MAE 0580/ MAC 6926 - Lista 3

4 de outubro de 2017

Notações e definições básicas

Neste resumo $(X_n)_{n\geq -k}$ é uma cadeia com memória de alcance variável gerada por (τ, p), sendo τ uma árvore de altura k. Denotamos por $c_\tau(X_{-k}^n)$ o contexto de τ associado à sequência $X_{-k}, \cdots, X_n$.

Dado $n \geq 1$ e dada uma amostra $X_{-k}, \cdots, X_n$, definimos a função de contagem

$$N_{0:n}(a_{-k}^0) = \sum_{t=0}^{n} \mathbf{1}_{\{X_{t-k}^t = a_{-k}^0\}}$$

para toda sequência $a_{-k}^0 \in A^{k+1}$.

O passado nas probabilidades de transição é indicado do símbolo mais recente ao símbolo mais remoto:

$$p(b|a_{-1}, ..., a_{-k}) = p(b|a_{-k}^{-1}) = \mathbb{P}\{X_0 = b|X_{-k}^{-1} = a_{-k}^{-1}\}, \text{ para } b \in A \text{ e } a_{-k}^{-1} = (a_{-k}, ..., a_{-1}) \in A^k.$$

Dada uma amostra $X_{-k}, \cdots, X_n$ de símbolos no alfabeto A com árvore de contextos τ, com altura de τ igual a k, definimos

$$\hat{\mathbb{P}}_\tau(X_0^n|X_{-k}^{-1}) = \prod_{\omega\in\tau}\prod_{a\in A} \hat{p}_n(a|\omega)^{N_{0:n}(\omega a)},$$

onde $\hat{p}_n(a|\omega) = \frac{N_{0:n}(\omega a)}{N_{0:n-1}(\omega)}$ é o estimador de máxima verossimilhança da matriz de probabilidades de transição dado τ.

Definimos o estimador

$$\tilde{p}_n(a|w) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{X_{t_k^w+1} = a\}}\,.$$

Desigualdade de Hoeffding para variáveis aleatórias binárias e i.i.d: Sejam $Y_1, Y_2, \ldots$ assumindo valores no alfabeto $\{0, 1\}$ com $\mathbb{P}(Y_n = 1) = p$. Então para todo $\delta > 0$,

$$\mathbb{P}(\sum_{i=1}^{n} Y_i > n(p + \delta)) \leq \exp\{-2n\delta^2\}$$

Algoritmo Contexto:

Dada uma amostra $X_{-k}^n = (X_{-k}, \ldots, X_n)$ de símbolos do alfabeto finito, definimos para toda sequência $w = w_{-k}^{-1} \in A^k$, com $1 < k < n$, a seguinte quantidade

$$\Delta_n(w_{-k}^{-1}) = \max_{a,b\in A}\left|\hat{p}_n(a|w_{-(k-1)}^{-1}) - \hat{p}_n(a|w_{-(k-1)}^{-1}b)\right|\,.$$

Fixado $\delta \in (0, 1)$, se $\Delta_n(w_{-k}^{-1}) < \delta$, então podamos os símbolos mais remotos (representados pela letra b) das sequências $\{bw_{-(k-1)}^{-1} : b \in A\}$. Caso contrário, mantemos as sequências $\{bw_{-(k-1)}^{-1} : b \in A\}$.

1. Seja $(\xi_n)_{n\geq 1}$ uma sequência de variáveis aleatórias independentes e identicamente distribuidas, assumindo valores no conjunto $A = \{0, 1\}$, com

$$\mathbb{P}(\xi_n = 1) = p\,.$$

Diga qual das afirmativas abaixo é verdadeira.

(a) $\mathbb{E}\left(2^{\sum_{m=1}^{n}\xi_m}\right) = (p + 1)^n$

(b) $\mathbb{E}\left(2^{\sum_{m=1}^{n}\xi_m}\right) = 2^{n(p+1)}$
(c) $\mathbb{E}\left(2^{\sum_{m=1}^{n}\xi_m}\right) = (1/2)^n$
(d) Nenhuma das respostas anteriores.

2. Seja $(X_n)_{n\geq 1}$ uma sequência de variáveis aleatórias iid, assumindo valores no alfabeto $A = \{0,1\}$, tal que $\mathbb{P}(X_n = 1) = p$, onde $0 < p < 1$. Usando a desigualdade de Hoeffding, diga qual das seguintes afirmações é verdadeira (use, se necessário, que $\log(0.01) = -4.6$, $\log(0.05) = -3$ e $\log(0.02) = -3.91$)

   (a) $\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p > 0.02\right) \leq 0.01, \forall n \geq 5.000$
   (b) $\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p > 0.02\right) \leq 0.05, \forall n \geq 4.000$
   (c) $\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p > 0.02\right) \geq 0.02, \forall n \geq 3.000$
   (d) Nenhuma das respostas anteriores.

3. Dada a amostra $(0,0,1,1,1,0,1,0,1,1,1,1,0,0,0,1)$, diga qual é a árvore de contextos $\hat{\tau}$ de altura menor ou igual a 2, obtida aplicando o Algoritmo Contexto, utilizando $\delta = 0.05$ no critério de poda.

   (a) $\hat{\tau} = \{1,0\}$
   (b) $\hat{\tau} = \{1,00,10\}$
   (c) $\hat{\tau} = \{0,11,01\}$
   (d) Nenhuma das respostas anteriores.

