统计中的计算方法
第二次作业

于慧倩
14300180118

2017 年 4 月

1. 已知隐马尔科夫链的状态转移概率如图一所示, 对输出状态 A,B 的输出概率为 e1(A) = 0.5,e1(B) = 0.5,e2 (A) = 0.1,e2 (B) = 0.9,e3 (A) = 0.9, e3 (B) = 0.1。完成 1-5 题:

   (a) 用向前方法计算出现状态 BAB 的概率。

   $$P(BAB) = 0.1 \times f_1(3) + 0.2 \times f_2(3) + 0.4 \times f_3(3)$$

   用 forward 方法求解 $f_1(3), f_2(3), f_3(3)$:

   $f_1(1) = 0.2 \times 0.5 = 0.1$

   $f_2(1) = 0.3 \times 0.9 = 0.27$

   $f_3(1) = 0.5 \times 0.1 = 0.05$

   $f_1(2) = 0.5 \times (0.3 \times 0.1 + 0) = 0.015$

   $f_2(2) = 0.1 \times (0.3 \times 0.1 + 0.4 \times 0.27 + 0.3 \times 0.05) = 0.0153$

   $f_3(2) = 0.9 \times (0.3 \times 0.1 + 0.4 \times 0.27 + 0.3 \times 0.05) = 0.1377$

   $f_1(3) = 0.5 \times (0.3 \times 0.015) = 0.00225$

   $f_2(3) = 0.9 \times (0.3 \times 0.015 + 0.4 \times 0.0153 + 0.3 \times 0.1377) = 0.046737$

   $f_3(3) = 0.1 \times (0.3 \times 0.015 + 0.4 \times 0.0153 + 0.3 \times 0.1377) = 0.005193$

   故

   $$P(BAB) = 0.1 \times f_1(3) + 0.2 \times f_2(3) + 0.4 \times f_3(3) = 0.0116496$$

   (b) 用向后方法计算出现状态 BAB 的概率。

   $$P(BAB) = 0.2 \times 0.5 \times b_1(1) + 0.3 \times 0.9 \times b_2(1) + 0.5 \times 0.1 \times b_3(1)$$

   用 backward 方法求解 $b_1(1), b_2(1), b_3(1)$:

   $b_1(3) = 0.1, b_2(3) = 0.2, b_3(3) = 0.4$

   $b_1(2) = 0.3 \times 0.5 \times 0.1 + 0.3 \times 0.9 \times 0.2 + 0.3 \times 0.1 \times 0.4 = 0.081$

   $b_2(2) = 0.4 \times 0.9 \times 0.2 + 0.4 \times 0.1 \times 0.4 = 0.088$

   $b_3(2) = 0.3 \times 0.9 \times 0.2 + 0.3 \times 0.1 \times 0.4 = 0.066$

   $b_1(1) = 0.3 \times 0.5 \times 0.081 + 0.3 \times 0.1 \times 0.088 + 0.3 \times 0.9 \times 0.066 = 0.03261$

   $b_2(1) = 0.4 \times 0.1 \times 0.088 + 0.4 \times 0.9 \times 0.066 = 0.02728$

   $b_3(1) = 0.3 \times 0.1 \times 0.088 + 0.3 \times 0.9 \times 0.066 = 0.02046$

故

$P(BAB) = 0.2\times0.5\times b_1(1)+0.3\times0.9\times b_2(1)+0.5\times0.1\times b_3(1) = 0.0116496$

(c) 对 BAB 计算对每个显示状态，隐状态 G_2 的概率。

$$\begin{aligned} P(s_1 = G_2|x_1 = B) &= \frac{P(s_1 = G_2, x_1 = B)}{P(x_1 = B)} \\ &= \frac{P(x_1 = B|s_1 = G_2)P(s_1 = G_2)}{\sum_{l=1,2,3} P(x_1 = B|s_1 = G_l)P(s_1 = G_l)} \\ &= \frac{0.9 \times 0.3}{0.5 \times 0.2 + 0.9 \times 0.3 + 0.1 \times 0.5} \\ &= 0.6428571 \end{aligned}$$

$$\begin{aligned} P(s_2 = G_2|x_2 = A) &= \frac{P(s_2 = G_2, x_2 = A)}{P(x_2 = A)} \\ &= \frac{P(x_2 = A|s_2 = G_2)P(s_2 = G_2)}{\sum_{l=1,2,3} P(x_2 = A|s_2 = G_l)P(s_2 = G_l)} \\ &= 0.09166667 \end{aligned}$$

$$\begin{aligned} P(s_3 = G_2|x_3 = B) &= \frac{P(s_3 = G_2, x_3 = B)}{P(x_3 = B)} \\ &= \frac{P(x_3 = B|s_3 = G_2)P(s_3 = G_2)}{\sum_{l=1,2,3} P(x_3 = B|s_3 = G_l)P(s_3 = G_l)} \\ &= 0.8686047 \end{aligned}$$

