MAE 0580/MAC 6926 - 1^a Prova (Turma A)

13 de setembro 2017

Esta é uma prova individual, sem consulta. Em cada questão uma e só uma opção é correta. A nota da prova será calculada pela fórmula Nota $= \max\{0, C - E/3\}$. Nesta expressão, C é o número de respostas certas e E, o número de respostas erradas. Questões deixadas em branco e respostas rasuradas *não serão consideradas* no cálculo da nota.

Notações e definições básicas

Seja $(X, Y) \in \mathcal{X} \times \{0, 1\}$ um par de variáveis aleatórias distribuidas de acordo com uma distribuição desconhecida P. Observamos uma sequencia $(X_i, Y_i)_{i=1,\dots,n}$ de pares i.i.d. tendo a mesma distribuição de (X, Y). O objetivo é construir uma função de classificação $g : \mathcal{X} \to \{0, 1\}$, tal que $g(X)$ esteja probabilísticamente próximo de Y.

O *risco* de g é definido como,

$$R(g) = \mathbb{P}(g(X) \neq Y).$$

O *risco empírico* de g calculado a partir de uma amostra é definido como,

$$R_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{g(X_i) \neq Y_i\}}.$$

Seja $\eta(x) = \mathbb{P}(Y = 1|X = x)$. O classificador de Bayes é definido como

$$f^*(x) = \mathbb{1}_{\{\eta(x) \geq 1/2\}}.$$

Desigualdade de Chebyshev: Seja $(Z_k)_{k\geq 1}$ uma sequência de variáveis aleatórias i.i.d assumindo valores em $\{0, 1\}$, tal que $\mathbb{P}(Z_k = 1) = p$, para todo k. A desigualdade de Chebyshev é dada por

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} Z_k - p\right| > \epsilon\right) \leq \frac{1}{4n\epsilon^2}$$

Desigualdade de Hoeffding: Seja $(Z_k)_{k\geq 1}$ uma sequência de variáveis aleatórias i.i.d assumindo valores em $\{0, 1\}$, tal que $\mathbb{P}(Z_k = 1) = p$, para todo k. A desigualdade de Hoeffding é dada por

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} Z_k - p\right| > \epsilon\right) \leq 2\exp(-2n\epsilon^2)$$

Dado um alfabeto finito A, para todo $k \geq 1$, definimos $\mathcal{M}_k(A)$ da seguinte maneira

$$\mathcal{M}_k(A) = \{p : A^k \times A \to [0, 1] : \forall a_{-k}^{-1} \in A^k, \sum_{b\in A} p(b|a_{-k}^{-1}) = 1\}.$$

Dado $n \geq 0, k \geq 0$ e dada uma amostra $X_{-k}, \cdots, X_n$ e uma sequência $a_{-k}^0 \in A^{k+1}$, definimos a função de contagem

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

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, gerada por uma cadeia de Markov de alcance k, definimos

$$\hat{\mathbb{P}}_k(X_0^n|X_{-k}^{-1}) = \prod_{a_{-1}^{-k}\in A^k} \prod_{b\in A} \hat{p}_n(b|a_{-1}^{-k})^{N_{0:n}(a_{-1}^{-k}b)},$$

onde $\hat{p}_n(b|a_{-1}^{-k}) = \frac{N_{0:n}(a_{-1}^{-k}b)}{\sum_{z\in A} N_{0:n}(a_{-1}^{-k}z)}$ é o estimador de máxima verossimilhança em $\mathcal{M}_k(A)$.

