Aprendizagem 2024/25

Homework IV – Group 51

(ist106221, ist105940)

I. Pen-and-paper

Consider the bivariate observations $\{\mathbf{x}_1 = \begin{bmatrix}1\\0\end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix}0\\2\end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix}3\\-1\end{bmatrix}\}$ and the multivariate

Gaussian mixture given by

$$\mathbf{u}_1 = \begin{bmatrix}2\\-1\end{bmatrix}, \quad \mathbf{u}_2 = \begin{bmatrix}1\\1\end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix}4 & 1\\1 & 4\end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}, \quad \pi_1 = 0.5, \quad \pi_2 = 0.5$$

Answer the following questions by presenting all intermediary steps, and use 3 decimal places in each.

1) [6v] Perform two epochs of the EM clustering algorithm and determine the new parameters

1

$P(x_i / \mu_k \Sigma_k) = \frac{1}{2\pi\sqrt{|\Sigma_k|}} \exp\left(-\frac{1}{2}(x_i - \mu_k)^T \Sigma^{-1} (x_i - \mu_k)\right)$

vamos calcular a ~~likel~~ likelihood para ~~a~~ dar todos os pontos

e sabemos que $P(x_i / \mu_k \Sigma_k) = N(x_i / \mu_k \Sigma_k)$

$\boxed{X_1 :}$

$x_1 - \mu_1 = \begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}2\\-1\end{bmatrix} = \begin{bmatrix}-1\\1\end{bmatrix}$

$\Sigma_1 = \begin{bmatrix}4 & 1\\1 & 4\end{bmatrix}$ $\det(\Sigma_1) = 15$ $\Sigma_1^{-1} = \frac{1}{15}\begin{bmatrix}4 & -1\\-1 & 4\end{bmatrix} \simeq$

$\simeq \begin{bmatrix}0.267 & -0.067\\-0.067 & 0.267\end{bmatrix}$

$P(x_1 / \overset{c=1}{\mu_1, \Sigma_1}) = N(x_1 / \mu_1 \Sigma_1) = \frac{1}{2\pi\sqrt{15}} \exp\left(-\frac{1}{2}[-1\ 1]\begin{bmatrix}0.267 & -0.067\\-0.067 & 0.267\end{bmatrix}\begin{bmatrix}-1\\1\end{bmatrix}\right)$

$\simeq 0.029$

cluster 2

$\boxed{X_1 :}$

$x_1 - \mu_2 = \begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}0\\-1\end{bmatrix}$ $\Sigma_2 = \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$ $\det(\Sigma_2) = 4$ $\Sigma_2^{-1} = \begin{bmatrix}0.5 & 0\\0 & 0.5\end{bmatrix}$

$P(x_1 / \overset{c=2}{\mu_2 \Sigma_2}) = \frac{1}{2\pi\sqrt{4}} \exp\left(-\frac{1}{2}[0\ -1]\begin{bmatrix}0.5 & 0\\0 & 0.5\end{bmatrix}\begin{bmatrix}0\\-1\end{bmatrix}\right) \simeq 0.062$

$\gamma_{ik} = \frac{\pi_k N(x_i / \mu_k \Sigma_k)}{\sum_{k=1}^{K} \pi_k N(x_i / \mu_k \Sigma_k)}$ $\gamma_{11} = \frac{\pi_1 N(x_1 / \mu_1 \Sigma_1)}{\pi_1 N(x_1 / \mu_1 \Sigma_1) + \pi_2 N(x_1 / \mu_2 \Sigma_2)} =$

$= \frac{0.5 \times 0.029}{0.5 \times 0.029 + 0.5 \times 0.062} \simeq \frac{0.015}{0.015 + 0.031} \simeq 0.326$

$$\gamma_{12} = \frac{0.062 \times 0.5}{0.062 \times 0.5 + 0.029 \times 0.5} \simeq \frac{0.031}{0.031 + 0.015} \simeq 0.674$$

para x_2

c_1

$$x_2 - \mu_1 = \begin{bmatrix} 0 \\ 2 \end{bmatrix} - \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$$

$$P(x_2 \mid c=1) = N(x_2 \mid \mu_1, \Sigma_1) = \frac{1}{2\pi\sqrt{15}} \exp\left(-\frac{1}{2} \begin{bmatrix} -2 & 3 \end{bmatrix} \begin{bmatrix} 0.267 & -0.067 \\ -0.067 & 0.267 \end{bmatrix} \begin{bmatrix} -2 \\ 3 \end{bmatrix}\right)$$

$$\simeq 0.05$$

c_2

$$x_2 - \mu_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

$$P(x_2 \mid c=2) = N(x_2 \mid \mu_2, \Sigma_2) = \frac{1}{2\pi\sqrt{4}} \exp\left(-\frac{1}{2} \begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix}\right)$$

