Notas do Curso Inferência em Processos Estocásticos

Prof. Antonio Galves
Transcrita por Karina Yuriko Yaginuma

1 Estimação de máxima verossimilhança para cadeias de Markov de ordem k

Seja $(X_n)_{n=0,1,2,\ldots}$ uma cadeia de Markov de ordem k assumindo valores em A finito, com proabilidades de transição dadas por

$$p = \{p(a|u); a \in A, u \in A^k\},$$

onde $u = u_1 u_2 \ldots u_k$.

A função de verossimilhança é dada por

$$\begin{aligned}
\mathbb{P}_p\{X_{-k}^n = a_{-k}^n\} &= \mathbb{P}_p\{X_{-k}^{-1} = a_{-k}^{-1}\} \prod_{t=0}^{n} \mathbb{P}_p\{X_t = a_t | X_{-k}^{t-1} = a_{-k}^{t-1}\} \\
&= \mathbb{P}_p\{X_{-k}^{-1} = a_{-k}^{-1}\} \prod_{t=0}^{n} \mathbb{P}_p\{X_t = a_t | X_{t-k}^{t-1} = a_{t-k}^{t-1}\} \\
&= \mathbb{P}_p\{X_{-k}^{-1} = a_{-k}^{-1}\} \prod_{a\in A} \prod_{u\in A^k} p(a|u)^{N_n(ua)}. \qquad (1)
\end{aligned}$$

onde $N_n(ua)$ é o número de vezes que observamos a sequência u seguida pelo símbolo a na amostra

$$N_n(ua) = \sum_{t=k+1}^{n} \mathbb{I}\{X_{t-k}^{t-1} = u\}.$$

Vamos assumir que $\mathbb{P}_p\{X_{-k}^{-1} = a_{-k}^{-1}\} = 1$. Queremos encontrar $\hat{p}_n = \{\hat{p}(a|u); a \in A, u \in A^k\}$ que maximiza a verossimilhança $\mathbb{P}_p\{X_{-k}^n = a_{-k}^n\}$.

$$\hat{p}_n = \arg\max\{q \in \mathcal{M}_k(A) : \mathbb{P}_q\{X_{-k}^n = a_{-k}^n\}\}, \qquad (2)$$

onde $\mathcal{M}_k(A)$ é a classe das cadeias de Markov de ordem k assumindo valores em A.

Denote por $L_p(a_{-k}^n)$ a função log da verossimilhança, definida por

$$L_p(a_{-k}^n) = \sum_{a\in A} \sum_{u\in A^k} N_n(ua) \log p(a|u). \qquad (3)$$

Queremos maximizar $L_p(a_{-k}^n)$ com a restrição que $\sum_{a\in A} p(a|u) = 1$, para todo $u \in A^k$. Para isso, vamos utilizar o método dos multiplicadores de Lagrange. Sejam $\underline{\lambda} = (\lambda_u)_{u\in A^k}$, $\lambda_u \in \mathbb{R}$ e $F(\underline{\lambda}, p)$ definido por

$$F(\underline{\lambda}, p) = \sum_{u\in A^k} \left\{ \sum_{a\in A} N_n(ua) \log p(a|u) + \lambda_u \Big[1 - \sum_{a\in A} p(a|u)\Big] \right\}. \qquad (4)$$

Derivando $F(\underline{\lambda}, p)$ em relação à λ_u e igualando zero, temos que

$$\begin{aligned}
\frac{\partial}{\partial \lambda_u} F(\underline{\lambda}, p) &= 1 - \sum_{a\in A} p(a|u) \\
1 - \sum_{a\in A} p(a|u) &= 0. \qquad (5)
\end{aligned}$$

Derivando $F(\underline{\lambda}, p)$ em relação à $p(a|u)$ e igualando zero, temos que

$$\begin{aligned}
\frac{\partial}{\partial p(a|u)} F(\underline{\lambda}, p) &= N_n(ua) \frac{1}{p(a|u)} - \lambda_u \\
\hat{p}_n(a|u) &= \frac{N_n(ua)}{\lambda_u} \qquad (6)
\end{aligned}$$

