Aprendizagem 2022/23

Homework IV – Group 084

I. Pen-and-paper

1)

①

$$\gamma_{ki} = p(C_k \mid x_i) = \frac{p(x_i \mid cluster=k)\,p(cluster=k)}{p(x_i)} = \frac{\pi_k \cdot \mathcal{N}(x_i \mid u_k, \Sigma_k)}{\sum_k \pi_k \cdot \mathcal{N}(x_i \mid u_k, \Sigma_i)}$$

$$N_k = \sum_{i=1}^{n} \gamma_k^{(i)} \qquad u_k = \frac{1}{N_k}\sum_{i=1}^{n} \gamma_k^{(i)} \cdot x_i \qquad \Sigma_k = \frac{1}{N_k}\sum_{i=1}^{n} \gamma_k^{(i)} (x_i - u_k)(x_i - u_k)^T$$

$$\pi_k = p(C_k) = \frac{N_k}{N}$$

$x_1 = \begin{bmatrix}1\\2\end{bmatrix}$:

$$\mathcal{N}\left([1\ 2]^T \mid \underbrace{[2\ 2]^T}_{u_1}, \underbrace{\begin{bmatrix}2&1\\1&2\end{bmatrix}}_{\Sigma_1}\right) \overset{cluster=1}{=} \frac{1}{2\pi}\cdot\frac{1}{\sqrt{3}}\cdot e^{-\frac{1}{2}\left(\begin{bmatrix}1\\2\end{bmatrix}-\begin{bmatrix}2\\2\end{bmatrix}\right)^T\begin{bmatrix}2&-1\\-1&2\end{bmatrix}\left(\begin{bmatrix}1\\2\end{bmatrix}-\begin{bmatrix}2\\2\end{bmatrix}\right)\cdot\frac{1}{3}} =$$

$$= \frac{1}{2\pi\sqrt{3}}\cdot e^{-\frac{1}{2}[-1\ 0]\begin{bmatrix}2&-1\\-1&2\end{bmatrix}\begin{bmatrix}-1\\0\end{bmatrix}\cdot\frac{1}{3}} = \frac{1}{2\pi\sqrt{3}}\cdot e^{-\frac{1}{2}\cdot 2\cdot\frac{1}{3}} = 0.0658$$

$$\mathcal{N}\left([1\ 2]^T \mid \underbrace{[0\ 0]^T}_{u_2}, \underbrace{\begin{bmatrix}2&0\\0&2\end{bmatrix}}_{\Sigma_2}\right) \overset{cluster=2}{=} \frac{1}{2\pi}\cdot\frac{1}{2}\cdot e^{-\frac{1}{2}\left(\begin{bmatrix}1\\2\end{bmatrix}-\begin{bmatrix}0\\0\end{bmatrix}\right)^T\begin{bmatrix}2&0\\0&2\end{bmatrix}\left(\begin{bmatrix}1\\2\end{bmatrix}-\begin{bmatrix}0\\0\end{bmatrix}\right)\cdot\frac{1}{4}} =$$

$$= \frac{1}{4\pi}\cdot e^{-\frac{1}{2}[1\ 2]\begin{bmatrix}2&0\\0&2\end{bmatrix}\begin{bmatrix}1\\2\end{bmatrix}\cdot\frac{1}{4}} = \frac{1}{4\pi}\cdot e^{-\frac{1}{2}\cdot 10\cdot\frac{1}{4}} = 0.0228$$

$$\gamma_1 = p(cluster=1 \mid x_1) = \frac{\pi_1\cdot\mathcal{N}(x_1\mid u_1,\Sigma_1)}{\sum_{k=1}^{2}\pi_k\cdot\mathcal{N}(x_1\mid u_k,\Sigma_k)} = \frac{0.5\cdot 0.0658}{0.5\cdot 0.0658 + 0.5\cdot 0.0228}$$

