Hidden Markov Model

Markov Model

A1 -> A2 -> A3 ···> Ai -> Ai+1 ···> An ->

M 个离散结点形成一条链

- 若每个结点有 K 个状态，则需要 K-1 + (M-1)K(K-1) 个参数，O(M)
- 如果是全连接，则需要 K^M -1 个参数，O(K^M)

A1 会有 K 个变化，比方当 A1=1 时，需要给出 A2 = i 这么多个参数, i 为 (0,k)，所以是 K(K-1) 个参数

链状模型可以极大减少我们需要学习的参数的数目

若状态 Ai 确定，则 Ai+1 只与 Ai 有关，与 A1,...,Ai-1 无关，伪随机过程

$$p(\theta^{t+1} \mid \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t)}) = p(\theta^{t+1} \mid \theta^{(t)})$$

例子：C 语言中的伪随机函数使用的是 Markov Model，实际上只要知道第 i 个随机数，就可以计算出第 i+1 个随机数

满足 Markov 模型的样本点 s1,s2...sn 与 |『独立同分布』只弱一点：样本间只有相邻元素有相关关系，并且，有些以『独立同分布』为条件的定理，是可以放松到 Markov Process 的。例如：大数定理。

Hidden Markov Model

HMM 是关于时序的概率模型，描述由一个隐藏的 Markov Chain 随机生成的不可观测的状态最忌序列，再有各个状态生成一个观测二产生观测随机序列的过程。

HMM 随机生成的状态序列，称为状态序列；每个状态生成一个观测，由此产生的光测随机序列，称为观测序列。序列的每个位置可以看做是一个时刻

$$
\begin{array}{ccccccccccc}
z_1 & \rightarrow & z_2 & \rightarrow & \cdots & \rightarrow & z_{n-1} & \rightarrow & z_n & \rightarrow & z_{n+1} & \rightarrow \\
\downarrow & & \downarrow & & & & \downarrow & & \downarrow & & \downarrow \\
x_1 & & x_2 & & & & x_{n-1} & & x_n & & x_{n+1}
\end{array}
$$

在 z1, z2 不可观察的前提下

- x1 和 z2 独立吗？不独立
- x1 和 x2 独立吗？不独立

HMM 的确定

由出事概率分布 π，状态转移分布 A 以及观测概率分布 B 确定。

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

假设我们的 HMM 模型如下

$$
\begin{array}{ccccccccccc}
z_1 & \rightarrow & z_2 & \rightarrow & \cdots & \rightarrow & z_{n-1} & \rightarrow & z_n & \rightarrow & z_{n+1} & \rightarrow \\
\downarrow & & \downarrow & & & & \downarrow & & \downarrow & & \downarrow \\
x_1 & & x_2 & & & & x_{n-1} & & x_n & & x_{n+1}
\end{array}
$$

- Q 是所有可能的状态的集合
  - N 是可能的状态数目
  - $Q = \{q_1, q_2, \ldots, q_N\}$
- V 是所有可能的观测的集合
  - M 是可能的观测数
  - $V = \{v_1, v_2, \ldots, v_M\}$

I 是长度为 T 的状态序列，O 是对应的观测序列

$$
\begin{aligned}
I = \{i_1, i_2, \ldots, i_T\} \\
O = \{o_1, o_2, \ldots, o_T\}
\end{aligned}
$$

A 是状态转移概率矩阵

$$
\begin{aligned}
A = [a_{ij}]_{N \times N} \\
a_{ij} = P(i_{t+1} \mid i_t = q_i)
\end{aligned}
$$

a_{ij} 是在时刻 t 处于 状态 q_i 的条件下时刻 t+1 转移到概率 q_j 的概率

B 是观测概率矩阵

$$B = [b_{ik}]_{N\times M}$$
$$b_{ik} = P(o_t = v_k \mid i_t = q_i)$$

b_{ik} 是在时刻 t 处于状态 q_i 的条件下生成观测 v_k 的概率

π 是初始状态概率向量

$$\pi = (\pi_i)$$
$$\pi_i = P(i_1 = q_i)$$

π_i 是时刻 t=1 处于状态 q_i 的概率

两个基本性质

$$\begin{array}{ccccccccccc} z_1 & \rightarrow & z_2 & \rightarrow & \cdots & \rightarrow & z_{n-1} & \rightarrow & z_n & \rightarrow & z_{n+1} & \rightarrow \\ \downarrow & & \downarrow & & & & \downarrow & & \downarrow & & \downarrow & \\ x_1 & & x_2 & & & & x_{n-1} & & x_n & & x_{n+1} & \end{array}$$