$$\simeq 0.048$$

$$\gamma_{21} = \frac{0.5 \times 0.005}{0.5 \times 0.005 + 0.5 \times 0.048} \simeq \frac{0.003}{0.003 + 0.024} \simeq 0.111$$

$$\gamma_{22} = \frac{0.5 \times 0.048}{0.005 \times 0.5 + 0.048 \times 0.5} \simeq \frac{0.024}{0.003 + 0.024} \simeq 0.889$$

para x_3

$$x_3 - \mu_1 = \begin{pmatrix} 3 \\ -1 \end{pmatrix} - \begin{pmatrix} 2 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$P(x_3 \mid c=1) = N(x_3 \mid \mu_1, \Sigma_1) = \frac{1}{2\pi\sqrt{15}} \exp\left(\begin{pmatrix} 1 & 0 \end{pmatrix} \begin{bmatrix} 0.267 & -0.067 \\ -0.067 & 0.267 \end{bmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix}\right)$$

$$\simeq 0.036$$

$$x_3 - \mu_2 = \begin{pmatrix} 3 \\ -1 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$$

$$p(x_3 \mid \mu_2 \Sigma_2) = \frac{1}{2\pi\sqrt{4}} \exp\left(-\frac{1}{2} (2 \ -2) \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix} \begin{pmatrix} 2 \\ -2 \end{pmatrix} \right) \simeq 0.011$$

$$\gamma_{31} = \frac{0.5 \times 0.036}{0.5 \times 0.036 + 0.5 \times 0.011} \simeq \frac{0.018}{0.018 + 0.006} \simeq 0.75$$

$$\gamma_{32} = \frac{0.5 \times 0.011}{0.5 \times 0.011 + 0.5 \times 0.036} \simeq \frac{0.006}{0.006 + 0.018} \simeq 0.25$$

$$\gamma \begin{bmatrix} 0.326 & 0.674 \\ 0.111 & 0.889 \\ 0.75 & 0.25 \end{bmatrix}$$

M step

$$N_k = \sum_{i=1}^{m} \gamma_{ik} \ \cancel{+0.326+0.111+0.75 = 1.187}$$

$$N_1 = 0.326 + 0.111 + 0.75 = 1.187$$

$$N_2 = 0.674 + 0.889 + 0.25 = 1.813$$

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{m} \gamma_{ik} x_i$$

$$\mu_1 = \frac{1}{1.187} \times \left[0.326 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 0.111 \begin{pmatrix} 0 \\ 2 \end{pmatrix} + 0.75 \begin{pmatrix} 3 \\ -1 \end{pmatrix} \right] \simeq \begin{bmatrix} 2.17 \\ -0.445 \end{bmatrix}$$

$$\mu_2 = \frac{1}{1.813} \times \left[0.674 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 0.889 \begin{pmatrix} 0 \\ 2 \end{pmatrix} + 0.25 \begin{pmatrix} 3 \\ -1 \end{pmatrix} \right] \simeq \begin{bmatrix} 0.785 \\ 0.843 \end{bmatrix}$$

$$\pi_k = \frac{N_k}{N} \qquad \pi_1 = \frac{1.187}{3} \simeq 0.396 \qquad \pi_2 = \frac{1.813}{3} \simeq 0.604$$

$$\Sigma_k = \frac{\sum_{i=1}^{n} \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{n} \gamma_{ik}}$$

$$\Sigma_1 = \frac{1}{1.187}\left(0.326\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}2.17\\-0.045\end{bmatrix}\right)\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}2.17\\-0.045\end{bmatrix}\right)^T + \right.$$

$$+ 0.111\left(\begin{bmatrix}0\\2\end{bmatrix} - \begin{bmatrix}2.17\\-0.045\end{bmatrix}\right)\left(\begin{bmatrix}0\\2\end{bmatrix} - \begin{bmatrix}2.17\\-0.045\end{bmatrix}\right)^T +$$

$$\left. + 0.75\left(\begin{bmatrix}3\\-1\end{bmatrix} - \begin{bmatrix}2.17\\-0.045\end{bmatrix}\right)\left(\begin{bmatrix}3\\-1\end{bmatrix} - \begin{bmatrix}2.17\\-0.045\end{bmatrix}\right)^T \right) \simeq$$

$$\simeq \begin{bmatrix}1.252 & -0.93\\-0.93 & 0.808\end{bmatrix}$$

$$\Sigma_2 = \frac{1}{1.813}\left(0.674\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.785\\0.843\end{bmatrix}\right)\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.785\\0.843\end{bmatrix}\right)^T + \right.$$

$$+ 0.889\left(\begin{bmatrix}0\\2\end{bmatrix} - \begin{bmatrix}0.785\\0.843\end{bmatrix}\right)\left(\begin{bmatrix}0\\2\end{bmatrix} - \begin{bmatrix}0.785\\0.843\end{bmatrix}\right)^T$$