(d) 对 BAB 计算隐状态为 $G_1G_2G_1$ 的概率。

由于从 G_2 转移到 G_1 的概率为 0，所以隐状态为 $G_1G_2G_1$ 的概率为 0.

(e) 对 BAB 计算最优的隐状态路径。

$$v_1(1) = 0.5 \times 0.2 = 0.1, v_2(1) = 0.9 \times 0.3 = 0.27, v_3(1) = 0.1 \times 0.5 = 0.05$$

$$v_1(2) = 0.5 \times \max\{0.1 \times 0.3, 0, 0\} = 0.015$$

$$v_2(2) = 0.1 \times \max\{0.3 \times 0.1, 0.4 \times 0.27, 0.3 \times 0.05\} = 0.108$$

$$v_3(2) = 0.9 \times \max\{0.3 \times 0.1, 0.4 \times 0.27, 0.3 \times 0.05\} = 0.972$$

$$v_1(3) = 0.5 \times \max\{0.3 \times 0.015, 0, 0\} = 0.00225$$

$$v_2(3) = 0.9 \times \max\{0.3 \times 0.015, 0.4 \times 0.108, 0.3 \times 0.972\} = 0.26244$$

$$v_3(3) = 0.1 \times \max\{0.2 \times 0.015, 0.4 \times 0.108, 0.3 \times 0.972\} = 0.02916$$

由此得到，最优隐状态路径为 $\mathrm{G}_2, \mathrm{G}_3, \mathrm{G}_2$，这条隐状态概率为 0.26244

2. 假设 HMM 隐状态为 A,B, 显示状态为 L,R, 对附件数据 assign2.csv, 估计 HMM 的参数, 并估计出现此隐状态的概率。数据中包含 2 条链的隐状态和显示状态。

   解:

   (a) 推导过程

   使用最大似然估计对隐状态已知的马尔可夫链进行参数估计:

   设转移矩阵为 A，发射矩阵为 E，其中 a_{ij} 代表从 j 转移到 i 状态的概率，e_{ij} 代表从 j 发射出 i 状态的概率。设 AA $\tilde{AA}$ 分别为第一条链、第二条链的转移频数矩阵，EE $\tilde{EE}$ 分别为第一条链、第二条链的发射频数矩阵,（其中 $AA_{i,j}$ 代表第一条链中隐状态由 j 变为 i 的频数，$EE_{i,j}$ 代表第一条链中隐状态 j 发射出显状态 i 的频数）。

   由于数据中包含两条链的隐状态和显状态，有利用最大似然估计:

$$
\begin{aligned}
Q(\theta|\theta^{(t)}) &= arg\max L(\theta|\boldsymbol{X}_1, \boldsymbol{S}_1, \boldsymbol{X}_2, \boldsymbol{S}_2) \\
&= arg\max \log P(\theta|\boldsymbol{X}_1, \boldsymbol{S}_1, \boldsymbol{X}_2, \boldsymbol{S}_2) \\
&= arg\max \{\log P(\theta|\boldsymbol{X}_1, \boldsymbol{S}_1) + \log P(\theta|\boldsymbol{X}_2, \boldsymbol{S}_2)\} \\
&= arg\max \log \prod_{k=1,2;l=1,2} (a_{k,l})^{AA_{k,l}} (a_{k,l})^{\tilde{AA}_{k,l}} (e_{k,l})^{EE_{k,l}} (e_{k,l})^{\tilde{EE}_{k,l}} \\
&= arg\max \{ \sum_{k=1,2;l=1,2} (AA_{k,l} + \tilde{AA}_{k,l}) \log a_{k,l} \\
&\quad + \sum_{k=1,2;l=1,2} (EE_{k,l} + \tilde{EE}_{k,l}) \log e_{k,l} \}
\end{aligned}
$$