齐次假设

$$P(i_t \mid i_{t-1}, o_{t-1}, i_{t-2}, o_{t-2}, \ldots, i_1, o_1) = P(i_t \mid i_{t-1})$$

这里 i_t 对应 z_n 输入的概率，o_{t-1} 这些对应 x_{n-1} 的观测，也就是说，z_n 的状态实际上只由 z_{n-1} 决定

观测独立性假设

$$P(o_t \mid i_t, i_{t-1}, o_{t-1}, i_{t-2}, o_{t-2}, \ldots, i_1, o_1) = P(o_t \mid i_t)$$

也就是说在时刻 t，得到观测 o_t 的概率只跟 i_t 有关（通过查 B 矩阵即可）

举例

假设有 3 个盒子，编号为 1 2 3，每个盒子都装有红白两种颜色的小球

盒子号	**1**	**2**	**3**

红球	5	4	7
白球	5	6	3

按照下面的方法抽取小球，得到球颜色的观测序列：

- 按照 $\pi = (0.2, 0.4, 0.4)$ 来选择一个盒子，抽取一个球后放回
- 按某条件概率选择新的盒子，重复上述过程
- 最终得到观测序列：红红白白红

各个参数

- 状态集合：Q={盒子1，盒子2，盒子3}
- 观测集合：V={红，白}
- 状态序列和观测序列的长度 T=5
- 初始概率分布 $\pi = (0.2, 0.4, 0.4)$
- 状态转移概率分布 A
- 观测概率分布 B (一行是一个分布)

$$\pi = \begin{pmatrix} 0.2 \\ 0.4 \\ 0.4 \end{pmatrix} \quad A = \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{pmatrix} \quad B = \begin{pmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{pmatrix}$$

三个基本问题

概率计算问题

给定模型 $\lambda = (A, B, \pi)$ 和观测序列 $O = \{o_1, o_2, \ldots, o_T\}$，计算模型 λ 下观测序列 O 出现的概率 $P(O \mid \lambda)$

学习问题

已知观测序列 $O = \{o_1, o_2, \ldots, o_T\}$，估计模型 $\lambda = (A, B, \pi)$ 的参数，使得在该模型下观测序列 $P(O \mid \lambda)$ 最大

预测问题

即解码问题，已知模型 $\lambda = (A, B, \pi)$ 和观测序列 $O = \{o_1, o_2, \ldots, o_T\}$，求对给

定观测序列条件概率 P(I\;|\;O) 最大的状态序列 I。

概率计算问题

直接计算法

按照概率公式，列举所有可能的长度为 T 的状态序列 $I=\{i_1,i_2,\ldots,i_T\}$，求各个状态序列 I 与观测序列 $O=\{o_1,o_2,\ldots,o_T\}$ 的联合概率 $P(O,I\mid\lambda)$，然后对所有可能的状态求和，从而得到 $P(O\mid\lambda)$

状态序列 $I=\{i_1,i_2,\ldots,i_T\}$ 的概率是：

$$P(I\mid\lambda)=\pi_{i_1}a_{i_1i_2}a_{i_2i_3}\ldots a_{i_{T-1}i_T}$$

对固定的状态序列 I，观测序列 O 的概率是

$$P(O\mid I,\lambda)=b_{i_1o_1}b_{i_2o_2}\ldots b_{i_To_T}$$

O 和 I 同时出现的联合概率

$$\begin{aligned}P(O,I\mid\lambda)&=P(O\mid I,\lambda)P(I\mid\lambda)\\&=\pi_{i_1}b_{i_1o_1}a_{i_1i_2}b_{i_2o_2}a_{i_2i_3}\ldots a_{i_{T-1}i_T}b_{i_To_T}\end{aligned}$$

对所有可能的状态序列 I 求和，得到观测序列 O 的概率 $P(O\mid\lambda)$

$$\begin{aligned}P(O\mid\lambda)&=\sum_I P(O,I\mid\lambda)=\sum_I P(O\mid I,\lambda)P(I\mid\lambda)\\&=\sum_{i_1,i_2,\ldots,i_T}\pi_{i_1}b_{i_1o_1}a_{i_1i_2}b_{i_2o_2}a_{i_2i_3}\ldots a_{i_{T-1}i_T}b_{i_To_T}\end{aligned}$$