$$\left. + 0.25\left(\begin{bmatrix}3\\-1\end{bmatrix} - \begin{bmatrix}0.785\\0.843\end{bmatrix}\right)\left(\begin{bmatrix}3\\-1\end{bmatrix} - \begin{bmatrix}0.785\\0.843\end{bmatrix}\right)^T \right) \simeq$$

$$\simeq \begin{bmatrix}0.996 & -1.076\\-1.076 & 1.389\end{bmatrix}$$

Epoch 2

Como na primeira iteração vamos usar

$$N(x_i \mid \mu_k \Sigma_k) = \frac{1}{2\pi\sqrt{|\Sigma_k|}} \exp\left(-\frac{1}{2}(x_i - \mu_k)^T \Sigma^{-1} (x_i - \mu_k)\right)$$

x_1 c_1

$$x_1 - \mu_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 2.17 \\ -0.445 \end{bmatrix} = \begin{bmatrix} -1.17 \\ 0.445 \end{bmatrix}$$

$$\det(\Sigma_1) \simeq 0.147 \qquad \Sigma_1^{-1} \simeq \begin{bmatrix} 5.497 & 6.327 \\ 6.327 & 8.517 \end{bmatrix}$$

$$p(x_1 \mid c=1, \mu_1 \Sigma_1) = N(x_1 \mid \mu_1 \Sigma_1) = \frac{1}{2\pi\sqrt{0.147}} \exp\left(-\frac{1}{2}\begin{bmatrix} -1.17 & 0.445 \end{bmatrix}\begin{bmatrix} 5.497 & 6.327 \\ 6.327 & 8.517 \end{bmatrix}\begin{bmatrix} -1.17 \\ 0.445 \end{bmatrix}\right)$$

$$\simeq 0.112$$

c_2

$$x_1 - \mu_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0.785 \\ 0.843 \end{bmatrix} = \begin{bmatrix} 0.215 \\ -0.843 \end{bmatrix}$$

$$\det(\Sigma_2) \simeq 0.226 \qquad \Sigma_2^{-1} \simeq \begin{bmatrix} 6.146 & 4.761 \\ 4.761 & 4.407 \end{bmatrix}$$

$$p(x_1 \mid c=2, \mu_2 \Sigma_2) = N(x_1 \mid \mu_2 \Sigma_2) \simeq 0.144$$

$$\gamma_{11} = \frac{0.396 \times 0.112}{0.396 \times 0.112 + 0.604 \times 0.144} \simeq \frac{0.044}{0.044 + 0.087} \simeq 0.336$$

$$\gamma_{12} = \frac{0.604 \times 0.144}{0.396 \times 0.112 + 0.604 \times 0.144} \simeq \frac{0.087}{0.044 + 0.087} \simeq 0.664$$

x_2

c_1 $x_2 - \mu_1 = \begin{bmatrix} -2.17 \\ 2.445 \end{bmatrix}$

$P(x_2 \mid \overset{c=1}{\cancel{\mu_1 \Sigma_1}}) = N(x_2 \mid \mu_1 \Sigma_1) \simeq 0.003$

c_2 $x_2 - \mu_2 = \begin{bmatrix} -0.785 \\ 1.157 \end{bmatrix}$

$P(x_2 \mid \overset{c=2}{\cancel{\mu_2 \Sigma_2}}) = N(x_2 \mid \mu_2 \Sigma_2) \simeq 0.199$

$\gamma_{21} = \dfrac{0.003 \times 0.396}{0.003 \times 0.396 + 0.199 \times 0.604} \simeq \dfrac{0.001}{0.001 + 0.12} \simeq \cancel{\dfrac{0.396}{0.008}}$

$\gamma_{22} = \dfrac{0.199 \times 0.604}{0.003 \times 0.396 + 0.199 \times 0.604} \simeq \dfrac{0.12}{0.001 + 0.12} \simeq 0.992$

x_3

c_1 $x_3 - \mu_1 = \begin{bmatrix} 0.83 \\ -0.555 \end{bmatrix}$

$P(x_3 \mid \overset{c=1}{\cancel{\mu_1 \Sigma_1}}) = N(x_3 \mid \mu_1 \Sigma_1) \simeq 0.31$

c_2 $x_3 - \mu_2 = \begin{bmatrix} 2.15 \\ -1.843 \end{bmatrix}$

$P(x_3 \mid \overset{c=2}{\cancel{\mu_2 \Sigma_2}}) = N(x_3 \mid \mu_2 \Sigma_2) \simeq 0.015$

$\gamma_{31} = \dfrac{0.31 \times 0.396}{0.31 \times 0.396 + 0.015 \times 0.604} \simeq \dfrac{0.123}{0.123 + 0.009} \simeq 0.932$