Pelas equações (5) e (6), temos que

$$\sum_{b\in A}\frac{N_n(ub)}{\lambda_u}=1\Rightarrow\lambda_u=\sum_{b\in A}N_n(ub). \tag{7}$$

Logo, o estimador de máxima verossimilhança $\hat{p}=\{\hat{p}(a|u);a\in A,u\in A^k\}$ é dado por

$$\hat{p}_n(a|u)=\frac{N_n(ua)}{\sum_{b\in A}N_n(ub)}. \tag{8}$$

Observação: Segue da Lei dos Grandes Números que

$$\hat{p}_n(a|u)=\frac{\frac{N_n(ua)}{n}}{\frac{N_{n-1}(u)}{n}}\overset{n\to\infty}{\longrightarrow}\frac{\mu(u)p(a|u)}{\mu(u)}, \tag{9}$$

onde $\mathbb{N}_{n-1}(u)=\sum_{b\in A}N_n(ub)$.

Teorema 1. *O estimador de máxima verossimilhança (EMV) é consistente,*

$$\hat{p}_n\xrightarrow{q.c.}p,\qquad quando\qquad n\longrightarrow\infty. \tag{10}$$

Da mesma maneira, podemos calcular os estimadores de máxima verossimilhança para as cadeias estocásticas com memória de alcance variável assumindo valores em A com $p_\tau=\{p_\tau(\cdot|w);w\in\tau\}$, onde τ é a árvore de contexto. O estimador é dado por

$$\hat{p}_\tau(a|w)=\frac{N_n(wa)}{\sum_{b\in A}N_{n-1}(wb)}. \tag{11}$$

Exercício 1: Seja $\tau=\{00,10,1\}$ uma árvore de contexto. Calcule a máxima verossimilhança da amostra

$$X_{-2}=0,X_{-1}=0,X_0=1,X_1=0,X_2=0,X_3=0,X_4=1,X_5=1,X_6=0,X_8=0.$$

Exercício 2: Seja $\tau\subset A_{-\infty}^{-1}\bigcup\{\cup_{k\geq1}A^{\{-k,\dots,-1\}}\}$. Vamos supor que τ é completo (i.e., para todo $x_{-\infty}^{-1}\in A_{-\infty}^{-1}$, existe y_{-k}^{-1}, $k\in\mathbb{N}\cup\{-\infty\}$, tal que $y_{-k}^{-1}=x_{-k}^{-1}$).
Suponha também que τ tem a propriedade de sufixo, portanto

1. se $x_{-n}^{-1}\in\tau$, então $x_{-m}^{-1}\notin\tau$ para todo $m\in\{1,2,...,n-1\}$,
2. se $x_{-\infty}^{-1}\in\tau$, então $x_{-m}^{-1}\notin\tau$ para todo $m\geq1$.

Seja $l:A_{-\infty}^{-1}\longrightarrow\mathbb{N}^*\cup\{+\infty\}$,

$$l(x_{-\infty}^{-1})=k\qquad\Longleftrightarrow\qquad x_{-k}^{-1}\in\tau. \tag{12}$$

Sejam $\Omega=A^{\mathbb{Z}}$, $\mathcal{F}=\sigma$-álgebra produto, $X_t:\Omega\longrightarrow A$ projeção da t-ésima coordenada e $\mathcal{F}_{-n}^{-1}=\sigma(X_{-n}^{-1})$ a σ-álgebra gerada por X_{-n}^{-1}. Mostre que l é tempo de parada se, e somente se, τ tem propriedade de sufixo.
Observação: Uma variável aleatória l é um tempo de parada se $\forall k\geq1$ o evento $\{l(X_{-\infty}^{-1})=k\}\in\mathcal{F}_{-k}^{-1}$.