$$= 0.7428$$

$$\gamma_2 = p(cluster=2 \mid x_1) = \frac{\pi_2\cdot\mathcal{N}(x_1\mid u_2,\Sigma_2)}{\sum_{k=1}^{2}\pi_k\cdot\mathcal{N}(x_1\mid u_k,\Sigma_k)} = \frac{0.5\cdot 0.0228}{0.5\cdot 0.0658 + 0.5\cdot 0.0228} =$$

$$= 0.2572$$

$x_2 = \begin{bmatrix}-1\\1\end{bmatrix}$:

$$\mathcal{N}\left([-1\ 1]^T \mid \underbrace{[2\ 2]^T}_{u_1}, \underbrace{\begin{bmatrix}2&1\\1&2\end{bmatrix}}_{\Sigma_1}\right) \overset{cluster=1}{=} \frac{1}{2\pi}\cdot\frac{1}{\sqrt{3}}\cdot e^{-\frac{1}{2}\left(\begin{bmatrix}-1\\1\end{bmatrix}-\begin{bmatrix}2\\2\end{bmatrix}\right)^T\begin{bmatrix}2&-1\\-1&2\end{bmatrix}\left(\begin{bmatrix}-1\\1\end{bmatrix}-\begin{bmatrix}2\\2\end{bmatrix}\right)\cdot\frac{1}{3}} =$$

$$= \frac{1}{2\pi\sqrt{3}}\cdot e^{-\frac{1}{2}[-3\ -1]\begin{bmatrix}2&-1\\-1&2\end{bmatrix}\begin{bmatrix}-3\\-1\end{bmatrix}\cdot\frac{1}{3}} = \frac{1}{2\pi\sqrt{3}}\cdot e^{-\frac{1}{2}\cdot 14\cdot\frac{1}{3}} = 0.0089$$

$$\mathcal{N}\left([-1\ 1]^T \mid \underbrace{[0\ 0]^T}_{u_2}, \underbrace{\begin{bmatrix}2&0\\0&2\end{bmatrix}}_{\Sigma_2}\right) \overset{cluster=2}{=} \frac{1}{2\pi}\cdot\frac{1}{2}\cdot e^{-\frac{1}{2}\left(\begin{bmatrix}-1\\1\end{bmatrix}-\begin{bmatrix}0\\0\end{bmatrix}\right)^T\begin{bmatrix}2&0\\0&2\end{bmatrix}\left(\begin{bmatrix}-1\\1\end{bmatrix}-\begin{bmatrix}0\\0\end{bmatrix}\right)\cdot\frac{1}{4}} =$$

$$= \frac{1}{4\pi}\cdot e^{-\frac{1}{2}[-1\ 1]\begin{bmatrix}2&0\\0&2\end{bmatrix}\begin{bmatrix}-1\\1\end{bmatrix}\cdot\frac{1}{4}} = \frac{1}{4\pi}\cdot e^{-\frac{1}{2}\cdot 4\cdot\frac{1}{4}} = 0.0483$$

$$\gamma_1 = \frac{\pi_1\cdot\mathcal{N}(x_2\mid u_1,\Sigma_1)}{\sum_{k=1}^{2}\pi_k\cdot\mathcal{N}(x_2\mid u_k,\Sigma_k)} = \frac{0.5\cdot 0.0089}{0.5\cdot 0.0089 + 0.5\cdot 0.0483} = 0.1558$$

$$\gamma_2 = \frac{\pi_2\cdot\mathcal{N}(x_2\mid u_2,\Sigma_2)}{\sum_{k=1}^{2}\pi_k\cdot\mathcal{N}(x_2\mid u_k,\Sigma_k)} = \frac{0.5\cdot 0.0483}{0.5\cdot 0.0089 + 0.5\cdot 0.0483} = 0.8442$$