$\gamma_{32} = \dfrac{0.015 \times 0.604}{0.015 \times 0.604 + 0.31 \times 0.396} \simeq \dfrac{0.009}{0.009 + 0.123} \simeq 0.068$

$$\gamma = \begin{bmatrix} 0.336 & 0.664 \\ 0.008 & 0.992 \\ 0.932 & 0.068 \end{bmatrix}$$

M step

$$N_k = \sum_{i=1}^{m} \gamma_{ik}$$

$$N_1 = 0.336 + 0.008 + 0.932 = 1.276$$

$$N_2 = 0.664 + 0.992 + 0.068 = 1.724$$

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{m} \gamma_{ik} x_i$$

$$\mu_1 = \frac{1}{1.276} \left(0.336 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 0.008 \begin{pmatrix} 0 \\ 2 \end{pmatrix} + 0.932 \begin{pmatrix} 3 \\ -1 \end{pmatrix} \right) \simeq \begin{pmatrix} 2.455 \\ -0.718 \end{pmatrix}$$

$$\mu_2 = \frac{1}{1.724} \left(0.664 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 0.992 \begin{pmatrix} 0 \\ 2 \end{pmatrix} + 0.068 \begin{pmatrix} 3 \\ -1 \end{pmatrix} \right) \simeq \begin{pmatrix} 0.503 \\ 1.111 \end{pmatrix}$$

$$\pi_k = \frac{N_k}{N} \qquad \pi_1 = \frac{1.276}{3} \simeq 0.425$$

$$\pi_2 = \frac{1.724}{3} \simeq 0.575$$

$$\Sigma_k = \frac{\sum_{i=1}^{m} \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{m} \gamma_{ik}}$$

$$\Sigma_1 = \frac{1}{1.276} \left(0.336 \left(\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 2.455 \\ -0.718 \end{pmatrix} \right) \left(\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 2.455 \\ -0.718 \end{pmatrix} \right)^T + \right.$$

$$+ 0.008 \left(\begin{pmatrix} 0 \\ 2 \end{pmatrix} - \begin{pmatrix} 2.455 \\ -0.718 \end{pmatrix} \right) \left(\begin{pmatrix} 0 \\ 2 \end{pmatrix} - \begin{pmatrix} 2.455 \\ -0.718 \end{pmatrix} \right)^T +$$

$$\left. + 0.932 \left(\begin{pmatrix} 3 \\ -1 \end{pmatrix} - \begin{pmatrix} 2.455 \\ -0.718 \end{pmatrix} \right) \left(\begin{pmatrix} 3 \\ -1 \end{pmatrix} - \begin{pmatrix} 2.455 \\ -0.718 \end{pmatrix} \right)^T \right) \simeq \begin{bmatrix} 0.812 & -0.429 \\ -0.429 & 0.24 \end{bmatrix}$$

$$\Sigma_2 = \frac{1}{1.724} \times \left[0.665 \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0.503 \\ 1.111 \end{bmatrix} \right) \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0.503 \\ 1.111 \end{bmatrix} \right)^T + \right.$$

$$+ 0.992 \left(\begin{bmatrix} 0 \\ 2 \end{bmatrix} - \begin{bmatrix} 0.503 \\ 1.111 \end{bmatrix} \right) \left(\begin{bmatrix} 0 \\ 2 \end{bmatrix} - \begin{bmatrix} 0.503 \\ 1.111 \end{bmatrix} \right)^T +$$

$$\left. + 0.068 \left(\begin{bmatrix} 3 \\ -1 \end{bmatrix} - \begin{bmatrix} 0.503 \\ 1.111 \end{bmatrix} \right) \left(\begin{bmatrix} 3 \\ -1 \end{bmatrix} - \begin{bmatrix} 0.503 \\ 1.111 \end{bmatrix} \right)^T \right]$$

$$\simeq \begin{bmatrix} 0.487 & -0.678 \\ -0.678 & 1.106 \end{bmatrix}$$

2) Using the final parameters computed in previous question:

a)[1v] Perform a hard assignment of observations to clusters under a MAP assumption.

1) vamos calcular a likelihood de cada ponto para cada cluster

$p(x_i \mid C=k) = N(x_i \mid \mu_k \Sigma_k)$

$$N(x_i \mid \mu_k \Sigma_k) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k)\right)$$

$\det(\Sigma_1) \simeq 0.011$ $\qquad \Sigma^{-1} \simeq \begin{pmatrix} 24.818 & 39 \\ 39 & 73.818 \end{pmatrix}$

$\det(\Sigma_2) \simeq 0.079$ $\qquad \Sigma^{-2} \simeq \begin{pmatrix} 14 & 8.582 \\ 8.582 & 6.165 \end{pmatrix}$

$x_1 - \mu_1 = \begin{bmatrix} -1.455 \\ 0.718 \end{bmatrix}$ $\qquad x_1 - \mu_2 \simeq \begin{bmatrix} 0.497 \\ -1.111 \end{bmatrix}$

