的联合概率为 $P(O, I|\lambda)$，对所有可能导致 O 的状态序列的联合

概率求和 $\sum_I P(O,I|\lambda)$，即得 $P(O|\lambda)$。所以问题的关键是求 $P(O,I|\lambda)$。根据条件概率公式有：

$$\begin{aligned}P(O,I|\lambda) &= P(O|I,\lambda)P(I|\lambda)\\ &= P(o_1|i_1)P(o_2|i_2)\cdots P(o_T|i_T)\times P(i_1|\lambda)P(i_2|i_1,\lambda)\cdots P(i_T|i_{T-1},\lambda)\\ &= e_{i_1}(o_1)\cdot e_{i_2}(o_2)\cdots e_{i_T}(o_T)\times \pi_i a_{i_1\to i_2}\cdots a_{i_{T-1}\to i_T}\\ &= \pi_i e_{i_1}(o_1)\cdot a_{i_1\to i_2}e_{i_2}(o_2)\cdots a_{i_{T-1}\to i_T}e_{i_T}(o_T)\end{aligned}$$

对所有可能的隐含状态序列 I 求 $\sum_I P(V,I|\lambda)$，即得 $P(V|\lambda)$。但上述直接求解的方法，需要计算的复杂度过高，事实上没有可操作性。

前向算法

定义：当第 t 个观测值的隐含状态为 q_i 时，前面的观测值分别为 $o_1,o_2,...,o_t$ 的概率，即：

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

为前向概率。其中 i_t 表示 t 时刻的隐含状态。当第 t 个观测值就是最后一个观测值 o_T 时，

$$\alpha_T(i_T) = P\big((o_1,o_2,\cdots,o_T),i_T|\lambda\big)$$

就是观测值序列 $O=(o_1,o_2,\cdots,o_T)$ 与最后一个隐含状态为 i_T 的联合概率。因此对所有可能的 i_T，求 $\sum_{i=1}^N \alpha_{i_T}(T)$，得 $P(O|\lambda)$。

1. 当 $t=1$ 时，前向概率为

$$\alpha_1(i_1) = \pi_i\cdot P(o_1|i_1) = \pi_i\cdot e_{i_1}(o_1)$$

2. 当 $t=2$ 时，前向概率为

$$\alpha_2(i_2) = \left[\sum_{i=1}^N \alpha_1(i_1)\cdot a_{i_1\to i_2}\right]\cdot e_{i_2}(o_2)$$

3. 当 $t=3$ 时，前向概率为

$$\alpha_3(i_3) = \left[\sum_{i=1}^N \alpha_2(i_2)\cdot a_{i_2\to i_3}\right]\cdot e_{i_3}(o_3)$$

4. 以此类推，当 $t=T$ 时，前向概率为

$$\alpha_T(i_T) = \left[\sum_{i=1}^N \alpha_{T-1}(i_{T-1})\cdot a_{i_{T-1}\to i_T}\right]\cdot e_{i_T}(o_T)$$

5. 最后对所有可能的终止状态 i_T 求和

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

前向算法本质上也属于动态规划的算法，也就是要通过找到局部状态递推的公式，这样一步步的从子问题的最优解拓展到整个问题的最优解。有前向概率，自然就有后向概率。

后向算法

定义：当第 t 个观测值的隐含状态为 q_i 时，后面的观测值分别观测到 $o_{t+1}, o_{t+2}, \cdots, o_T$ 的概率，即：

$$\beta_t(i_t) = P\big((o_{t+1}, o_{t+2}, \cdots, o_r)|i_t, \lambda\big)$$

为后向概率。

1. 当 $t = T$ 时，由于 o_T 后已无观测值，所以后向概率相当于必然事件的概率，所以

$$\beta_T(i_T) = P(\Omega|i_T, \lambda) = 1$$

2. 当 $t = T-1$ 时，后向概率为

$$\begin{aligned}\beta_{T-1}(i_{T-1}) &= \sum_{i=1}^{N} P(o_T|i_{T-1}, \lambda) \\ &= \sum_{i=1}^{N} a_{i_{T-1}\to i_T} \cdot e_{i_T}(o_T)\end{aligned}$$