2 Seleção da ordem de uma cadeia de Markov

Dada uma amostra $X_0,X_1,...,X_n$. Sabemos que ela foi gerada por uma cadeia de Markov de alcance fixo finito. Porém, não sabemos qual é esse alcance $\bar{k}$. Mas sabemos que $\bar{k}<<n$, isto é, $n\geq|A|^{\bar{k}}$.
Vamos supor que a probabilidade da primeira sequência X_0^{k-1} ser escolhida seja 1, para $1\leq k\leq\log_{|A|}n$. Então conseguimos calcular

$$\hat{\mathbb{P}}_{MV}^{(k)}\{X_0^n\}=\prod_{a_{-k}^{-1}\in A^k}\prod_{b\in A}\hat{p}^{(k)}(b|a_{-k}^{-1})^{N_n(a_{-k}^{-1}b)}. \tag{13}$$

Como encontrar $\bar{k}$?

Suponha que conhecemos $\bar{k}$ (hipótese nula), então

$$p^{(\bar{k}+1)}(b|a_{-\bar{k}+1}^{-1}) \quad = \quad p^{(\bar{k})}(b|a_{-\bar{k}}^{-1}). \tag{14}$$

Portanto, para $b \in A$ e $a_{-\bar{k}}^{-1} \in A^{\bar{k}}$, temos que

$$\mathbb{P}_{\bar{k}}(X_n = b|X_{n-\bar{k}}^{n-1} = a_{-\bar{k}}^{-1}) = \mathbb{P}_{\bar{k}}(X_n = b|X_{n-\bar{k}}^{n-1} = a_{-\bar{k}}^{-1}, X_{n-(\bar{k}+1)} = c), \text{ para todo } c \in A. \tag{15}$$

Como visto anteriormente, os estimadores de máxima verossimilhança são dados por

$$\hat{p}^{(\bar{k})}(b|a_{-\bar{k}}^{-1}) = \frac{N_n(a_{-\bar{k}}^{-1}b)}{\sum_{c\in A} N_{n-1}(a_{-\bar{k}}^{-1}c)} \qquad \text{e} \qquad \hat{p}^{(\bar{k}+1)}(b|a_{-(\bar{k}+1)}^{-1}) = \frac{N_n(a_{-(\bar{k}+1)}^{-1}b)}{\sum_{c\in A} N_{n-1}(a_{-(\bar{k}+1)}^{-1}c)}.$$

Queremos obter um critério estatístico para saber se $\hat{p}^{(\bar{k})}$ e $\hat{p}^{(\bar{k}+1)}$ são suficientemente próximos para suportar a hipótese nula. Podemos pensar no mesmo problema para as cadeias estocásticas com memória de alcance variável, uma maneira de selecionar a árvore probabilística de contexto é dada pelo seguinte algoritmo:

1. Catalogar todas as sequências de comprimento $d(n) = \log_{|A|} n$ que aparecem na amostra;
2. Defina $\delta \in (0, 1)$ (pequeno);
3. Para cada sequência $a_{-d(n)}^{-1}$ catalogada:
   (a) Defina $k = d(n)$
   (b) Calcule $M(a_{-k}^{-1}) = \max_{b\in A} |\hat{p}_n(b|a_{-k}^{-1}) - \hat{p}_n(b|a_{-(k-1)}^{-1})|$:
      i. Se $M(a_{-k}^{-1}) < \delta$, atualize o valor de k por $k-1$ e volte para 3b,
      ii. caso contrário ou se $k = 1$, pare e volte para 3;
4. Repita esse procedimento para todas as sequências catalogadas.