$x_3 = \begin{bmatrix}1\\0\end{bmatrix}$:

$$\mathcal{N}\left([1\ 0]^T \mid \overset{u_1}{[2\ 2]^T}, \overset{\Sigma_1}{\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix}}\right) \overset{\text{cluster 1}}{=} \frac{1}{2\pi} \cdot \frac{1}{\sqrt{3}} \cdot e^{-\frac{1}{2}\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}2\\2\end{bmatrix}\right)^T \begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}2\\2\end{bmatrix}\right)\cdot\frac{1}{3}} =$$

$$= \frac{1}{2\pi\sqrt{3}} \cdot e^{-\frac{1}{2}[-1\ -2]\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}\begin{bmatrix}-1\\-2\end{bmatrix}\cdot\frac{1}{3}} = \frac{1}{2\pi\sqrt{3}} \cdot e^{-\frac{1}{2}\cdot 6 \cdot \frac{1}{3}} = 0.0338$$

$$\mathcal{N}\left([1\ 0]^T \mid \overset{u_2}{[0\ 0]^T}, \overset{\Sigma_2}{\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}}\right) \overset{\text{cluster 2}}{=} \frac{1}{2\pi} \cdot \frac{1}{2} \cdot e^{-\frac{1}{2}\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0\\0\end{bmatrix}\right)^T \begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0\\0\end{bmatrix}\right)\cdot\frac{1}{4}} =$$

$$= \frac{1}{4\pi} \cdot e^{-\frac{1}{2}[1\ 0]\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix}\cdot\frac{1}{4}} = \frac{1}{4\pi} \cdot e^{-\frac{1}{2}\cdot 2 \cdot \frac{1}{4}} = 0.0620$$

$$\gamma_1 = \frac{\pi_1 \cdot \mathcal{N}(x_3 \mid u_1, \Sigma_1)}{\sum_{k=1}^{2} \pi_k \cdot \mathcal{N}(x_3 \mid u_k, \Sigma_k)} = \frac{0.5 \cdot 0.0338}{0.5 \cdot 0.0338 + 0.5 \cdot 0.0620} = 0.3528$$

$$\gamma_2 = \frac{\pi_2 \cdot \mathcal{N}(x_3 \mid u_2, \Sigma_2)}{\sum_{k=1}^{2} \pi_k \cdot \mathcal{N}(x_3 \mid u_k, \Sigma_k)} = \frac{0.5 \cdot 0.0620}{0.5 \cdot 0.0338 + 0.5 \cdot 0.0620} = 0.6472$$

<u>Re-estimation:</u> $N_1 = \sum_{i=1}^{3} \gamma_1^{(i)} = 0.7428 + 0.1558 + 0.3528 = 1.2514$

$$N_2 = \sum_{i=1}^{3} \gamma_2^{(i)} = 0.2572 + 0.8442 + 0.6472 = 1.7486$$

$$u_1 = \frac{1}{N_1}\sum_{i=1}^{3} \gamma_1^{(i)} \cdot x_i = \frac{1}{1.2514} \cdot \left(0.7428 \cdot \begin{bmatrix}1\\2\end{bmatrix} + 0.1558 \cdot \begin{bmatrix}-1\\1\end{bmatrix} + 0.3528 \cdot \begin{bmatrix}1\\0\end{bmatrix}\right) = \begin{bmatrix}0.7510\\1.3117\end{bmatrix}$$

$$u_2 = \frac{1}{N_2}\sum_{i=1}^{3} \gamma_2^{(i)} \cdot x_i = \frac{1}{1.7486} \left(0.2572 \cdot \begin{bmatrix}1\\2\end{bmatrix} + 0.8442 \cdot \begin{bmatrix}-1\\1\end{bmatrix} + 0.6472 \cdot \begin{bmatrix}1\\0\end{bmatrix}\right) = \begin{bmatrix}0.0344\\0.7770\end{bmatrix}$$

$$\Sigma_1 = \frac{1}{N_1}\sum_{i=1}^{3} \gamma_1^{(i)} (x_i - u_1)(x_i - u_1)^T = \frac{1}{1.2514} \cdot \left(0.7428 \left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0.7510\\1.3117\end{bmatrix}\right)\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0.7510\\1.3117\end{bmatrix}\right)^T +\right.$$

