Aprendizagem

Instituto Superior Técnico

outubro de 2023

Homework 4 - Report

Joana Pimenta (103730), Rodrigo Laia (102674)

Pen and Paper

1. Fórmulas utilizadas:

$$\gamma_{ki} = p(c_k|\mathbf{x}_i) = \frac{p(c_k)p(\mathbf{x}_i|c_k)}{p(\mathbf{x}_i)} \tag{1}$$

$$p(\mathbf{x}_i) = p(c_1)p(\mathbf{x}_i|c_1) + p(c_2)p(\mathbf{x}_i|c_2) \tag{2}$$

NAO SEI SE POSSO PROPRIAMENTE METER ESTA EQ ASSIM ????????

$$p(\mathbf{x}_i|c_k) = \begin{cases} p_k \cdot \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) & \text{se } y_1 = 1 \\ (1-p_k) \cdot \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) & \text{se } y_1 = 0 \end{cases} \tag{3}$$

E-step:

Cálculo das probabilidades $p(\mathbf{x}_i)$

$$p(\mathbf{x}_1) = 0.05185$$

$$p(\mathbf{x}_2) = 0.02775$$

$$p(\mathbf{x}_3) = 0.04337$$

$$p(\mathbf{x}_4) = 0.05243$$

Cálculos dos γ_{ki}

$$\gamma_{k=1,i=1} = 0.19259$$

$$\gamma_{k=2,i=1} = 0.80741$$

$$\gamma_{k=1,i=2} = 0.63135$$

$$\gamma_{k=2,i=2} = 0.36865$$

$$\gamma_{k=1,i=3} = 0.55181$$

$$\gamma_{k=2,i=3} = 0.44819$$

$$\gamma_{k=1,i=4} = 0.16892$$

$$\gamma_{k=2,i=4} = 0.83108$$

M-step:
Cada observação $\mathbf{x}_i$ permite atualizar os parâmetros com peso γ_{ki}. Assim calculamos os novos parâmetros atualizados para cada cluster utilizando as seguintes fórmulas.

$$N_k = \sum_{i=1}^{4} \gamma_{ki} \tag{4}$$

$$\pi_k = \frac{N_k}{N} \tag{5}$$

EU INVENTEI ISTO, PODE NAO TAR CERTO!!!!!!!

$$P_k(y_1 = 1) = \frac{\sum_{i=1}^{4} \gamma_{ki} p(y_1 = 1\mathbf{x}_i)}{\sum_{i=1}^{4} \gamma_{ki}} \tag{6}$$

$$P_k(y_1 = 1) = \sum_{i=1}^{4} \gamma_{ki} \tag{7}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{4} \gamma_{ki} \mathbf{x}_i \tag{8}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{i=1}^{4} \gamma_{ki} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \tag{9}$$

Parâmetros atualizados:

$$\pi_1 = 0.38617$$

$$\pi_2 = 0.61383$$

$$P_{k=1}(y_1 = 1) = 0.23404$$

$$P_{k=2}(y_1 = 1) = 0.66732$$

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0.02651 \\ 0.50713 \end{bmatrix}$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 0.30914 \\ 0.21042 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.14137 & -0.10541 \\ -0.10541 & 0.09605 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.11698 & -0.11835 \\ -0.11835 & 0.13530 \end{bmatrix}$$

2. Considerando o critério de *máximo a posteriori* Para decidir a que cluster pertence a observação $\mathbf{x}_{new}$ precisamos de calcular o seu posterior. Para isso utilizamos a seguinte fórmula:

$$p(cluster = k|\mathbf{x}_{new}) = \frac{p(cluster = k)p(\mathbf{x}_{new}|cluster = k)}{p(\mathbf{x}_{new})} \tag{10}$$

Cálculos:

$$p(\mathbf{x}_{new}) = 0.09667$$

$$p(cluster = 1|\mathbf{x}_{new}) = 0.02531$$

$$p(cluster = 2|\mathbf{x}_{new}) = 0.97469$$

Como $p(cluster = 2|\mathbf{x}_{new}) > p(cluster = 1|\mathbf{x}_{new})$, então a observação $\mathbf{x}_{new}$ pertence ao cluster 2.

3. Neste exercício assumimos que o cluster atribuído a cada observação é é escolhido pelo critério de *maximum likelihood.* Assim, o cluster escolhido é dado por:

$$cluster = \arg\max_k p(\mathbf{x}_i|cluster = k) \tag{11}$$

SERA Q ISTO TA CERTO????

$$p(\mathbf{x}_i|cluster = k) = \frac{p(\mathbf{x}_i, cluster = k)}{p(cluster = k)} \tag{12}$$

Assim,

Observação	$p(\mathbf{x}_i \mid cluster = 1)$	$p(\mathbf{x}_i \mid cluster = 2)$	Cluster atribuído
$\mathbf{x_1}$	0.59941	1.29174	2
$\mathbf{x_2}$	1.00197	0.25471	1
$\mathbf{x_3}$	1.13789	1.43871	2
$\mathbf{x_4}$	0.05377	1.26294	2

Coeficiente de Silhueta:

$$s_i = 1 - \frac{a(\mathbf{x}_i)}{b(\mathbf{x}_i)} \tag{13}$$

em que $a(\mathbf{x}_i)$ é a distância média entre $\mathbf{x}_i$ e as outras observações no mesmo cluster e $b(\mathbf{x}_i)$ é a distância média entre $\mathbf{x}_i$ e as observações no outro cluster.

A silhueta de um cluster é dada pela média dos coeficientes de silhueta de todas as observações pertencentes a esse cluster.

A silhueta da solução é por sua vez dada pela média das silhuetas de todos os clusters.

Neste caso a distância considerada é a distância de Manhattan, logo:

$$d(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{n} |u_i - v_i| \tag{14}$$

<table>
<tr><th>cluster</th><th>x_i</th><th>$s(x_i)$</th><th>$s(c_k)$</th><th>$s(sol)$</th></tr>
<tr><td>1</td><td>x_2</td><td>1</td><td>1</td><td rowspan="4">0.61420</td></tr>
<tr><td rowspan="3">2</td><td>x_1</td><td>0.59259</td><td rowspan="3">0.22840</td></tr>
<tr><td>x_3</td><td>-0.5</td></tr>
<tr><td>x_4</td><td>0.59259</td></tr>
</table>

4.

Programming - Código Python e Resultados Obtidos

1. Código Utilizado:

2. Código Utilizado:

3. Código Utilizado:

4.