Vamos calcular a seguinte probabilidade

$$\mathbb{P}\left\{\max_{b\in A} |\hat{p}_n(b|a_{-k}^{-1}) - \hat{p}_n(b|a_{-(k-1)}^{-1})| > \delta\right\}. \tag{16}$$

Denote por $w = a_{-k}^{-1}$ e $w' = a_{-(k-1)}^{-1}$. Seja

$$\xi_i = I\{X_{T_i-(k-1)}^{T_i} = w, X_{T_i+1} = b\},$$

onde T_i = instantes da conclusão da i-ésima visita ao contexto w,

$$\begin{aligned} T_1 &= \inf\{t \geq k : X_{t-(k-1)}^{t} = w\}, \\ T_i &= \inf\{t \geq T_{i-1} : X_{t-(k-1)}^{t} = w\}, \text{ para } n \geq 2. \end{aligned}$$

Se $N_n(w) = N$, então $\hat{p}_n(b|w) = \sum_{i=1}^{N} \xi_i$ e $\hat{p}_n(b|w') = \sum_{i=1}^{N} \xi_i'$ (ξ_i' equivalente para w'). Suponha que $H_0 : p(b|w) = p(b|w')$ (hipótese nula), temos que

$$\mathbb{P}\{\max_{b\in A} |\hat{p}_n(b|w) - \hat{p}_n(b|w')| > \delta\} \leq \sum_{b\in A} \mathbb{P}\{|\hat{p}_n(b|w) - \hat{p}_n(b|w')| > \delta\}. \tag{17}$$

Notação: $p = p(b|w)$, $p' = p(b|w')$, $\hat{p} = \hat{p}(b|w)$ e $\hat{p}' = \hat{p}(b|w')$. Pela hipótese nula, temos que $p = p'$, então

$$\begin{aligned} \mathbb{P}\{\max_{b\in A} |\hat{p}_n(b|w) - \hat{p}_n(b|w')| > \delta\} &\leq \sum_{b\in A} \mathbb{P}\{|\hat{p}_n(b|w) - \hat{p}_n(b|w')| > \delta\} \\ &= \sum_{b\in A} \mathbb{P}\{|\hat{p} - p + p' - \hat{p}'| > \delta\} \\ &\leq \sum_{b\in A} \mathbb{P}\{|\hat{p} - p| + |p' - \hat{p}'| > \delta\}. \end{aligned} \tag{18}$$

Lema 1. *Seja Z uma variável aleatória tal que $0 < Z \leq Z_1 + Z_2$, onde Z_1 e Z_2 são duas variáveis aleatórias. Então,*

$$\mathbb{P}\{Z > \delta\} \leq \mathbb{P}\{Z_1 + Z_2 > \delta\} \leq \mathbb{P}\left\{Z_1 > \frac{\delta}{2}\right\} + \mathbb{P}\left\{Z_2 > \frac{\delta}{2}\right\}. \tag{19}$$

Pelo Lema 1, temos que

$$\begin{aligned} \mathbb{P}\{\max_{b\in A} |\hat{p}_n(b|w) - \hat{p}_n(b|w')| > \delta\} &\leq \mathbb{P}\{|\hat{p} - p| + |p' - \hat{p}'| > \delta\} \\ &\leq \mathbb{P}\left\{|\hat{p} - p| > \frac{\delta}{2}\right\} + \mathbb{P}\left\{|\hat{p}' - p'| > \frac{\delta}{2}\right\}. \end{aligned} \tag{20}$$

Precisamos apenas analisar um termo de (20), já que os dois termos são iguais. Denote por $\epsilon = \frac{\delta}{2}$,

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}\xi_i - p\right| > \epsilon\right\} = \mathbb{P}\left\{\left|\sum_{i=1}^{N}\xi_i - Np\right| > N\epsilon\right\} \tag{21}$$

Sob H_0 temos que $\mathbb{P}\{\xi_i = 1\} = p$, queremos majorar (21) de tal forma que $\searrow 0$ quando $n \to \infty$.
Dificuldade extra: N é aleatória.
Vamos considerar primeiro o caso em que N **não** é aleatório, $N = n$ fixado

$$\mathbb{P}\{|\sum_{i=1}^{n}\xi_i - np| > n\epsilon\} \leq \eta(n) \searrow 0. \tag{22}$$