分析：加和符号中有 2T 个因子，I 的遍历个数为 N^T，时间复杂度为 O(T N^T)，复杂度过高

前向算法

定义：给定 λ，定义到时刻 t，部分观测序列为 $o_1,o_2,\ldots,o_t$ 且隐状态为 q_i 的概率为 **前向概率**，记作

$$\alpha_t(i)=P(o_1,o_2,\ldots,o_t,i_t=q_i\mid\lambda)$$

可以递推计算前向概率 $\alpha_t(i)$ 以及观测序列概率 $P(O \mid \lambda)$

初值：$\alpha_1(i) = \pi_i b_{io_1}$

递推，对于 $t = 1, 2, \dots, T-1$，这里的 a_{ji} 是矩阵 A 中从状态 j 转移到状态 i 的值

$$\alpha_{t+1}(i) = \left(\sum_{j=1}^{N} \alpha_t(j) a_{ji}\right) b_{io_{t+1}}$$

最终

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

前向概率算法的时间复杂度是 O(TN^2)

例子

还是用刚才的小球的例子，计算观测向量 O=红白红 的出现概率，其中

$$\pi = \begin{pmatrix} 0.2 \\ 0.4 \\ 0.4 \end{pmatrix} \quad A = \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{pmatrix} \quad B = \begin{pmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{pmatrix}$$

计算初值，o_1表示红色，o_2表示白色

$$\begin{aligned} \alpha_1(\text{盒子}1) &= \pi_1 b_{1_{o_1}} = 0.2 \times 0.5 = 0.1 \\ \alpha_1(\text{盒子}2) &= \pi_2 b_{2_{o_1}} = 0.4 \times 0.4 = 0.16 \\ \alpha_1(\text{盒子}3) &= \pi_3 b_{2_{o_1}} = 0.4 \times 0.7 = 0.28 \end{aligned}$$

递推

$$\begin{aligned} \alpha_2(i=1) &= \left(\sum_{j=1}^{N} \alpha_1(j) a_{j1}\right) b_{1_{o_2}} \\ &= (0.1 \times 0.5 + 0.16 \times 0.3 + 0.28 \times 0.2) \times 0.5 \\ &= 0.077 \end{aligned}$$

$$\alpha_2(2) = 0.1104$$
$$\alpha_2(3) = 0.606$$

$$\alpha_3(1) = 0.04187$$
$$\alpha_3(2) = 0.03551$$
$$\alpha_3(3) = 0.05284$$

最终

$$\begin{aligned} P(O \mid \lambda) &= \sum_{i=1}^{3} \alpha_3(i) \\ &= 0.04187 + 0.03551 + 0.05284 \\ &= 0.13022 \end{aligned}$$

后向算法

定义：给定 λ，定义到时刻 t 隐状态为 q_i 的前提下，从 t+1 到 T 的部分观测序列概率为 $o_{t+1}, o_{t+2}, \ldots, o_T$ 的概率为**后向概率**，记作

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

可以递推计算前向概率 $\beta_t(i)$ 以及观测序列概率 $P(O \mid \lambda)$

初值：$\beta_T(i) = 1$

递推，对于 $t = T-1, T-2, \ldots, 1$，这里的 a_{ij} 是矩阵 A 中从状态 i 转移到状态 j 的值

$$\beta_t(i) = \left(\sum_{j=1}^{N} a_{ij} b_{j o_{t+1}} \beta_{t+1}(j) \right)$$

最终

$$P(O \mid \lambda) = \sum_{i=1}^{N} \pi_i b_{i o_1} \beta_1(i)$$

为了计算在时刻 t 隐状态为 q_i 条件下，时刻 t+1 之后观测序列为 $o_{t+1}, o_{t+2}, \ldots, o_T$ 的后向概率 $\beta_t(i)$，只需要考虑在时刻 t+1 所有可能的 N 个状态

q_j 的转移概率(a_{ij} 项)，以及在此状态下的观测 $o_{t+!}$ 的观测概率($b_{jo_{t+1}}$ 项)，然后考虑状态 q_j 之后的观测序列的后向概率 $\beta_{t+1}(j)$

前向后向概率的关系

根据定义，可得

$$P(i_t = q_i, O \mid \lambda) = \alpha_t(i)\beta_t(i)$$
$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i)$$

单个状态的概率

给定模型 λ 和观测 O，求在时刻 t 处于状态 q_i 的概率

$$\gamma_t(i) = P(i_t = q_t \mid O, \lambda)$$

根据前向后向概率的定义

$$P(i_t = q_i, O \mid \lambda) = \alpha_t(i)\beta_t(i)$$
$$\gamma_t(i) = P(i_t = q_t \mid O, \lambda) = \frac{P(i_t = q_i, O \mid \lambda)}{P(O \mid \lambda)}$$
$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O \mid \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}$$