3. 当 $t = T-2$ 时，后向概率为

$$\begin{aligned}\beta_{T-2}(i_{T-2}) &= P\big((o_{T-1}, o_T)|i_{T-2}, \lambda\big) \\ &= P(i_{T-1}|i_{T-2})P(o_{T-1}|i_{T-1}) \times \sum_{i=1}^{N} P(o_T|i_{T-1}, \lambda) \\ &= a_{i_{T-2}\to i_{T-1}} \cdot e_{i_{T-1}}(o_{T-1}) \cdot \beta_{T-1}(i_{T-1})\end{aligned}$$

4. 以此类推，当 $t = 2$ 时，后向概率为

$$\begin{aligned}\beta_2(i_2) &= P\big((o_3, \cdots, o_{T-1}, o_T)|i_2, \lambda\big) \\ &= P(i_3|i_2)P(o_3|i_3) \times \sum_{i=1}^{N} P\big((o_4, \cdots, o_{T-1}, o_T)|i_3, \lambda\big) \\ &= a_{i_2\to i_3} \cdot e_{i_3}(o_3) \cdot \beta_3(i_3)\end{aligned}$$

5. 当 $t=1$ 时，后向概率为

$$\begin{aligned}\beta_1(i_1) &= P\big((o_2,o_3,\cdots,o_{T-1},o_T)|i_1,\lambda\big)\\ &= P(i_2|i_1)P(o_2|i_2)\times\sum_{i=1}^{N}P\big((o_3,\cdots,o_{T-1},o_T)|i_2,\lambda\big)\\ &= a_{i_1\to i_2}\cdot e_{i_2}(o_2)\cdot\beta_2(i_2)\end{aligned}$$

6. 当 $t=0$ 时，后向概率为

$$\begin{aligned}\beta(0) &= P\big((o_1,o_2,\cdots,o_{T-1},o_T)|\lambda\big)\\ &= P(i_1|\lambda)P(o_1|i_1)\times\sum_{i=1}^{N}P\big((o_2,\cdots,o_{T-1},o_T)|i_1,\lambda\big)\\ &= \pi_i\cdot e_{i_1}(o_1)\cdot\beta_1(i_1)\end{aligned}$$

7. 最后对所有可能的起始状态 i_1 求和

$$P(O|\lambda)=\sum_{i=1}^{N}\pi_i\cdot e_{i_1}(o_1)\cdot\beta_1(i_1)$$

前向后向算法

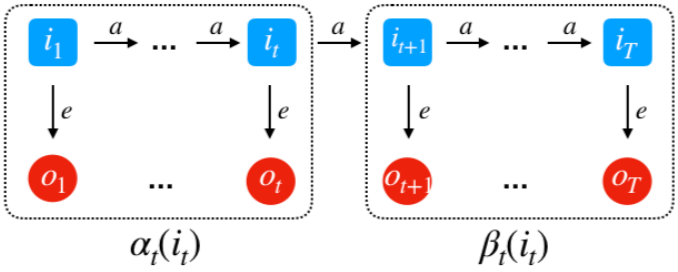

图 2: 前向概率与后向概率的关系

后向概率的动态规划递推公式和前向概率是相反的。且存在如下关系

$$\begin{aligned}P(O,i_t|\lambda) &= P(O|i_t,\lambda)\cdot P(i_t|\lambda)\\ &= P\big((o_1,o_2,\cdots,o_t,o_{t+1},\cdots,o_T)|i_t,\lambda\big)\cdot P(i_t|\lambda)\\ &= P\big((o_1,o_2,\cdots,o_t)|i_t,\lambda\big)\cdot P\big((o_{t+1},\cdots,o_T)|(o_1,o_2,\cdots,o_t),i_t,\lambda\big)\cdot P(i_t|\lambda)\\ &= P\big((o_1,o_2,\cdots,o_t),i_t|\lambda\big)\cdot P\big((o_{t+1},\cdots,o_T)|i_t,\lambda\big)\\ &= \alpha_t(i_t)\cdot\beta_t(i_t)\end{aligned}$$