4. Seja X_0^n uma amostra gerada por uma cadeia de alcance variável assumindo valores no alfabeto $A = \{0,1\}$. Sob a hipótese nula

   H_0 : A sequência 00 é um contexto e $p(.\,|\,00)$ é a probabilidade de transição associada a esse contexto,

   e, fixado $\delta > 0$, assinale a alternativa correta:

   (a) $\mathbb{P}(|\hat{p}_n(1|00) - \hat{p}_n(1|000)| > \delta) \leq \mathbb{P}(|\hat{p}_n(1|00) - p(1|00)| > \frac{\delta}{2}) + \mathbb{P}(|\hat{p}_n(1|000) - p(1|000)| > \frac{\delta}{2})$
   (b) $\mathbb{P}(|\hat{p}_n(1\,|\,00) - \hat{p}_n(1\,|\,000)| > \delta) \leq \mathbb{P}(|\hat{p}_n(1\,|\,00) - p(1\,|\,00)| > \frac{\delta}{2}) + \mathbb{P}(|\hat{p}_n(1\,|\,0) - p(1\,|\,0)| > \frac{\delta}{2})$
   (c) $\mathbb{P}(|\hat{p}_n(1\,|\,00) - \hat{p}_n(1\,|\,000)| > \delta) > \mathbb{P}(|\hat{p}_n(1\,|\,0) - p(1\,|\,0)| > \frac{\delta}{2})$
   (d) Nenhuma das respostas anteriores.

5. Dada uma amostra $X_{-k}, \cdots, X_n$ de símbolos do alfabeto $A = \{0,1\}$, diga qual das seguintes afirmações é verdadeira:

   (a) $N_{0:n}(0^k0) + N_{0:n}(0^k1) = N_{0:n-1}(0^k)$
   (b) $N_{0:n}(0^k0) = N_{0:n}(0^k1)$
   (c) $N_{0:n}(0^k0) + N_{0:n}(0^k1) = N_{0:n+1}(0^k)$
   (d) Nenhuma das respostas anteriores.

6. Seja $(X_n)_{n\geq 0}$ uma cadeia com memória de alcance variável assumindo valores no alfabeto $A = \{0,1\}$. Seja $X_{-2}^{11} = 00111000101111$ uma amostra gerada por essa cadeia. Aplicando o Algoritmo Contexto, obtemos $\hat{\tau} = \{1,00,10\}$. Para obter tal resultado, o valor de δ utilizado foi:

(a) $\delta = 0.05$
(b) $\delta = 0.03$
(c) $\delta = 0.01$
(d) Nenhuma das respostas anteriores.

7. Seja $(X_n)_{n\geq -1}$ uma cadeia com memória de alcance variável assumindo valores em $A = \{0, 1\}$ tendo como árvore de contextos

$$\tau = \{\{X_{-1} = 1\}, \{X_{-2} = 1, X_{-1} = 0\}, \{X_{-2} = 0, X_{-1} = 0\}\}$$

e família de probabilidades de transição p desconhecida. Dada a amostra

$$X_{-1}^{14} = (1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1)$$

gerada por esta cadeia, os estimadores de máxima verossimilhança para $p(1|1), p(1|00)$ e $p(1|01)$ são, respectivamente:

(a) $\hat{p}_n(1|1) = 1/2$, $\hat{p}_n(1|00) = 2/5$ e $\hat{p}_n(1|01) = 1/4$.
(b) $\hat{p}_n(1|1) = 1/3$, $\hat{p}_n(1|00) = 3/5$ e $\hat{p}_n(1|01) = 1/4$.
(c) $\hat{p}_n(1|1) = 1/3$, $\hat{p}_n(1|00) = 3/5$ e $\hat{p}_n(1|01) = 3/4$.
(d) Nenhuma das respostas anteriores.

8. Seja τ a árvore de contextos indexada pelos símbolos do alfabeto $A = \{0, 1\}$ assim definida

$$\tau = \{\{\omega_{-1} = 0\}, \{\omega_{-2} = 0, \omega_{-1} = 1\}, \{\omega_{-2} = 1, \omega_{-1} = 1\}\}.$$

Associamos a τ a família de probabilidade de transição p assim definida

$$p(0 \,|\, 0) = 0.2, p(0 \,|\, 10) = 0.4, p(0 \,|\, 11) = 0.6.$$

Simulamos uma amostra X_{-1}^{5} da cadeia estocástica assumindo valores no alfabeto $A = \{0, 1\}$ usando o seguinte algoritmo:

Passo 1. $X_{-1} = 0$
Passo 2. Para $n \geq 0$, definimos $X_n = \mathbb{1}_{\{U_n \geq p(0|c_\tau(X_{-1}^{n-1}))\}}$ em que $U_0, U_1, \cdots$ é uma sequência de v.a. i.i.d. com distribuição uniforme no intervalo $[0, 1]$.
Supondo que o sorteio das variáveis uniformes produziu a sequência

$$U_0 = 0.77, U_1 = 0.92, U_2 = 0.12, U_3 = 0.44, U_4 = 0.59, U_5 = 0.14.$$

diga qual das opções abaixo é correta

(a) $X_0 = 0, X_1 = 0, X_2 = 1, X_3 = 0, X_4 = 0, X_5 = 1$
(b) $X_0 = 1, X_1 = 0, X_2 = 0, X_3 = 1, X_4 = 0, X_5 = 1$
(c) $X_0 = 1, X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 0$
(d) Nenhuma das respostas anteriores.