$x_2 - \mu_1 = \begin{bmatrix} -2.455 \\ 2.718 \end{bmatrix}$ $\qquad x_2 - \mu_2 = \begin{bmatrix} -0.503 \\ 0.889 \end{bmatrix}$

$x_3 - \mu_1 = \begin{bmatrix} 0.545 \\ -0.282 \end{bmatrix}$ $\qquad x_3 - \mu_2 = \begin{bmatrix} 2.497 \\ -2.111 \end{bmatrix}$

$p(x_1 \mid c=1) = N(x_1 \mid \mu_1 \Sigma_1) \simeq 0.382$

$p(x_1 \mid c=2) = N(x_1 \mid \mu_2 \Sigma_2) \simeq 0.256$

$p(x_2 \mid C=1) = N(x_2 \mid \mu_1 \Sigma_1) \simeq 0$

$p(x_2 \mid C=2) = N(x_2 \mid \mu_2 \Sigma_2) \simeq 0.391$

$p(x_3 \mid C=1) = N(x_3 \mid \mu_1 \Sigma_1) \simeq 1.266$

$p(x_3 \mid C=2) = N(x_3 \mid \mu_2 \Sigma_2) \simeq 0$

agora calculamos a probabilidade conjunta

$$P(C=1, X_1) = 0.382 \times 0.425 \simeq 0.162$$

$$P(C=2, X_1) = 0.256 \times 0.575 \simeq 0.147$$

$$P(C=1, X_2) = 0 \times 0.425 \simeq 0$$

$$P(C=2, X_2) = 0.391 \times 0.575 \simeq 0.225$$

$$P(C=1, X_3) = \cancel{0.4}\ 1.266 \times 0.425 \simeq 0.538$$

$$P(C=2, X_3) = 0 \times 0.975 \simeq 0$$

Como

$$P(C=1, X_1) > P(C=2, X_1) \quad X_1 \in C=1$$

$$P(C=1, X_2) < P(C=2, X_2) \quad X_2 \in C=2$$

$$P(C=1, X_3) > P(C=2, X_3) \quad X_3 \in C=1$$

b) [2v] Compute the silhouette of the larger cluster (the one that has more observations assigned to it) using the Euclidean distance.

---

2 b) calcular a silhueta do cluster 1

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

$$\cancel{a(i) = \sqrt{}}$$

$$d(x_1, x_3) = \sqrt{(3-1)^2 + (-1+0)^2} = \sqrt{5} \simeq 2{,}236$$

$$a(1) = 2{,}236 = a(3)$$

$$d(x_1, x_2) = \sqrt{(0-1)^2 + (2-0)^2} = \sqrt{5} \simeq 2.236$$

$$b(1) = 2.236$$

$$d(x_3, x_2) = \sqrt{(0-3)^2 + (2-(-1))^2} = \sqrt{18} \simeq 4.243$$

$$b(3) = 4.243$$

$$s(1) = \frac{b(1) - a(1)}{\max(a(1), b(1))} = \frac{2.236 - 2.236}{2.236} = 0$$

$$s(3) = \frac{b(3) - a(3)}{\max(a(3), b(3))} = \frac{4.243 - 2.236}{4.243}$$

$$\simeq 0{,}473$$

$$s = \frac{s(1) + s(3)}{2} = \frac{0 + 0{,}473}{2} = \cancel{0.236}\ 0.237$$

II. Programming and critical analysis

In the next exercise you will use the `accounts.csv` dataset. This dataset contains account details of bank clients, and the target variable *y* is binary ('has the client subscribed a term deposit?'). Select the first 8 features and remove duplicates and null values.

> ***Hint:*** You can use `get_dummies()` to change the feature type (e.g. `pd.get_dummies(data, drop_first=True)`).

```python
import pandas as pd
import numpy as np
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
from sklearn.preprocessing import MinMaxScaler
from copy import deepcopy
# Load the data
data = pd.read_csv("accounts.csv", delimiter=',')
# Convert 'deposit' column from 'yes'/'no' to 1/0
data['deposit'] = data['deposit'].map({'yes': 1, 'no': 0})
# Separate features (X) and target (y)
X = data.drop('deposit', axis=1)
# Remove duplicates and null values
X = X.iloc[:, :8]
X = X.drop_duplicates().dropna()
X.reset_index(drop=True, inplace=True)
data2 = deepcopy(X)
```

3) Normalize the data using MinMaxScaler:

a) [4v] Using sklearn, apply k-means clustering (without targets) on the normalized data with k={2,3,4,5,6,7,8}, max_iter=500 and random_state=42. Plot the different sum of squared errors (SSE) using the _inertia attribute of k-means according to the number of clusters.