两个状态的联合概率

给定模型 λ 和观测 O，求在时刻 t 处于状态 q_i 并且时刻 t+1 处于时刻 q_j 的概率

$$\xi_t(i,j) = P(i_t = q_i, i_{t+1} = q_j \mid O, \lambda)$$

根据前向后向概率的定义

$$
\begin{aligned}
\xi_t(i,j) &= P(i_t = q_i, i_{t+1} = q_j \mid O, \lambda) \\
&= \frac{P(i_t = q_i, i_{t+1} = q_j, O \mid \lambda)}{P(O \mid \lambda)} \\
&= \frac{P(i_t = q_i, i_{t+1} = q_j, O \mid \lambda)}{\sum_{i=1}^N \sum_{j=1}^N P(i_t = q_i, i_{t+1} = q_j, O \mid \lambda)}
\end{aligned}
$$

$$
P(i_t = q_i, i_{t+1} = q_j, O \mid \lambda) = \alpha_t(i) a_{ij} b_{j o_{t+1}} \beta_{t+1}(j)
$$

学习算法

- 若训练数据包括观测序列和状态序列，则 HMM 的学习是监督学习 -> MLE
- 若训练数据只有观测序列，则 HMM 的学习需要使用 EM 算法，是非监督学习

监督学习方法

假设已给定训练数据包含 S 个长度相同的观测序列和对应的状态序列 $\{(O_1, I_1), (O_2, I_2), \ldots, (O_s, I_s)\}$，那么，可以直接给出 HMM 的参数估计，直接用频率去估计概率

转移概率 a_{ij} 的估计

设样本中时刻 t 处于状态 i，时刻 t+1 转移到状态 j 的频数是 A_{ij}，则

$$
\hat{a}_{ij} = \frac{A_{ij}}{\sum_{j=1}^N A_{ij}}
$$

观测概率 b_{ik} 的估计

设样本中状态 i 并且观测为 k 的频数为 B_{ik}，则

$$
\hat{b}_{ij} = \frac{B_{ik}}{\sum_{j=1}^N B_{ik}}
$$

初始状态的估计

π_i 的估计为 S 个样本中初始状态为 q_i 的概率

非监督学习 Baum-Wellch 算法

所有观测数据写成 $O=(o_1,o_2,\dots,o_T)$，所有隐数据写成 $I=(i_1,i_2,\dots,i_T)$，完全数据是 $(O,I)=(o_1,o_2,\dots,o_T,i_1,i_2,\dots,i_T)$，完全数据的对数似然函数是 $lnP(O,I\mid\lambda)$

假设 $\bar{\lambda}$ 是 HMM 参数当前的估计值，λ 为待求的参数，$P(O,I\mid\bar{\lambda}$ 可以看作是对数似然函数在当前条件下取值的概率，用来求期望

$$Q(\bar{\lambda},\lambda)=\sum_I lnP(O,I\mid\lambda)P(O,I\mid\bar{\lambda})$$

EM 过程

根据

$$\begin{aligned}P(O,I\mid\lambda)&=P(O\mid I,\lambda)P(I\mid\lambda)\\&=\pi_{i_1}b_{i_1o_1}a_{i_1i_2}b_{i_2o_2}a_{i_2i_3}\dots a_{i_{T-1}i_T}b_{i_To_T}\end{aligned}$$

函数可写成（把π,a,b分开）

$$Q(\bar{\lambda},\lambda)=\sum_I lnP(O,I\mid\lambda)P(O,I\mid\bar{\lambda})$$

$$=\sum_I ln\pi_{i_1}P(O,I\mid\bar{\lambda})+\sum_I\left(\sum_{t=1}^{T-1}lna_{i_ti_{t+1}}\right)P(O,I\mid\bar{\lambda})+\sum_I\left(\sum_{t=1}^{T}lnb_{i_to_t}\right)P(O,I$$

极大化 Q，求得参数 A,B,π

由于这三个参数分别位于三个项中，可分别极大化

初始概率

$$\sum_I ln\pi_{i_1}P(O,I\mid\bar{\lambda})=\sum_{i=1}^{N}ln\pi_{i_1}P(O,i_1=i\mid\bar{\lambda})$$

注意到 π_i 满足加和为 1，利用拉格朗日乘子法，得到

$$\sum_{i=1}^{N} ln\pi_{i_1} P(O, i_1 = i \mid \bar{\lambda}) + \gamma \left(\sum_{i=1}^{N} \pi_i - 1 \right)$$

