# Terceira Prova
# Fundamentos Estatísticos para Ciência dos Dados

11/07/2017

1. **10 pontos:** O vento é uma força vetorial que está sempre mudando de direção e de intensidade (magnitude). Por causa, disso, sua medição num instante de tempo qualquer é tratada como uma variável aleatória. O modelo usualmente adotado para a intensidade $X$ do vento (uma v.a. contínua) é a densidade

$$f(x) = \begin{cases} \frac{x}{\phi} e^{-x^2/(2\phi)}, & \text{se } x > 0 \\ 0, & \text{se } x \leq 0 \end{cases}$$

Estamos considerando apenas a magnitude escalar da intensidade do vento, ignorando a sua direção.

- **1 ponto:** Encontre o MLE de $\phi$ se uma amostra i.i.d. de medições $X_1, X_2, \ldots, X_n$ do vento é feita.
- **2 pontos:** Encontre uma estatística suficiente para estimar $\phi$.
- **1 ponto:** O MLE é função da estatística suficiente?
- **1 ponto:** Sabe-se que $\mathbb{E}(X) = \sqrt{\phi\pi/2}$ e $\mathbb{V}(X) = \phi(4-\pi)/2$. Obtenha então $\mathbb{E}(X^2)$ usando a expressão da variância em termos de $\mathbb{E}(X)$ e de $\mathbb{E}(X^2)$.
- **2 pontos:** Obtenha a informação de Fisher $\mathbb{I}(\phi)$ através da derivada segunda da log-verossimilhança.
- **1 ponto:** Diga qual é a distribuição aproximada do MLE quando o tamanho da amostra é grande.
- **2 pontos:** Use a distribuição assintótica do MLE para construir um intervalo de confiança de 95% para o parâmetro $\theta$. OBS: se $Z \sim N(0,1)$ então $\mathbb{P}(|Z| \leq 2) \approx 0.95$

**Solução:** A densidade conjunta dos dados é

$$f(\mathbf{x}phi) = \frac{\prod_i x_i}{\phi^n} \exp\left(-\frac{1}{2\phi}\sum_i x_i^2\right)$$

e portanto com log-verossimilhança

$$\ell(\phi) = \sum_i log(x_i) - nlog(\phi) - \frac{1}{2\phi}\sum_i x_i^2 \,.$$

Derivando com respeito a $\phi$ temos

$$\frac{\partial \ell}{\partial \phi} = -\frac{n}{\phi} + \frac{1}{2\phi^2}\sum_i x_i^2.$$

Igualando a zero, obtemos o MLE

$$\hat{\phi} = \frac{1}{2n}\sum_i x_i^2 \,.$$

A densidade conjunta pode ser escrita como $f(\mathbf{x}$
$phi) == k(\mathbf{x})g(\boldsymbol{\phi})h(T(\mathbf{x}), \phi)$ onde $T(\mathbf{x}) = \sum_i x_i^2$. Pelo teorema da fatoração de Neyman-Fisher, $T(\mathbf{X}) = \sum_i X_i^2$ é uma estatística suficiente para estimar $\phi$. Claramente, o MLE é função da estatística suficiente.

Como $\mathbb{E}(X^2) = \mathbb{V}(X) + (\mathbb{E}[X])^2$, então $\mathbb{E}(X^2) = \phi(4-\pi)/2 + \phi\pi/2 = 2\phi$.

Quanto à informação de Fisher,

$$\mathbb{I}(\phi) = -\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \phi^2}\right) = -n/\phi^2 + n/\phi^3 \mathbb{E}(X^2) = n/\phi^2 \, .$$

O resultado fundamental da teoria de MLE afirma que $\hat{\phi} \approx N(\phi, \phi^2/n)$.

O I.C. com aproximadamente 95% de confiança é $(\hat{\phi} \pm 2 * \hat{\phi}/\sqrt{n})$.

---

2. **8 pontos:** Uma amostra de tamanho $n$ vem de uma mistura de duas gaussianas, $N(\mu_a, 1)$ e $N(\mu_b, 1)$. Com probabilidade $\alpha$ o dado vem da distribuição do tipo $a$. Obtenha as expressões para as iterações do algoritmo EM.