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans

# Convert categorical features to numerical using get_dummies
data1 = pd.get_dummies(X, drop_first=True)

# Normalize the data using MinMaxScaler
scaler = MinMaxScaler()
X_scaled = scaler.fit_transform(data1)

# Apply k-means clustering with k={2,3,4,5,6,7,8}
sse = []
for k in range(2, 9):
    kmeans = KMeans(n_clusters=k, max_iter=500, random_state=42)
    kmeans.fit(X_scaled)
    sse.append(kmeans.inertia_)

# Plot the SSE for each k
import matplotlib.pyplot as plt
plt.plot(range(2, 9), sse, marker='o')
plt.xlabel('Number of clusters')
plt.ylabel('Sum of Squared Errors (SSE)')
plt.title('SSE vs. Number of Clusters')
plt.show()
```

b) [1.5v] According to the previous plot, how many underlying customer segments (clusters) should there be? Explain based on the trade-off between the clusters and inertia.

No gráfico, vemos uma queda acentuada no SSE de 2 até 6 clusters, com a redução da taxa de diminuição a partir do ponto de 6 clusters. A partir deste ponto, adicionar mais clusters resulta em ganhos menores na redução do SSE. Assim, o número ideal de clusters parece ser 6, pois oferece uma boa divisão dos dados com uma inércia (ou variabilidade intra-cluster) razoável, evitando uma complexidade desnecessária com mais clusters.

c) [1.5v] Would k-modes be a better clustering approach? Explain why based on the dataset features.

k-modes seria uma melhor opção do que k-means, porque os dados fornecidos apresentam muitas variáveis categóricas e apenas duas variáveis numéricas.

O modelo k-means é bom a operar em modelos numéricos mas tem dificuldades em categorias. O que não acontece com o k-modes sendo assim esta uma melhor opção para avaliar estes dados.

4) Normalize the data using StandardScaler:

```python
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler

# Scale the data using StandardScaler
scaler = StandardScaler()
X_scaled_standard = scaler.fit_transform(pd.get_dummies(X, drop_first=True))
```

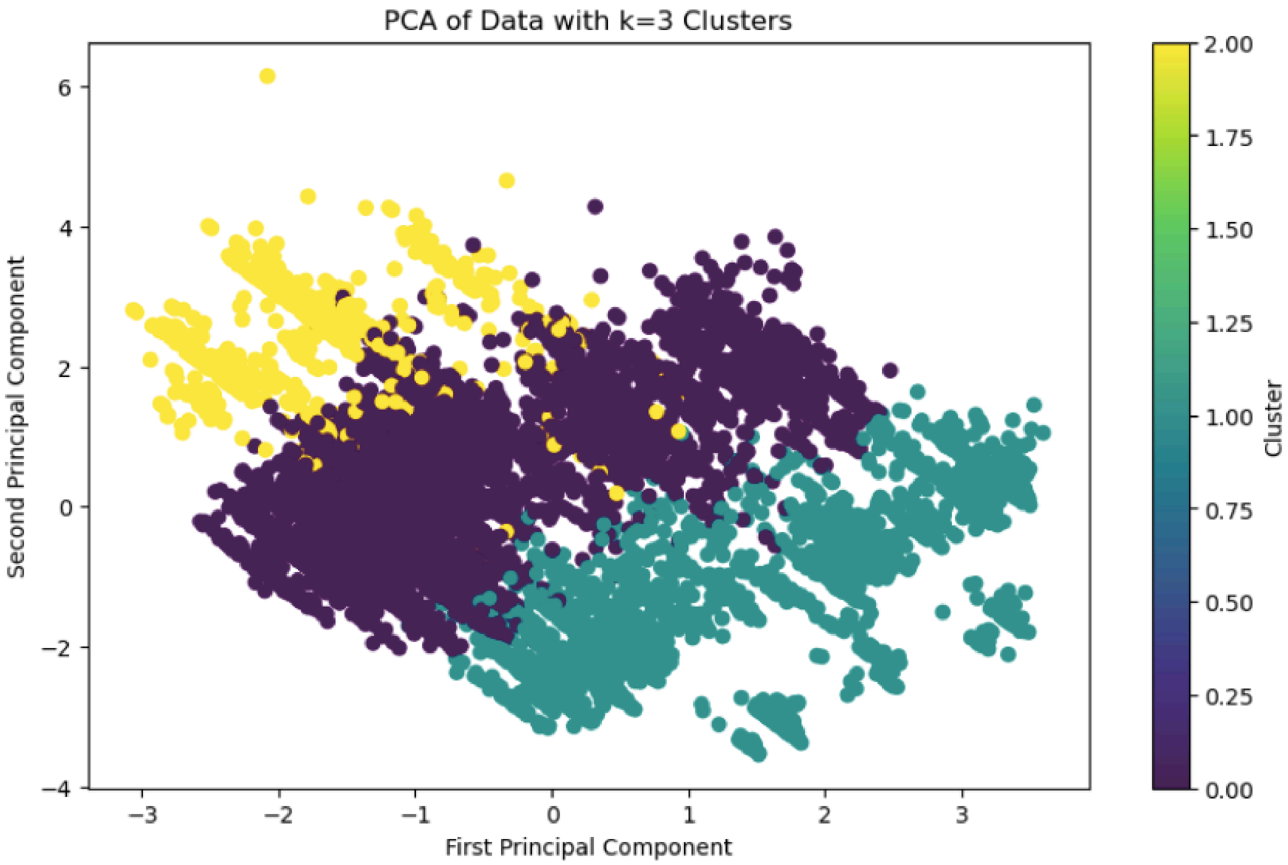
a) [1v] Apply PCA to the data. How much variability is explained by the top 2 components?