所以对所有导致第 t 个观测值的隐含状态，求和得

$$\sum_{i=1}^{N} P(V, i_t|\lambda) = P(V|\lambda) = \sum_{i=1}^{N} \alpha_t(i_t) \cdot \beta_t(i_t)$$

再回过来看前向算法的结果 $P(V|\lambda) = \sum_{i=1}^{N} \alpha_T(i_T)$，以及后向算法的结果 $P(V|\lambda) = \sum_{i=1}^{N} \pi_i \cdot e_{i_1}(o_1) \cdot \beta_1(i_1)$，前向算法相当于从前至后 $1 \to T$ 递归求值，而后向算法相当于从后至前 $T \to 1$ 递归求值。而上式中，相当于将序列从 t 点分割，$1 \to t$ 利用前向概率计算，而 $t+1 \to T$ 利用后向概率计算，然后取乘积得全序列，当导致第 t 个观测值的隐含状态为 i_t 时的概率，最终对所有可能的 i_t 求和即得 $P(V|\lambda)$。此算法即前向后向算法。

0.2.3.3 学习问题 (Learning)

问题描述：通过观测序列 $O = (o_1, o_2, \cdots, o_T)$, 估计模型 $\lambda = (\boldsymbol{A}, \boldsymbol{E}, \boldsymbol{\Pi})$ 的参数，使在该模型下观测序列的概率 $P(O|\lambda)$ 最大。

有监督学习

HMM 模型参数求解根据已知的条件可以分为两种情况。第一种情况较为简单，就是已知 l 个长度为 T 的观测序列 O 和对应的隐藏状态序列 I，即 $\{(O_1, I_1), (O_2, I_2), \cdots, (O_l, I_l)\}$ 是已知的，此时可以用最大似然估计来求解模型参数。

假设样本从隐含状态 q_i 转移到 q_j 的频率计数是 A_{ij}，那么状态转移概率矩阵为：

$$\boldsymbol{A} = [a_{q_i \to q_j} = \frac{A_{ij}}{\sum_{s=1}^{N} A_{is}}]_{N \times N}$$

假设样本隐含状态为 q_i 到观测值 v_k 的频率计数是 B_{ik}，那么观测概率矩阵为：

$$\boldsymbol{E} = [e_{q_i}(v_k) = \frac{B_{ik}}{\sum_{l=1}^{M} B_{il}}]_{N \times M}$$

假设所有样本中初始隐含状态为 q_i 的频率计数为 C_{q_i}，那么初始概率分布为：

$$\boldsymbol{\Pi} = \left(\frac{C_{q_1}}{\sum_{s=1}^{N} C_{q_s}}, \cdots, \frac{C_{q_i}}{\sum_{s=1}^{N} C_{q_s}}, \cdots, \frac{C_{q_N}}{\sum_{s=1}^{N} C_{q_s}} \right)$$

可见这种情况下求解模型还是很简单的。但是在很多时候，我们无法得到 HMM 样本观察序列对应的隐藏序列，仅仅 $\{O_1, O_2, \cdots, O_l\}$ 是已知的。这种情形的解法中最常用的是鲍姆-韦尔奇算法。

有监督学习——鲍姆-韦尔奇 (Baum-Welch) 算法

假设有 m 个独立的观测序列 $\{O_1, O_2, \cdots, O_m\}$，在模型 λ 的下，有对数似然函数

$$L(\lambda) = \sum_{i=1}^{m} \ln P(O_i|\lambda) = \sum_{i=1}^{m} \ln \sum_{I} P(O_i, I|\lambda)$$

