

Aprendizagem

Instituto Superior Técnico

outubro de 2023

Homework 4 - Report

Joana Pimenta (103730), Rodrigo Laia (102674)

Pen and Paper

1. Fórmulas utilizadas:

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

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

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

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.

$$N_k = \sum_{i=1}^4 \gamma_{ki} \quad (4)$$

$$\pi_k = \frac{N_k}{N} \quad (5)$$

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

Nota: A probabilidade $p(y_1 = 1|\mathbf{x}_i)$ é 1 se y_1 de \mathbf{x}_i for 1 e 0 caso contrário.

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i=1}^4 \gamma_{ki} \mathbf{x}_i \quad (7)$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{i=1}^4 \gamma_{ki} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \quad (8)$$

Parâmetros atualizados:

$$\pi_1 = 0.38617$$

$$\pi_2 = 0.61383$$

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

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

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

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

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

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.10829 & -0.08865 \\ -0.08865 & 0.10412 \end{bmatrix}$$

2. Para calcular os posteriors da observação \mathbf{x}_{new} utilizamos a seguinte fórmula:

$$p(cluster = k | \mathbf{x}_{new}) = \frac{p(cluster = k)p(\mathbf{x}_{new} | cluster = k)}{p(\mathbf{x}_{new})} \quad (9)$$

Em que $p(cluster = k)$ é dado por π_k e $p(\mathbf{x}_{new} | cluster = k)$ é dado pela fórmula (3).

Cálculos:

$$p(\mathbf{x}_{new}) = 0.03048$$

$$p(cluster = 1 | \mathbf{x}_{new}) = 0.08029$$

$$p(cluster = 2 | \mathbf{x}_{new}) = 0.91971$$

Assim conclui-se que a observação \mathbf{x}_{new} pertence ao cluster 2 com probabilidade 0.91971 e ao cluster 1 com probabilidade 0.08029.

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

$$cluster = \arg \max_k p(\mathbf{x}_i | cluster = k) \quad (10)$$

Em que $p(\mathbf{x}_i | cluster = k)$ é dado pela fórmula (3).

Assim,

Observação	$p(\mathbf{x}_i cluster = 1)$	$p(\mathbf{x}_i cluster = 2)$	Cluster atribuído
\mathbf{x}_1	0.23147	0.94954	2
\mathbf{x}_2	1.26633	0.08874	1
\mathbf{x}_3	1.43811	0.45417	1
\mathbf{x}_4	0.02077	0.72331	2

Coefficiente de Silhueta:

$$s_i = 1 - \frac{a(\mathbf{x}_i)}{b(\mathbf{x}_i)} \quad (11)$$

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:

$$d(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^n |u_i - v_i| \quad (12)$$

Assim, as silhuetas obtidas foram:

cluster	\mathbf{x}_i	$a(\mathbf{x}_i)$	$b(\mathbf{x}_i)$	$s(\mathbf{x}_i)$	$s(\text{cluster})$	$s(\text{sol})$
1	\mathbf{x}_2	0.9	2.7	0.(6)	0.58(3)	0.702(7)
	\mathbf{x}_3	0.9	1.7(9)	0.4(9)		
2	\mathbf{x}_1	0.3(9)	2.25	0.8(2)	0.8(2)	
	\mathbf{x}_4	0.3(9)	2.25	0.8(2)		

4. A purity é dada por:

$$\text{purity} = \frac{1}{N} \sum_{k=1}^K \max_j |c_k \cap t_j| = \frac{1}{N} \left(\max_j |c_1 \cap t_j| + \max_j |c_2 \cap t_j| \right) \quad (13)$$

Uma vez que temos uma purity de 0.75 e um número total de observações de 4, então $\frac{1}{N} (\max_j |c_1 \cap t_j| + \max_j |c_2 \cap t_j|) = 0.75 \times 4 = 3$

Logo podemos ter os seguintes casos:

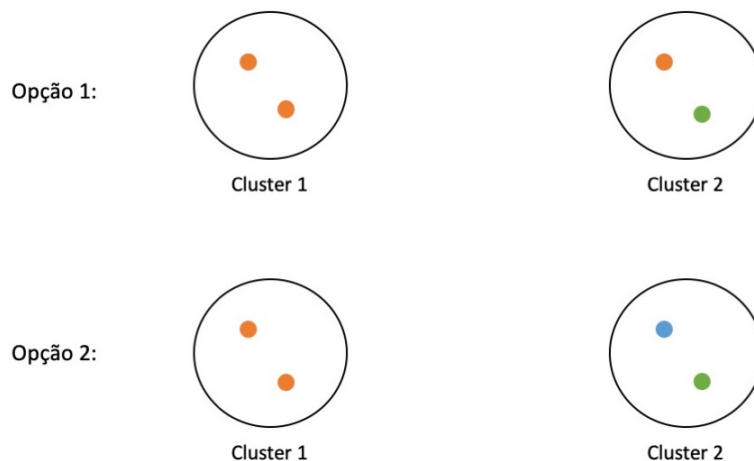
- (a) $\max_j |c_1 \cap t_j| = 3$ e $\max_j |c_2 \cap t_j| = 0$
- (b) $\max_j |c_1 \cap t_j| = 2$ e $\max_j |c_2 \cap t_j| = 1$
- (c) $\max_j |c_1 \cap t_j| = 1$ e $\max_j |c_2 \cap t_j| = 2$
- (d) $\max_j |c_1 \cap t_j| = 0$ e $\max_j |c_2 \cap t_j| = 3$

As opções (a) e (d) não são possíveis porque os clusters 2 e 1 só têm 2 observações cada um.

Opção (b)

Neste caso, as observações do cluster 1 são as duas classificadas corretamente. No cluster 2 uma é corretamente identificada e a outra não. Assim, as observações no cluster 1 têm a mesma classificação; uma das observações do cluster 2 tem classificação diferente das do cluster 1 e a outra pode ter classificação igual às do cluster 1 (opção 1) ou diferente, sendo que neste caso é também diferente da classificação da outra observação do cluster 2 (opção 2). Assim, conclui-se que o número verdadeiro de classes pode ser 2 ou 3.

Para visualizar melhor as opções possíveis fizemos os seguintes esquemas (bolas de cores diferentes representam classes verdadeiras diferentes):



Opção (c)

O raciocínio é semelhante ao da opção (b). Conclui-se que o número verdadeiro de classes pode ser 2 ou 3.

Programming - Código Python e Resultados Obtidos

1. Código Utilizado:
2. Código Utilizado:
3. Código Utilizado:
- 4.