```python
# Apply PCA
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled_standard)

# Variability explained by the top 2 components
explained_variance = pca.explained_variance_ratio_
print(f"Variability explained by the top 2 components: {explained_variance.sum() * 100:.2f}%")
```

```
Variability explained by the top 2 components: 22.75%
```

b) [1v] Apply k-means clustering with k=3 and random_state=42 (all other arguments as default) and use the original 8 features. Next, provide a scatterplot according to the first 2 principal components. Can we clearly separate the clusters?Justify.

```python
# Apply k-means clustering with k=3
kmeans_3 = KMeans(n_clusters=3, max_iter=500, random_state=42)
clusters = kmeans_3.fit_predict(X_scaled_standard)

# Scatter plot of the first 2 principal components
plt.figure(figsize=(10, 6))
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=clusters, cmap='viridis', marker='o')
plt.xlabel('First Principal Component')
plt.ylabel('Second Principal Component')
plt.title('PCA of Data with k=3 Clusters')
plt.colorbar(label='Cluster')
plt.show()
```

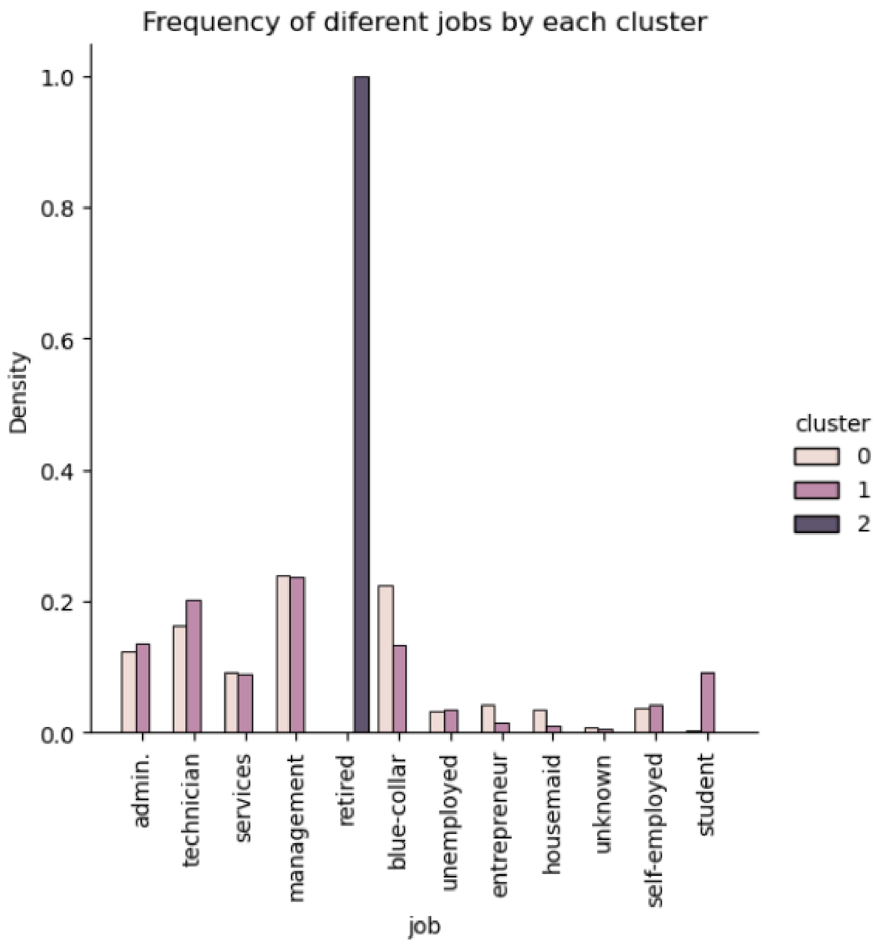


As duas principais componentes explicam aproximadamente 11.68% e 11.08% da variância. Isto significa que os dois juntos contam 22.75% da variância total do modelo.

Conseguimos separar os clusters porque existe uma visível diferença entre a localidade de cada cluster . Isto significa que o pca entre estas características dá-nos uma separação suficiente entre clusters.

c. [2v] Plot the cluster conditional features of the frequencies of “job” and ”education” according to the clusters obtained in the previous question (2b.). Use sns.displot (see Data Exploration notebook), with multiple="dodge", stat='density', shrink=0.8 and common_norm=False. Describe the main differences between the clusters in no more than half a page.