Usando a desigualdade de Chebyshev

$$\begin{aligned} \mathbb{P}\left\{\left|\sum_{i=1}^{n}\xi_i - np\right| > n\epsilon\right\} &= \mathbb{P}\left\{(\sum_{i=1}^{n}\xi_i - np)^2 > (n\epsilon)^2\right\} \\ &\leq \frac{Var(\sum_{i=1}^{n}\xi_i)}{n^2\epsilon^2}. \end{aligned} \tag{23}$$

Sob hipótese de que $\xi_1, \xi_2, ..., \xi_n$ são indepedentes, temos que

$$\begin{aligned} \mathbb{P}\left\{\left|\sum_{i=1}^{n}\xi_i - np\right| > n\epsilon\right\} &\leq \frac{nVar(\xi_1)}{n^2\epsilon^2} \\ &= \frac{p(1-p)}{n\epsilon^2} \\ &\leq \frac{1/4}{n\delta^2}. \end{aligned} \tag{24}$$

Podemos utilizar também, para encontrar um limitante superior para (21), a desigualdade dos Grandes Desvios. Para isso, defina $S = \sum_{i=1}^{n}\xi_i$, $\mu = np$ e $t = n\delta$.
O evento $\{|S - \mu| > t\}$ é equivalente à união disjunta dos eventos $\{S > \mu + t\} \cup \{S < \mu - t\}$, logo

$$\begin{aligned} \mathbb{P}\{|\sum_{i=1}^{n}\xi_i - np| > n\delta\} &= \mathbb{P}\left\{\{S > \mu + t\} \cup \{S < \mu - t\}\right\} \\ &= \mathbb{P}\{S > \mu + t\} + \mathbb{P}\{S < \mu - t\}. \end{aligned} \tag{25}$$

Queremos calcular a $\mathbb{P}\{S > \mu + t\} = \mathbb{P}\{e^{\lambda S} > e^{\lambda(\mu+t)}\}$, onde $\lambda > 0$. Pela desigualdade de Markov, temos que

$$\mathbb{P}\{e^{\lambda S} > e^{\lambda(\mu+t)}\} \leq e^{-\lambda(\mu+t)}\mathbb{E}(e^{\lambda S}). \tag{26}$$

Utilizando a estrutura de $S = \sum_{i=1}^{n}\xi_i$ e suponha que $\xi_i's$ são iid, com $\mathbb{P}\{\xi_i = 1\} = p$ e $\mathbb{P}\{\xi_i = 0\} = 1 - p$, então

$$\begin{aligned} \mathbb{E}\left(e^{\lambda S}\right) &= \mathbb{E}\left(e^{\lambda\sum_{i=1}^{n}\xi_i}\right) = \mathbb{E}\left(e^{\lambda\xi_1}\right)^n \\ &= \left[e^{\lambda}p + (1-p)\right]^n. \end{aligned} \tag{27}$$

Portanto,

$$\begin{aligned}\mathbb{P}\{e^{\lambda S} > e^{\lambda(\mu+t)}\} &\leq e^{-\lambda(\mu+t)}[e^{\lambda}p + (1-p)]^n \\ &= \exp\left\{-n\left[\lambda(p+\delta) - \ln\{e^{\lambda}p + (1-p)\}\right]\right\}. \end{aligned} \tag{28}$$

Defina $\Phi(\lambda) = \lambda(p+\delta) - \ln\{e^{\lambda}p + (1-p)\}$, logo

$$\mathbb{P}\{e^{\lambda S} > e^{\lambda(\mu+t)}\} \leq e^{-n\Phi(\lambda)}. \tag{29}$$

Queremos que $\Phi(\lambda) > 0$ e que seja o maior possível. Para isso, vamos analisar o comportamento de $\Phi(\lambda)$, $\lambda \geq 0$.