   - **1 ponto:** Defina o parâmetro desconhecido $\boldsymbol{\theta}$, os dados observados $\mathbf{y}$ e os dados latentes (ocultos ou faltantes) $\mathbf{z}$.
   - **1 ponto:** Obtenha a log-verossmilhança $\log L^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z})$ dos dados completos.
   - **2 pontos:** Obtenha $\gamma_i = \mathbb{E}(Z_i|\mathbf{y}, \boldsymbol{\theta}^{(0)}) = \mathbb{P}(Z_i = 1|\mathbf{y}, \boldsymbol{\theta}^{(0)})$ onde $\boldsymbol{\theta}^{(0)}$ é o valor corrente do parâmetro.
   - **2 pontos:** Obtenha $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \mathbf{y}) = \mathbb{E}_{\boldsymbol{\theta}^{(0)}}(\log L^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}))$.
   - **2 pontos:** Derive $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \mathbf{y})$ com relação a $\boldsymbol{\theta}$ para encontrar as expressões iterativas do algoritmo EM.

   **Solução:** $\boldsymbol{\theta} = (\mu_a, \mu_b, \alpha)$, os dados observados $\mathbf{y} = (y_1, \ldots, y_n)$ e os dados não-observados $\mathbf{z} = (z_1, \ldots, z_n)$, variáveis binárias indicando se a $i$-ésima observação $y_i$ pertence à população $a$ ou não. Isto é, $z_i = 1$ se $i \in \{\text{pop}a\}$ e $z_i = 0$, caso contrário.

   A log-verossmilhança dos dados completos é dada por

$$\ell^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) = \sum_{i=1}^{n} [Z_i \log f_a(y_i) + (1 - Z_i) \log f_b(y_i)]$$

   Então

$$Q(\theta|\boldsymbol{\theta}_0, \mathbf{y}) = \sum_{i=1}^{n} [\gamma_i \log f_a(y_i) + (1 - \gamma_i) \log f_b(y_i)]$$

   onde

$$\gamma_i = \frac{\alpha^{(0)} f_a(y_i|\mu_a^{(0)})}{\alpha^{(0)} f_a(y_i|\mu_a^{(0)}) + (1 - \alpha^{(0)}) f_b(y_i|\mu_b^{(0)})}$$

   e $f_a(y|\mu_a)$ é a densidade de uma $N(\mu_a, 1)$avaliada no ponto $y$. Derivando $Q(\theta|\boldsymbol{\theta}_0, \mathbf{y})$,obtemos as estimativas do passo M:

$$\hat{\alpha} = \frac{1}{n} \sum_i \gamma_i^{(0)} \, ,$$

$$\hat{\mu}_a = \frac{\sum_i \gamma_i^{(0)} y_i}{\sum_i \gamma_i^{(0)}}$$

   e

$$\hat{\mu}_a = \frac{\sum_i (1 - \gamma_i^{(0)}) y_i}{\sum_i (1 - \gamma_i^{(0)})}$$

   com uma fórmula análoga para $\mu_b$.

---

3. **2 pontos:** Sejam $X_1, X_2, \ldots, X_n$ variáveis aleatórias i.i.d com distribuição $U(0, \theta)$. com $\theta > 0$ sendo um parâmetro desconhecido. Seja $\overline{X}$ a média aritmética das v.a.'s.

- Explique porquê $\overline{X}$ é um estimador viciado de $\theta$.
- Ache um estimador não-viciado de $\theta$.

**Solução:** O estimador $\overline{X}$ é viciado para estimar $\theta$ pois

$$\mathbb{E}(\overline{X}) = \mathbb{E}\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n}\sum_i \mathbb{E}\left(X_i\right) = \frac{1}{n}\sum_i \mathbb{E}\left(X_i\right) = \frac{1}{n} n \frac{\theta}{2} = \frac{\theta}{2}\,.$$

Um estimador não-viciado seria então $\hat{\theta} = 2\overline{X}$