

Aprendizado de Máquina e Reconhecimento de Padrões (UTFPR/CPGEI) - Lista de Exercícios 2

Tópicos: Modelos Estatísticos e Classificadores baseados em Teoria de Decisão Bayesiana

Parte 1 – Exercícios Teóricos:

- 1) Demonstrar, para distribuição Normal unidimensional, a estimativa através do método MLE, para a variância e para a média, considerando:
 - a. Média conhecida e variância desconhecida;
 - b. Média desconhecida e variância conhecida.

Explique o motivo da estimativa para a média ser *unbiased* e para variância ser *biased*, ou seja:

$$E[\hat{\sigma}_{ML}^2] = \frac{N-1}{N} \sigma^2$$

- 2) Considerando que a média seja desconhecida e que possui

$$p(\boldsymbol{\mu}) = \frac{1}{(2\pi)^{l/2} \sigma_{\mu}^l} \exp\left(-\frac{1}{2} \frac{\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2}{\sigma_{\mu}^2}\right)$$

Mostre que o MAP resulta em:

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln\left(\prod_{k=1}^N p(\mathbf{x}_k | \boldsymbol{\mu}) p(\boldsymbol{\mu})\right) = \mathbf{0} \longrightarrow \hat{\boldsymbol{\mu}}_{MAP} = \frac{\boldsymbol{\mu}_0 + \frac{\sigma_{\mu}^2}{\sigma^2} \sum_{k=1}^N \mathbf{x}_k}{1 + \frac{\sigma_{\mu}^2}{\sigma^2} N}$$

Parte 2 – Exercícios Práticos:

- 1) Gere dois conjuntos de dados, \mathbf{X} (treinamento) e \mathbf{X}_1 (teste), com $N=1000$ exemplos constituídos por vetores tridimensionais que pertencem a três classes equiprováveis ω_1 , ω_2 , and ω_3 . Os parâmetros das classes são:

$$\begin{aligned} m_1 &= [0, 0, 0]^T \\ m_2 &= [1, 2, 2]^T \\ m_3 &= [3, 3, 4]^T \end{aligned} \quad S_1 = S_2 = S_3 = \begin{bmatrix} 0.8 & 0 & 0 \\ 0 & 0.8 & 0 \\ 0 & 0 & 0.8 \end{bmatrix} = \sigma^2 I$$

- a) Utilizando \mathbf{X} , calcule o estimador ML da média e das matrizes de covariância para as três classes. Como assume-se que as três matrizes de covariância devem ser as mesmas, calcule para as três classes e realize uma média entre as três como a estimativa final.
- b) Utilize o classificador com base em distância Euclidiana para classificar os padrões de \mathbf{X}_1 , utilizando os valores computados (MLE) no item anterior.

- c) Utilize o classificador com base em distância de Mahalanobis para classificar os padrões de \mathbf{X}_1 , utilizando os valores computados (MLE) no item anterior.
- d) Utilize o classificador de Bayes para classificar os padrões de \mathbf{X}_1 , utilizando os valores computados (MLE) no item anterior.
- e) Para cada caso, calcule o erro de classificação e compare os resultados? Por que os desempenhos são próximos?

2) Repita o exercício anterior para:

$$\begin{aligned} m_1 &= [0, 0, 0]^T \\ m_2 &= [1, 2, 2]^T \\ m_3 &= [3, 3, 4]^T \end{aligned} \quad S_1 = S_2 = S_3 = \begin{bmatrix} 0.8 & 0.2 & 0.1 \\ 0.2 & 0.8 & 0.2 \\ 0.1 & 0.2 & 0.8 \end{bmatrix} \neq \sigma^2 I$$

- a) Repita agora para $P_1=1/2$ e $P_2=P_3=1/4$. Por que o desempenho do classificador Bayesiano é superior nesse caso?
- b) Repita agora para $P_1=P_2=P_3=1/3$ e

$$S_1 = \begin{bmatrix} 0.8 & 0.2 & 0.1 \\ 0.2 & 0.8 & 0.2 \\ 0.1 & 0.2 & 0.8 \end{bmatrix}, S_2 = \begin{bmatrix} 0.6 & 0.01 & 0.01 \\ 0.01 & 0.8 & 0.01 \\ 0.01 & 0.01 & 0.6 \end{bmatrix}, S_3 = \begin{bmatrix} 0.6 & 0.1 & 0.1 \\ 0.1 & 0.6 & 0.1 \\ 0.1 & 0.1 & 0.6 \end{bmatrix}$$