   令其对参数求导等于零，得到:

$$
a_{k,l} = \frac{AA_{k,l} + \tilde{AA}_{k,l}}{AA_{k,l} + AA_{k',l} + \tilde{AA}_{k,l} + \tilde{AA}_{k',l}}
$$

$$
e_{k,l} = \frac{EE_{k,l} + \tilde{EE}_{k,l}}{AA_{k,l} + AA_{k',l} + \tilde{AA}_{k,l} + \tilde{AA}_{k',l}}
$$

$$
(k = 1, 2; l = 1, 2)
$$

(b) 编写程序 H2_2.R 计算，最终得到转移矩阵与发射矩阵：

i. 转移矩阵

$$\begin{pmatrix} 0.854546 & 0.193182 \\ 0.145455 & 0.806818 \end{pmatrix}$$

ii. 发射矩阵

$$\begin{pmatrix} 0.6636364 & 0.2727273 \\ 0.3363636 & 0.7272727 \end{pmatrix}$$

(c) 求该隐状态出现的概率, 即 $P(s_i = l|\boldsymbol{X})$ 出现的概率：

$$\begin{aligned} P(s_i = l|\boldsymbol{X}) &= \frac{P(\boldsymbol{X}, s_i = l)}{P(\boldsymbol{X})} \\ &= \frac{P(\boldsymbol{X}, s_i = l)}{\sum_k P(\boldsymbol{X}, s_i = l_k)} \\ &= \frac{f_{s_i}(l) b_{s_i}(l)}{\sum_k f_{s_i}(l_k) b_{s_i}(l_k)} \end{aligned}$$

由程序 H2_2.R 得到该隐状态出现的概率如 P_x_s.

3. 同 (二), 对附件数据 assign3.csv, 估计 HMM 的参数, 并估计最优隐状态路径。数据中包含 2 条链的隐状态。

(a) 估计参数

i. 用向前方法计算:

$$f_l(i) = e_l(x_i) \sum_k f_k(i-1) \cdot a_{kl}; (i = 2 : L)$$

ii. 用向后方法计算:

$$B(s_i) = \sum_{s_{i+1}} P(s_{i+1}|s_i) P(x_{i+1}|s_{i+1}) B(s_{i+1}); (i = 1 : L-2)$$

$$B(s_{L-1}) = P(x_L|s_{L-1}) = \sum_{s_L} P(s_L|s_{L-1}) P(x_L|x_L)$$

iii. 有两条观察到的显链，所以利用 EM 算法得到参数估计值：

A. E-step:

$$\begin{aligned} & Q(\theta|\theta^t) \\ = & \sum_j \sum_{s^j \in S^j} P(s^j|X^j, t\theta^t) \log P(x^j, X^j|\theta) \\ = & \sum_j \sum_{s^j \in S^j} P(s^j|X^j, t\theta^t) \log(\pi^j_{s_1} \prod_{t=1}^{T} a_{s_{t-1}s_t} e_{s_t}(x_t)) \end{aligned}$$

B. M-step:

初始发射概率:

$$\pi_{1,i} = \frac{1}{2} \sum_{j=1}^{2} P(S^j_1 = i|X^j, \theta^t)$$

转移数量矩阵:

$$\begin{aligned} A_{kl} &= \sum_{j=1}^{2} \frac{1}{P(X^j)} \sum_{i=1}^{L} P(X^j, s_{i-1} = k, s_i = l) \\ &= \sum_{j=1}^{2} \frac{1}{P(X^j)} \sum_{i=1}^{L} f^j_k(i-1) a^j_{kl} e^j_l(x_i) b^j_l(i) \end{aligned}$$

发射数量矩阵:

$$E_k(b) = \sum_{j=1}^{2} \frac{1}{P(X^j)} \sum_{i, x_i = b} f^j_k(i) b^j_k(i)$$

从而得到转移矩阵与发射矩阵的迭代公式:

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}}, e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}$$

(b) 迭代结果

如程序 H2_3.R 最终得到转移矩阵与发射矩阵如下:

i. 转移矩阵

$$\begin{pmatrix} 0.8875018 & 0.04368625 \\ 0.1124982 & 0.9563138 \end{pmatrix}$$

ii. 发射矩阵

$$\begin{pmatrix} 0.8375897 & 0.3603209 \\ 0.1624103 & 0.6396791 \end{pmatrix}$$

(c) 估计最优隐状态路径

估计出两条链的最优隐状态如下:

i. L_1= B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B A A A A A A A A A A A A A A A A A A A B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B.

ii. L_2=B B B B B B B B B B B B B B B B B B B B B B B B B B B B B A A A A A A A A A A A A A A A A A A A A A A A A A A B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B.