由于隐含状态 I 未知，所以无法进行最大似然估计。对 $L(\lambda)$ 做如下处理

$$\begin{aligned} L(\lambda) &= \sum_{i=1}^{m} \ln \sum_{I} P(O_i, I|\lambda) \\ &= \ln \sum_{I} P(O, I|\lambda) \end{aligned}$$

将 $P(I|O, \lambda)$ 引入上式得

$$L(\lambda) = \ln \sum_{I} P(I|O, \lambda) \frac{P(O, I|\lambda)}{P(I|O, \lambda)}$$

根据 Jensen 不等式

$$\ln \sum_{i} \lambda_i y_i \geq \sum_{i} \lambda_i \ln y_i \quad (\lambda_i \geq 0, \sum_{i} \lambda_i = 1)$$

得到其下界

$$L(\lambda) \geq \sum_{I} P(I|O, \lambda) \ln \frac{P(O, I|\lambda)}{P(I|O, \lambda)}$$

对 $L(\lambda)$ 求极值，可转变为对上述不等式中的下界求极值，因为随着下界的提高，$L(\lambda)$ 会被逼至非常小的范围，直至收敛于其极值。同时，因为下界表达式中对数函数里已经没有求和项，对参数求导并令导数为 0 时一般可以得到公式解。

但是 $L(\lambda)$ 的下界表达式中有两个的 λ(对数符号外和分数分母上的可视为同一个)。所以寻找 λ 使 $L(\lambda)$ 的下界最大时，需要固定其中一个，显然固定 $P(I|O, \lambda)$ 中的 λ 最合适，因为当给定一个参数 $\overline{\lambda}$ 时，$P(I|O, \overline{\lambda})$ 为常数 (可通过解码问题求解)，所以下界表达式可改写为

$$\sum_{I} P(I|O, \overline{\lambda}) \ln \frac{P(O, I|\lambda)}{P(I|O, \overline{\lambda})}$$

由于对似然函数求极值时，常数部分是对结果没有影响。所以在给定一个参数 $\overline{\lambda}$ 时，$P(O|\overline{\lambda})$ 是一个常数 (可通过评估问题求解)，引入下界表达式

得

$$\sum_I P(I|O,\overline{\lambda}) \cdot P(O|\overline{\lambda}) \ln \frac{P(O,I|\lambda)}{P(I|O,\overline{\lambda})}$$

同理也可将 $P(I|O,\overline{\lambda})$ 作常数处理，并根据条件概率公式得

$$\sum_I P(O,I|\overline{\lambda}) \ln P(O,I|\lambda)$$

上式可视为含有两个自变量 $\lambda, \overline{\lambda}$ 的函数

$$\mathrm{Q}(\lambda,\overline{\lambda}) = \sum_I P(O,I|\overline{\lambda}) \ln P(O,I|\lambda)$$

将上式中对数符号内外的概率都变成 $P(O,I|\lambda)$ 的形式，是因为该表达式的结果已在评估问题中解决，即

$$P(O,I|\lambda) = \pi_i e_{i_1}(o_1) \cdot a_{i_1 \to i_2} e_{i_2}(o_2) \cdots a_{i_{T-1} \to i_T} e_{i_T}(o_T),$$

同时转变后, 函数 Q 中将不再含有未知量 I, 而仅与模型参数 $\lambda = (\boldsymbol{A}, \boldsymbol{E}, \boldsymbol{\Pi})$ 有关。所以从一个随机参数 $\overline{\lambda_1}$ 开始，代入上述不等式，然后对函数 Q 中 $P(O,I|\lambda)$ 部分的 λ 求最大似然估计 $\overline{\lambda_2}$，借以在第二次估计中，替换 $\overline{\lambda_1}$，依次迭代下去，逐步提高下限函数的值，直至收敛。这就是鲍姆-韦尔奇算法的基本思想。