对上式相对于 π_i 求偏导，得到

$$P(O, i_1 = i \mid \bar{\lambda}) + \gamma \pi_i = 0;$$

对 i 求和，得到

$$\gamma = -P(O \mid \bar{\lambda})$$

从而得到初始状态概率

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

转移概率和观测概率

第二项可写成

$$\sum_{I} \left(\sum_{t=1}^{T-1} ln a_{i_t i_{t+1}} \right) P(O, I \mid \bar{\lambda}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T-1} ln a_{ij} P(O, i_t = i, i_{t+1} = j \mid \bar{\lambda})$$

仍然使用拉格朗日乘子法，得到

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

同理，对于观测概率，得到

$$b_{ik} = \frac{\sum_{t=1}^{T} P(O, i_t = i \mid \bar{\lambda}) I(o_t = v_k)}{\sum_{t=1}^{T} P(O, i_t = i \mid \bar{\lambda})} \frac{\sum_{t=1, o_t=v_k}^{T} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}$$

预测算法

预测的近似算法

在每个时刻 t 选择在该时刻最有可能出现的状态 i_t^*，从而得到一个隐状态序列 $I^* = \{i_1^*, i_2^*, \ldots, i_T^*\}$，将它作为预测的结果

给定模型和观测序列，时刻 t 处于状态 q_i 的概率为

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O \mid \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}$$

选择概率最大的 i 作为最有可能的状态（会出现此状态在实际中可能不会发生的情况）

Viterbi 算法

用动态规划解 HMM 预测问题，用 DP 求概率最大的路径，一条路径对应一个状态序列。

定义变量 $\delta_t(i)$：在时刻 t 状态为 i 的所有路径中，概率的最大值

定义

$$\delta_t(i) = max_{i_1,i_2,\ldots,i_{t-1}} P(i_t = i, i_{t-1}, \ldots, i_1, o_t, \ldots, o_1 \mid \lambda)$$

递推

$$\begin{gathered}\delta_1(i) = \pi_i b_{io_1} \\ \delta_{t+1}(i) = max_{i_1,i_2,\ldots,i_t} P(i_{t+1} = i, i_t, \ldots, i_1, o_{t+1}, \ldots, o_1 \mid \lambda) \\ = max_{1 \le i \le N} \left(\delta_t(j) a_{ji}\right) b_{io_{t+1}}\end{gathered}$$

终止

$$P^* = max_{1 \le i \le N} \delta_T(i)$$

例子

还是用刚才的小球的例子，观测向量 O=红白红，试求最优状态序列

$$\pi = \begin{pmatrix} 0.2 \\ 0.4 \\ 0.4 \end{pmatrix} \quad A = \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{pmatrix} \quad B = \begin{pmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{pmatrix}$$

初始化，在 t=1 时，对于每一个状态 i，求状态为 i 观测到 $o_1 = red$ (矩阵 B 的第一列) 的概率，记为 $\delta_1(i)$

$$\delta_1(i) = \pi_i b_{io_1} = \pi_i b_{i,red}$$

求得

$$\delta_1(1) = 0.2 \times 0.5 = 0.1$$
$$\delta_1(2) = 0.4 \times 0.4 = 0.16$$
$$\delta_1(3) = 0.4 \times 0.8 = 0.28$$

在 t=2 时，对每个状态 i，求在 t=1 时状态为 j 观测为红并且在 t=2 时状态为 i 观测为白的路径的最大概率，记为 $\delta_2(i)$，这里 $o_2 = white$ 则：

$$\delta_{t+1}(i) = max_{1\leq j\leq 3}\left(\delta_1(j)a_{ji}\right) b_{io_2} = max_{1\leq j\leq 3}\left(\delta_1(j)a_{j1}\right) b_{i,white}$$

求得

$$\delta_2(1) = max_{1\leq j\leq 3}\left(\delta_1(j)a_{j1}\right) b_{i,white}$$
$$= max\{0.10 \times 0.5, 0.16 \times 0.3, 0.28 \times 0.2\} \times 0.5 = 0.028$$

同理可得

$$\delta_2(2) = 0.0504$$
$$\delta_2(3) = 0.042$$

继续可以求得

$$\delta_3(1) = 0.00756$$
$$\delta_3(2) = 0.01008$$
$$\delta_3(3) = 0.0147$$

所以最大是 $\delta_3(3) = 0.0147$，根据每一步最大，得到的序列是(3,3,3)