1. Seja $X = (X(1), X(2), X(3))$ uma variável aleatória que toma valores em $\mathcal{X} = \{0,1\}^3$, isto é $\mathcal{X}$ é o conjunto de todas as sequências ordenadas com três elementos, assumindo os valores 0 ou 1. Seja $Y \in \{0,1\}$ uma variável de classificação assim definida: $Y = X(1)$. Definimos a função de classificação $g : \mathcal{X} \to \{0,1\}$ da seguinda maneira
$$g(x) = \mathbb{1}_{\left\{\frac{x(1)}{2}+\frac{x(2)}{4}+\frac{x(3)}{8}\geq 0.75\right\}}.$$
Assumindo que $\mathbb{P}(X = x) = \left(\frac{1}{2}\right)^3$, para todo $x \in \mathcal{X}$, diga qual das afirmações abaixo é verdadeira:
   a) $R(g) = 0.5$
   b) $R(g) = 0.25$
   c) $R(g) = 0.75$
   d) Nenhuma das respostas anteriores.
2. Seja $X \in \mathcal{X} = [0,1]$ uma variável aleatória e seja $Y \in \{0,1\}$ uma variável de classificação assim definida
$$Y = Z\mathbb{1}_{\{X\geq 1/2\}} + (1-Z)\mathbb{1}_{\{X<1/2\}},$$
onde Z é uma variável aleatória com valores em $\{0,1\}$, independente de X e tal que $\mathbb{P}(Z = 1) = 0.9$. Queremos calcular o classificador de Bayes para o par (X,Y). Diga qual das afirmações abaixo é verdadeira:
   a) $f^*(x) = \mathbb{1}_{\{x\geq 0.9\}}$
   b) $f^*(x) = \mathbb{1}_{\{x\geq 0.5\}}$
   c) $f^*(x) = \mathbb{1}_{\{x\geq 0.1\}}$
   d) Nenhuma das respostas anteriores.
3. Seja $(X_1,Y_1),...,(X_{10},Y_{10})$, com $(X_i,Y_i) \in [0,1]\times\{0,1\}$ uma sequência de variáveis aleatórias. Seja $g : [0,1] \to \{0,1\}$ uma funcão definida por $g(x) = \mathbb{1}_{\left\{|x-\frac{1}{2}|\leq 0.2\right\}}$. Dado uma amostra
$$\begin{aligned}
X_1 &= 0.4, \quad Y_1 = 0\\
X_2 &= 0.1, \quad Y_2 = 0\\
X_3 &= 0.7, \quad Y_3 = 1\\
X_4 &= 0.3, \quad Y_4 = 1\\
X_5 &= 0.2, \quad Y_5 = 0\\
X_6 &= 0.8, \quad Y_6 = 0\\
X_7 &= 0.3, \quad Y_7 = 1\\
X_8 &= 0.9, \quad Y_8 = 0\\
X_9 &= 0.2, \quad Y_9 = 1\\
X_{10} &= 0.5, \quad Y_{10} = 1,
\end{aligned}$$
diga qual das seguintes afirmações é verdadeira
   a) $R_n(g) = 2/10$
   b) $R_n(g) = 1/10$
   c) $R_n(g) = 5/10$
   d) Nenhuma das respostas anteriores.
4. Seja $(X_i,Y_i)_{i=1,...,n}$ uma sequência de variáveis aleatórias i.i.d com $X_i \in \mathcal{X}$ e $Y_i \in \{0,1\}$, e seja $g(x) = \mathbb{1}_{\{x\geq p\}}$, com $p \in [0,1]$ fixado. Usando a desigualdade de Hoeffding queremos obter uma majoração para
$$\mathbb{P}(|R_n(g) - R(g)| > 0.01).$$
Diga qual das seguintes afirmações é verdadeira (use, se necessário, que $\ln(0.005) = -5.2983$ e $\ln(0.025) = -3.6889$)
   a) $\mathbb{P}\left(|R_n(g) - R(g)| > 0.01\right) \leq 0.01, \quad \forall n \geq 30000$
   b) $\mathbb{P}\left(|R_n(g) - R(g)| > 0.01\right) \leq 0.05, \quad \forall n \geq 15000$
   c) $\mathbb{P}\left(|R_n(g) - R(g)| > 0.01\right) \geq 0.01, \quad \forall n \geq 30000$

d) Nenhuma das respostas anteriores.

5. Seja $(X_i, Y_i)_{i=1,\dots,n}$ uma sequência de variáveis aleatórias i.i.d com $X_i \in \mathcal{X}$ e $Y_i \in \{0,1\}$, e seja $g(x) = \mathbb{1}_{\{x \geq p\}}$, com $p \in [0,1]$ fixado. Fixamos $\epsilon \in [0, 1/2]$ e queremos obter uma majoração para $\mathbb{P}(|R_n(g) - R(g)| > \epsilon)$. Sejam $\delta_c(n,\epsilon)$ e $\delta_h(n,\epsilon)$ as majoracões fornecidas pelas desigualdades de Chebyshev e Hoeffding respectivamente. Diga qual das seguintes afirmações é verdadeira

   a) $\lim\limits_{n\to\infty} \frac{\delta_c(n,\epsilon)}{\delta_h(n,\epsilon)} = +\infty$

   b) $\lim\limits_{n\to\infty} \frac{\delta_c(n,\epsilon)}{\delta_h(n,\epsilon)} = 0$

   c) $\lim\limits_{n\to\infty} \frac{\delta_c(n,\epsilon)}{\delta_h(n,\epsilon)} = 1$

   d) Nenhuma das respostas anteriores.