将 $P(V,Q|\lambda)$ 代入不等式得

$$\begin{aligned}
\mathrm{Q}(\lambda,\overline{\lambda}) &= \sum_I P(O,I|\overline{\lambda}) \ln \left(\pi_i e_{i_1}(o_1) \cdot a_{i_1 \to i_2} e_{i_2}(o_2) \cdots a_{i_{T-1} \to i_T} e_{i_T}(o_T) \right) \\
&= \sum_I P(O,I|\overline{\lambda}) \left(\ln \pi_i + \sum_{t=1}^{T-1} \ln a_{i_t \to i_{t+1}} + \sum_{t=1}^{T} \ln e_{i_t}(o_t) \right) \\
&= \sum_I P(O,I|\overline{\lambda}) \ln \pi_i + \sum_I P(O,I|\overline{\lambda}) \sum_{t=1}^{T-1} \ln a_{i_t \to i_{t+1}} + \sum_I P(O,I|\overline{\lambda}) \sum_{t=1}^{T} \ln e_{i_t}(o_t)
\end{aligned}$$

接下来分别对 π, a, e 求偏导，每次都只留下与相应参数有关的项。

1. 求 π

上式第一项可改写为

$$\sum_I P(O,I|\overline{\lambda}) \ln \pi_i = \sum_{j=1}^{N} P(O, i_1 = q_j|\overline{\lambda}) \ln \pi_j$$

由于 π_j 还满足 $\sum_{j=1}^{N}\pi_j=1$，因此根据拉格朗日乘子法，得到拉格朗日函数：

$$\sum_{j=1}^{N}P(O,i_1=q_j|\overline{\lambda})\ln\pi_j+\gamma\Big(\sum_{j=1}^{N}\pi_j-1\Big)$$

对拉格朗日函数求偏导并令结果为 0

$$\frac{\partial}{\partial\pi_j}\Big[\sum_{j=1}^{N}P(O,i_1=q_j|\overline{\lambda})\ln\pi_j+\gamma\Big(\sum_{j=1}^{N}\pi_j-1\Big)\Big]=0$$

得

$$P(O,i_1=q_j|\overline{\lambda})+\gamma\pi_j=0$$

对 j 求和得

$$\begin{aligned}\sum_{j=1}^{N}P(O,i_1=q_j|\overline{\lambda})+\gamma\sum_{j=1}^{N}\pi_j&=0\\ \gamma\sum_{j=1}^{N}\pi_{q_j}&=-\sum_{j=1}^{N}P(O,i_1=q_j|\overline{\lambda})\\ \gamma&=-P(O|\overline{\lambda})\end{aligned}$$

再将 γ 代回上式，得

$$\pi_j=\frac{P(O,i_1=q_j|\overline{\lambda})}{P(O|\overline{\lambda})}$$

2. 求 a

第二项可改写为

$$\sum_{I}P(O,I|\overline{\lambda})\sum_{t=1}^{T-1}\ln a_{i_t\to i_{t+1}}=\sum_{j}^{N}\sum_{k}^{N}P(O,i_t=q_j,i_{t+1}=q_k|\overline{\lambda})\sum_{t=1}^{T-1}\ln a_{i_t\to i_{t+1}}$$

与求 π 时类似，利用约束条件 $\sum_{i=1}^{N}a_{i_t\to i_{t+1}}=1$ 的拉格朗日乘子法，求得

$$a_{q_j\to q_k}=\frac{\sum_{t=1}^{T-1}P(O,i_t=q_j,i_{t+1}=q_k|\overline{\lambda})}{\sum_{t=1}^{T-1}P(O,i_t=q_j|\overline{\lambda})}$$

3. 求 e

第三项可改写为

$$\sum_{I}P(O,I|\overline{\lambda})\sum_{t=1}^{T}\ln e_{i_t}(o_t)=\sum_{j=1}^{N}\sum_{k=1}^{M}P(O,i_t=q_j,o_t=v_k|\overline{\lambda})\sum_{t=1}^{T}\ln e_{i_t}(o_t)$$