```
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler


data2['cluster'] = clusters

sns.displot(data2, x="job", hue="cluster", multiple="dodge", stat="density", shrink=0.8, common_norm=False)
plt.title('Frequency of diferent jobs by each cluster')
plt.xticks(rotation=90)
plt.show()

sns.displot(data2, x="education", hue="cluster", multiple="dodge", stat="density", shrink=0.8, common_norm=False)
plt.title('Frequency of diferent educations by each cluster')
plt.show()
```



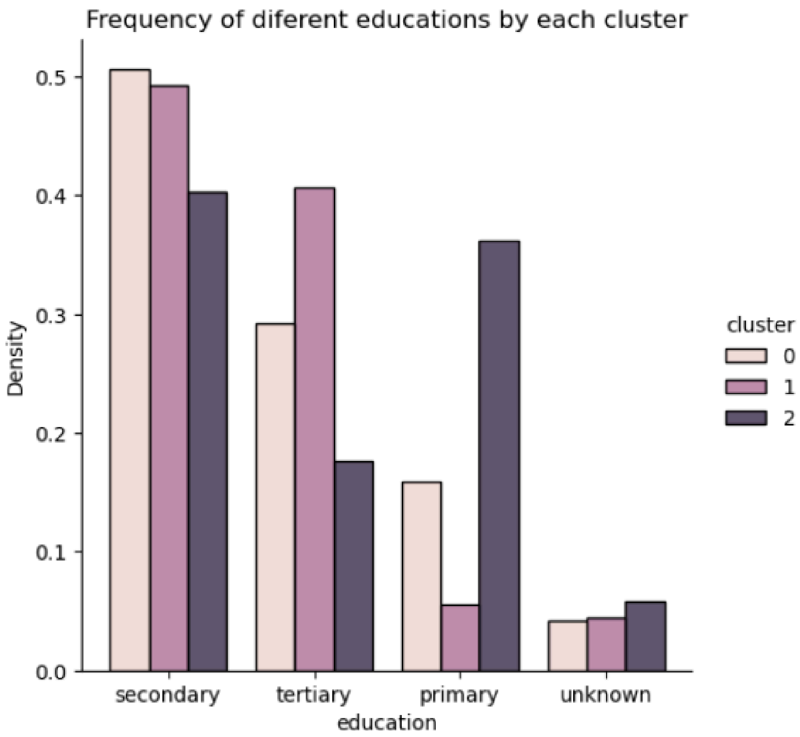


Grafico 1 :

- Cluster 0: Apresenta uma distribuição mais equilibrada entre diversas profissões, com destaque para administração,técnicos ,serviços , "blue-collar" , entrepreneurs , housemaid .- Cluster 1: Destaca-se pela alta frequência de management e administração, seguido de serviços, técnicos , admin e student.

- Cluster 2: Apresenta uma concentração significativa de pessoas reformadas.

Grafico 2 :

- Cluster 0: Apresenta uma distribuição mais equilibrada entre os níveis de educação, com uma predominância para o ensino secundário. Isso corrobora a hipótese de que este cluster é mais heterogêneo em termos de idade e nível socioeconômico.

- Cluster 1:Destaca-se pela alta frequência de pessoas com ensino terciário e secundário.

- Cluster 2: Apresenta uma concentração significativa de pessoas com ensino primário e secundário. Isso sugere que este cluster é composto por uma geração mais idosa que não frequentava o ensino terciário .

Tendo em vista a análise dos clusters, podem-se identificar perfis distintos de acordo com a ocupação e o nível educacional, trazendo insights valiosos sobre as características demográficas e socioeconômicas de cada grupo. Para o Cluster 0, tem-se um perfil mais heterogêneo, que varia entre várias profissões e níveis de escolaridade, ainda que sobrepondo ligeiramente as pessoas com ensino secundário completo e terciário. Esse não é o caso do Cluster 1, em que se obtém uma alta concentração de indivíduos em ocupações administrativas, de gestão e um predomínio de pessoas com ensino terciário completo: trata-se, portanto, de um perfil mais qualificado e socioeconomicamente elevado (management e technical).Por fim, o Cluster 2 concentra-se em pessoas reformadas, com uma distribuição educacional centrada no ensino primário, e secundário, reforçando a hipótese de que esse grupo inclui principalmente aqueles fora do mercado de trabalho ativo. Essas distinções ajudam a entender a composição dos grupos e a adaptar estratégias para cada perfil identificado.

END