3) Gere um conjunto \mathbf{X} com $N=500$, seguindo:

$$\begin{aligned} p(x) &= \sum_{j=1}^3 P_j p(x|j) & m_1 &= [1, 1]^T \\ & & m_2 &= [3, 3]^T \\ S_1 &= 0.1I, S_2 = 0.2I, S_3 = 0.3I & m_3 &= [2, 6]^T \\ P_1 &= 0.4, P_2 = 0.4, \text{ and } P_3 = 0.2. \end{aligned}$$

- a) Aplique o algoritmo EM para definir os parâmetros do conjunto de dados acima, assumindo que esses parâmetros são previamente desconhecidos. Utilize:

$$J = 3, m_{1,ini} = [0, 2]^T, m_{2,ini} = [5, 2]^T, m_{3,ini} = [5, 5]^T$$

$$S_{1,ini} = 0.15I, S_{2,ini} = 0.27I, S_{3,ini} = 0.4I \text{ and } P_{1,ini} = P_{2,ini} = P_{3,ini} = 1/3$$

- b) Altere os parâmetros iniciais e verifique o impacto no resultado final. Compare também, os valores obtidos com os valores verdadeiros das pdfs.

4) Gere um conjunto de dados \mathbf{x} em uma dimensão tal que $\mathbf{x}_i \in \mathbb{R}$, $i = 1, 2, \dots, N$, seguindo a seguinte pdf:

$$p(x) = \frac{1}{3} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{x^2}{2\sigma_1^2}\right) + \frac{2}{3} \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x-2)^2}{2\sigma_2^2}\right)$$

$$\sigma_1^2 = \sigma_2^2 = 0.2$$

- a) Utilize o método Janela de Parzen, com $h = 0.01$; $N = 1000$ e $h = 0.1$; $N = 10,000$ para estimar a pdf. Qual a relação entre h e N ? Qual a relação entre h e σ ?
- b) Considere o mesmo conjunto de dados com $N=1000$. Utilize o modelo k-vizinhos-mais-próximos com $k = 21$ e estime a pdf. Qual o impacto da variação de k ? Justifique.
- 5) Gere um conjunto de dados \mathbf{X}_1 com $N_1=50$ exemplos constituídos por vetores 5-D que pertencem a duas classes equiprováveis ω_1 e ω_2 . De modo similar, gere \mathbf{X}_2 (teste), com $N=10000$. Os parâmetros das classes são:

$$\mathbf{m}_1 = [0, 0, 0, 0, 0]^T \text{ and } \mathbf{m}_2 = [1, 1, 1, 1, 1]^T$$

$$S_1 = \begin{bmatrix} 0.8 & 0.2 & 0.1 & 0.05 & 0.01 \\ 0.2 & 0.7 & 0.1 & 0.03 & 0.02 \\ 0.1 & 0.1 & 0.8 & 0.02 & 0.01 \\ 0.05 & 0.03 & 0.02 & 0.9 & 0.01 \\ 0.01 & 0.02 & 0.01 & 0.01 & 0.8 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 0.9 & 0.1 & 0.05 & 0.02 & 0.01 \\ 0.1 & 0.8 & 0.1 & 0.02 & 0.02 \\ 0.05 & 0.1 & 0.7 & 0.02 & 0.01 \\ 0.02 & 0.02 & 0.02 & 0.6 & 0.02 \\ 0.01 & 0.02 & 0.01 & 0.02 & 0.7 \end{bmatrix}$$

- a) Para cada classe e para cada uma das cinco dimensões, utilize o MLE para determinar as médias e variâncias unidimensionais de forma independente:

$$m_{1j}, m_{2j}, j = 1, 2, \dots, 5$$

$$\sigma_{1j}^2, \sigma_{2j}^2, j = 1, 2, \dots, 5$$

- b) Utilizando os valores anteriores, aplique o Naive Bayes Classifier para classificar os padrões do conjunto de teste, sendo a distribuição por dimensão dada por:

$$p(x|\omega_i) = \prod_{j=1}^5 \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp\left(-\frac{(x(j) - m_{ij})^2}{2\sigma_{ij}^2}\right), i = 1, 2$$

- c) Utilizando um classificador Bayesiano (não o Naive!) com m_1, m_2, S_1 e S_2 determinados por MLE, repita a classificação para os dados de teste e compare com o desempenho de classificação anterior.