6. Seja $(X_n)_{n\in\mathbb{Z}}$ uma cadeia de Markov de alcance 1 assumindo valores em $A = \{0,1\}$. Diga qual é o maior valor que $\mathbb{P}(X_0^{10} = (0,1,1,0,1,0,0,0,1,0,0) | X_{-1} = 0)$ pode assumir

   (a) $(1/2)^{11}$

   (b) $(1/2)^5(3/4)^3(1/4)^3$

   (c) $(4/7)^4(3/7)^3(3/4)^3(1/4)$

   (d) Nenhuma das respostas anteriores.

7. Seja $(X_n)_{n\in\mathbb{Z}}$ uma cadeia com memória de alcance variável, assumindo valores no alfabeto $A = \{0,1\}$, tendo como árvore de contextos $\tau = \{\{X_{-1} = 0\}, \{X_{-2} = 0, X_{-1} = 1\}, \{X_{-2} = 1, X_{-1} = 1\}\}$ e tendo família associada de probabilidades de transição definida por

$$p(1|0) = \alpha, \quad p(1|11) = \beta, \quad p(1|10) = \gamma,$$

   onde α, β, γ são três parâmetros pertenecentes ao intervalo aberto $(0,1)$. Dada uma amostra $X_0^{12} = (0,1,1,1,0,0,0,1,0,1,1,1,1)$ gerada por esta cadeia, diga qual das seguintes afirmações é correta:

   a) $\mathbb{P}(X_0^{12} = 0,1,1,1,0,0,0,1,0,1,1,1,1 | X_{-1} = 0) = \alpha^3(1-\alpha)^3\beta^3(1-\beta)\gamma^2(1-\gamma)$

   b) $\mathbb{P}(X_0^{12} = 0,1,1,1,0,0,0,1,0,1,1,1,1 | X_{-1} = 0) = \alpha^3(1-\alpha)^3\beta(1-\beta)^3\gamma(1-\gamma)^2$

   c) $\mathbb{P}(X_0^{12} = 0,1,1,1,0,0,0,1,0,1,1,1,1 | X_{-1} = 0) = (\alpha\beta\gamma)^{13}$

   d) Nenhuma das respostas anteriores.

8. Dado $A = \{1,2,3\}$, seja $p \in \mathcal{M}_1(A)$ assim definida:

$$\begin{array}{c c} & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \begin{pmatrix} 1/4 & 1/2 & 1/4 \\ 1/3 & 1/3 & 1/3 \\ 1/4 & 1/4 & 1/2 \end{pmatrix} \end{array}.$$

   Queremos simular uma cadeia de Markov $(X_n)_{n\geq 0}$, tendo essa matriz de probabilidades de transição, usando o seguinte algoritmo:
   Passo 1. Escolho X_0;
   Passo 2. Para $n \geq 1$, definimos $X_n = F(X_{n-1}, U_n)$, onde $(U_n)_{n\geq 1}$ é uma sequência de variáveis aleatórias i.i.d com distribuição uniforme em $[0,1]$ e $F(x,u)$ é uma função definida por:

$$F(x,u) = \begin{cases} 1, & \text{se } 0 \leq u < h_1(x) \\ 2, & \text{se } h_1(x) \leq u < h_2(x) \\ 3, & \text{se } h_2(x) \leq u \leq 1\,. \end{cases}$$

   Diga qual das linhas abaixo, definindo $h_1(3)$ e $h_2(3)$, está correta:

   (a) $h_1(3) = 1/3$, $h_2(3) = 2/3$

   (b) $h_1(3) = 1/4$, $h_2(3) = 3/4$

   (c) $h_1(3) = 1/4$, $h_2(3) = 1/2$

   (d) Nenhuma das respostas anteriores.