同样利用约束条件为 $\sum_{t=1}^{M} e_{q_i}(o_t) = 1$ 的拉格朗日乘子法，求得

$$e_{q_j}(v_k) = \frac{\sum_{t=1}^{T} P(O, i_t = q_j, o_t = v_k|\overline{\lambda})}{\sum_{t=1}^{T} P(O, i_t = q_j|\overline{\lambda})}$$

在评估问题中，利用前向后向算法，可得在 t 位置上的状态为 q_i 的概率为

$$\begin{aligned}\gamma_t(q_i) &= P(i_t = q_i|O, \lambda) \\ &= \frac{P(i_t = q_i, O|\lambda)}{P(O|\lambda)} \\ &= \frac{\alpha_t(i_t) \cdot \beta_t(i_t)}{\sum_{i=1}^{N} \alpha_t(i_t) \cdot \beta_t(i_t)}\end{aligned}$$

而在 t 位置上的状态为 q_i，在 $t+1$ 位置上的状态为 q_j 的概率为

$$\begin{aligned}\xi_t(q_i, q_j) &= P(i_t = q_i, i_{t+1} = q_j|O, \lambda) \\ &= \frac{P(i_t = q_i, i_{t+1} = q_j, O|\lambda)}{P(O|\lambda)} \\ &= \frac{\alpha_t(i_t) \cdot a_{i_t \to i_{t+1}} e_{i_{t+1}}(o_{t+1}) \cdot \beta_{t+1}(i_{t+1})}{\sum_{i=1}^{N} \alpha_t(i_t) \cdot a_{i_t \to i_{t+1}} e_{i_{t+1}}(o_{t+1}) \cdot \beta_{t+1}(i_{t+1})}\end{aligned}$$

利用 $\gamma_t(q_i)$ 和 $\xi_t(q_i, q_j)$，可得 π, a, e 的求解公式

$$\begin{aligned}\pi_i &= \gamma_1(q_i) \\ a_{q_i \to q_j} &= \frac{\sum_{t=1}^{T} \xi_t(q_i, q_j)}{\sum_{t=1}^{T} \gamma_t(q_i)} \\ e_{q_i}(v_k) &= \frac{\sum_{t=1, v_t = v_k}^{T} \gamma_t(q_i)}{\sum_{t=1}^{T} \gamma_t(q_i)}\end{aligned}$$

完成一次参数估计后，将新的 λ 估计带回函数 Q，进行下一轮最大似然参数估计，直至参数取值收敛。

鲍姆-韦尔奇算法，本质上也就是就是最大期望 (Expectation-Maximization，EM) 算法，是一类通过迭代进行极大似然估计的优化算法，通常作为牛顿迭代法 (Newton-Raphson method) 的替代用于对包含隐变量或缺失数据的概率模型进行参数估计。

0.3 马尔可夫链的应用实例

0.3.1 CpG 岛预测

CpG 双核苷酸在人类基因组中的分布很不均一，而在基因组的某些区段, CpG 保持或高于正常概率。CpG 岛主要位于基因的启动子（promotor）和第一外显子区域，约有 60% 以上基因的启动子含有 CpG 岛。CpG 岛的 GC 含量大于 50%, 长度 500~1000bp。许多基因, 尤其是管家基因的启动子区，其中通常存在一些富含双核苷酸 “CG” 的区域，称为 “CpG 岛”（CpG island）。研究碱基 G 和 C 在整个基因组内的含量和分布有十分重要的意义。例如在人类基因组内，GC 的含量大约为 40%；这些 GC 并不是平均分布在基因组内，在某些 DNA 片段上其含量可高达 60% 以上，而在另一些区域则只有 33% 左右。这种 GC 含量的差别，在基因表达的调控和基因突变上都可能扮演着重要的角色。