Note que quando $\lambda = 0$, temos que $\Phi(0) = 0$. A primeira e segunda derivada de $\Phi(\lambda)$ são dadas por

$$\Phi'(\lambda) = (p+\delta) - \frac{pe^{\lambda}}{pe^{\lambda} + (1-p)} \qquad \text{e} \qquad \Phi''(\lambda) = \frac{(pe^{\lambda})^2}{(pe^{\lambda} + (1-p))^2} - \frac{pe^{\lambda}}{pe^{\lambda} + (1-p)}.$$

Como $\Phi'' < 0$, então λ^* o ponto de inflexão da função Φ é o λ ótimo. Logo

$$\mathbb{P}\{e^{\lambda S} > e^{\lambda(t+\mu)}\} \leq e^{-n\Phi(\lambda^*)}, \tag{30}$$

onde $\Phi(\lambda^*)$ é ótimo.

Agora vamos analisar o caso em que $N = N_n(w)$ é aleatório,

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^{N}\xi_i \geq N(p+\delta)\right\} &= \mathbb{P}\left\{B, N > m\right\} + \mathbb{P}\left\{B, N < m\right\} \\ &\leq \mathbb{P}\left\{B, N > m\right\} + \mathbb{P}\left\{N < m\right\}. \end{aligned} \tag{31}$$

onde $B = \left\{\sum_{i=1}^{N}\xi_i \geq N(p+\delta)\right\}$ e $m = m(n)$ determinístico. Seja $\chi_t = I\{X_{t-(k-1)}^{t} = w\}$, então $N_n(w) = \sum_{t=k-1}^{n}\chi_t$.

$\chi_1, \chi_2, ...$ não são independentes pois a cadeia $(X_n)_{n=0,1,...}$ é de Markov e dependendo do valor de w, a ocorrência de w num certo instante pode influenciar a ocorrência de w no futuro. Por esse motivo precisamos usar propriedades de Martingais.

Vamos calcular a $\mathbb{E}[N_n(w)]$,

$$\begin{aligned}\mathbb{E}[N_n(w)] &= \sum_{t=k-1}^{n}\mathbb{E}(I\{X_{t-(k-1)}^{t} = w\}) \\ &= n\mathbb{P}\{X_{-k}^{-1} = w\} = np(w). \end{aligned} \tag{32}$$

A igualdade em (32) sugere utilizar $m = m(n) = n\alpha$, onde $\alpha = p(w) - \delta$ e mostrar que

$$\mathbb{P}\{N_n(w) < n(p(w)-\delta)\} \overset{n\to\infty}{\to} 0. \tag{33}$$

Feito isso, voltamos para o outro termo da soma em (31)

$$\begin{aligned}\mathbb{P}\{B, N > m\} &= \mathbb{P}\left\{\sum_{i=1}^{N}\xi_i > N(p+\delta), N \geq m\right\} \\ &\leq \mathbb{P}\left\{\sum_{i=1}^{N}\xi_i \geq m(p+\delta)\right\} \\ &\leq \mathbb{P}\left\{\sum_{i=1}^{n}\xi_i \geq m(p+\delta)\right\}, \text{pois } N_n \leq n. \end{aligned} \tag{34}$$

Como escolher $m(n) \leq n$ de maneira que (33) continue valendo? Se $m(n) = n(p-\delta)$,

$$\begin{aligned}\mathbb{P}\{\sum_{i=1}^{n}\xi_i \geq n(p+\delta)(p-\delta)\} &= \mathbb{P}\{\sum_{i=1}^{n}\xi_i \geq n(p^2-\delta^2)\} \\ &\leq e^{-n\Phi(\lambda^*)}, \end{aligned} \tag{35}$$

onde Φ é definido com $(p^2 - \delta^2)$. Como w é um contexto, temos que

$$\mathbb{P}\left\{X_t = b | X_{t-|w|}^{t-1} = w, X_0^{(t-|w|)-1}\right\} = \mathbb{P}\left\{X_t = b | X_{t-|w|}^{t-1} = w\right\}. \tag{36}$$

De fato, $\xi_1, \xi_2, ...$ são independentes. Pois, como $\xi_i = I\{X_{T_i+1} = b\}$ onde $T_i =$ instante do fim da i-ésima visita a w, a propriedade de Markov garante a independência dos ξ_i's.