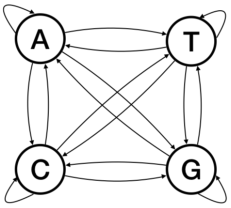

图 3: DNA 的马尔可夫链

双核苷酸在序列构成上是重要的, 因此需要一个模型在生成序列时一个碱基的概率依赖于前一个碱基，就像经典的一阶马尔可夫模型所规定的那样。在图3所示的模型中，每个箭头都与一个状态转移概率相关联，决定了一个碱基跟随另一个碱基的概率，用 a_{st} 表示为

$$a_{st} = P(x_i = t|x_{i-1} = s)$$

连续使用该概率公式，一条长度为 L 的序列的概率可表示为

$$\begin{aligned} P(x) &= P(x_L, x_{L-1}, \cdots, x_1) \\ &= P(x_L|x_{L-1}, \cdots, x_1)P(x_{L-1}|x_{L-2}, \cdots, x_1)\cdots P(x_1) \end{aligned}$$

根据一阶马尔可夫链的关键性质：当前状态 (字符) 的概率仅依赖于前一个

状态 (字符) 的概率。所以 $P(x_L|x_{L-1},\cdots,x_1)=P(x_L|x_{L-1})$，于是

$$\begin{aligned}P(x)&=P(x_L|x_{L-1})P(x_{L-1}|x_{L-2})\cdots P(x_2|x_1)P(x_1)\\&=P(x_1)\prod_{i=2}^{L}a_{x_{i-1}x_i}\end{aligned}$$

以上方程可用于计算任意马尔可夫链生成的特定序列的概率。注意观察上式，我们会发现序列的起始字符的概率形式上独立于概率链之外。为了让模型更贴合 DNA 序列的直观，可以在模型中加入一个额外的状态对应于序列的起始 (并非第一个字符)，类似地，也可以独立地为序列结尾建模。至此，DNA 序列的马尔可夫模型可以图4来表示，在该模型中起始和结尾都是沉默状态，因此在转移概率矩阵中无需添加专门的行和列以对应起始和结尾 (即增加 ATCG 之外的字符)。

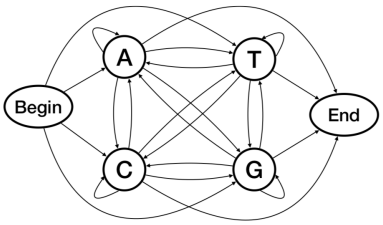

图 4: 起始与结尾独立建模的 DNA 马尔可夫链模型

回到 CpG 岛的问题上，根据 DNA 序列的马尔可夫模型，我们可以利用已知的 CpG 岛序列，对转移概率矩阵中的各项参数作出估计。同理，我们也可以对非 CpG 岛序列，估计转移概率矩阵中的各项参数。下图是来自一组人类 DNA 序列中提取的共 48 个推测的 CpG 岛序列，建立的两个马尔可夫链模型，一个针对 CpG 岛 ("+" 模型)，一个针对非 CpG 岛 ("-" 模型)。

+	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

-	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

图 5: CpG 岛与非 CpG 岛马尔可夫链的转移概率矩阵

其中转移概率的 (最大似然) 估计公式为

$$a_{st}^{+} = \frac{c_{st}^{+}}{\sum_{N=A,T,C,G} c_{sN}^{+}}$$

其中 c_{st}^{+} 表示在 CpG 岛区域内碱基 t 尾随碱基 s 的次数。在转移概率矩阵中，每一行的和为 1，因为某一碱基之后出现任意碱基是必然事件；特别地，该矩阵是非对称的，两个矩阵中 C 之后出现 G 的概率都低于 G 之后出现 C。

用马尔可夫模型来判断一条 DNA 序列是否属于 CpG 岛，我们可以利用似然比检验。为方便计算，似然比作对数处理，得对数似然比

$$S(x) = \log \frac{P(x|\text{model } +)}{P(x|\text{model } -)} = \sum_{i=1}^{L} \log \frac{a_{x_{i-1}x_i}^{+}}{a_{x_{i-1}x_i}^{-}}$$

那么对所有 CpG 岛区域和非 CpG 岛区域计算对数似然比，再经长度归一化后，可以得到图6所示的直方图 (取 2 为底的对数后，单位称为比特)，其中灰色表示非 CpG 岛区域，深灰色表示 CpG 岛区域。进而，对于一条待分析的序列，可根据 CpG 岛区域对数似然比的分布，进行假设检验，以判断是否属于 CpG 岛区域。其中需要特别注意的是，判断错误的情形可能与不充分或不正确的模型有关，也可能有训练数据的错误标注有关。例如当前数据中就有一个标记为 CpG 岛区域，但其对数似然比值却在非 CpG 岛区域的分布范围的最左侧，显然需要重新审视该条序列的标注结果。

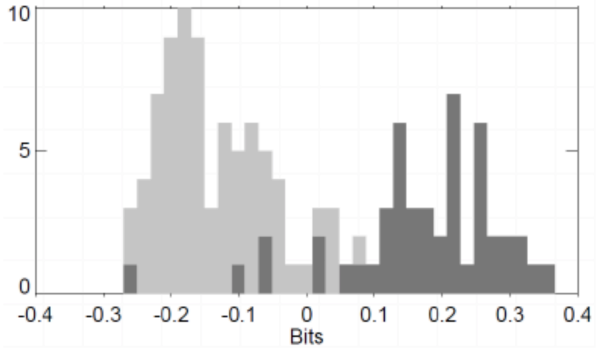

图 6: 长度归一后所有序列对数似然比的直方图

那么如何从一条长序列中找到其中隐藏的 CpG 岛区域呢？我们可以将长序列从起始位置开始分割成预设长度的 (如 100bp) 子序列，利用上述

马尔可夫模型计算子序列的对数似然比；向后移动一个步长 (如 10bp)，再计算新子序列的对数似然比；依次计算所有子序列的对数似然比，理论上 CpG 岛区域的对数似然比将取正值 (图6)。这样就可以将 CpG 岛区域从长序列中检出，但是这种方法难以对 CpG 岛区域的边界进行确定。为解决这一问题，就需要借助隐马尔可夫模型 (HMM)。

根据 HMM 模型的定义，我们可将 DNA 序列的不同碱基作为观测状态，将 CpG 岛区域和非 CpG 岛区域作为两种隐含状态 (+ 和 −)。隐含状态依状态转移概率矩阵相互转换，公式依观测概率矩阵发射不同的观测状态 (即 A、T、C、G 四种碱基)。因此，识别一条长序列中的 CpG 岛区域问题就转变成了 HMM 模型中的解码问题，可以通过 Viterbi 算法解决。

0.3.2 启动子识别

0.3.3 基因预测

利用 HMM 模型对 DNA 序列的隐含状态进行建模，是一种普遍适用的方法。模型的结构因需要解决的问题而异，CpG 岛的预测中，隐含状态有两个分别对应 CpG 岛和非 CpG 岛。如果这两种隐含状态理解为基因编码区和非基因编码区，即得对基因预测问题的最简单的一种 HMM 模型 (most simple gene predictor, MSGP)。类似的利用 Viterbi 算法进行解码。

但是基因预测问题显然要比 CpG 岛的预测要复杂，特别是针对真核生物的基因预测问题。因此对应的 HMM 模型结构也要复杂的多，其隐含状态至少有三种。

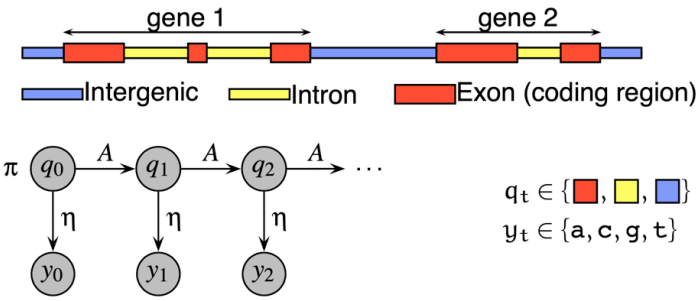

图 7: 真核生物基因预测的 HMM 模型