$$\left. + 0.1558 \left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.7510\\1.3117\end{bmatrix}\right)\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.7510\\1.3117\end{bmatrix}\right)^T + 0.3528 \left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.7510\\1.3117\end{bmatrix}\right)\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.7510\\1.3117\end{bmatrix}\right)^T\right) =$$

$$= \begin{bmatrix}0.4360 & 0.0776\\ 0.0776 & 0.7784\end{bmatrix}$$

$$\Sigma_2 = \frac{1}{N_2}\sum_{i=1}^{3} \gamma_2^{(i)} (x_i - u_2)(x_i - u_2)^T = \frac{1}{1.7486} \left(0.2572 \left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0.0344\\0.7770\end{bmatrix}\right)\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0.0344\\0.7770\end{bmatrix}\right)^T +\right.$$

$$\left. + 0.8442 \left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.0344\\0.7770\end{bmatrix}\right)\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.0344\\0.7770\end{bmatrix}\right)^T + 0.6472 \left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.0344\\0.7770\end{bmatrix}\right)\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.0344\\0.7770\end{bmatrix}\right)^T\right) =$$

$$= \begin{bmatrix}0.9988 & -0.2154\\ -0.2154 & 0.4675\end{bmatrix}$$

$$\pi_1 = \frac{N_1}{N} = \frac{1.2514}{1.2514 + 1.7486} = 0.4171$$

$$\pi_2 = \frac{N_2}{N} = \frac{1.7486}{1.2514 + 1.7486} = 0.5829$$

2)

a)

2a) $x_1=\begin{bmatrix}1\\2\end{bmatrix};\ x_2=\begin{bmatrix}-1\\1\end{bmatrix};\ x_3=\begin{bmatrix}1\\0\end{bmatrix}$

$P(c_K|x_n)=\dfrac{P(x_n|c_K)\cdot P(c_K)}{P(c_K)}$; Tendo em conta a MAP assumption queremos $\underset{c_K}{\operatorname{argmax}}\ P(x_n|c_K)\cdot P(c_K)$

$\mu_1=\begin{bmatrix}0,7510\\1,3112\end{bmatrix};\ \mu_2=\begin{bmatrix}0,0344\\0,7770\end{bmatrix};\ \Sigma_1=\begin{bmatrix}0,4360 & 0,0776\\0,0776 & 0,7784\end{bmatrix};\ \Sigma_2=\begin{bmatrix}0,9988 & -0,2154\\-0,2154 & 0,4675\end{bmatrix}$

$\det(\Sigma_1)=0,3334;\ \det(\Sigma_2)=0,4205;\ \Sigma_1^{-1}=\begin{bmatrix}2,3347 & -0,2328\\-0,2328 & 1,3077\end{bmatrix};$

$\Sigma_2^{-1}=\begin{bmatrix}1,1118 & 0,5122\\0,5122 & 2,3753\end{bmatrix};\ \pi_1=0,4171;\ \pi_2=0,5829$

x_1:

$P(x_1|c=1)=N(x_1|\mu_1,\Sigma_1)=\frac{1}{2\pi}\cdot\frac{1}{\sqrt{\det(\Sigma_1)}}\cdot\exp\left(-\frac{1}{2}(x_1-\mu_1)^T\Sigma_1^{-1}(x_1-\mu_1)\right)=$

$=\frac{1}{2\pi\cdot\sqrt{0,3334}}\cdot\exp\left(-\frac{1}{2}\left(\begin{bmatrix}1\\2\end{bmatrix}-\begin{bmatrix}0,7510\\1,3112\end{bmatrix}\right)^T\begin{bmatrix}2,3347 & -0,2328\\-0,2328 & 1,3077\end{bmatrix}\left(\begin{bmatrix}1\\2\end{bmatrix}-\begin{bmatrix}0,7510\\1,3112\end{bmatrix}\right)\right)$

$=\frac{1}{2\pi\cdot0,5774}\exp\left(-\frac{1}{2}\cdot0,6853\right)=0,1957$

$P(x_1|c=2)=N(x_1|\mu_2,\Sigma_2)=\frac{1}{2\pi}\cdot\frac{1}{\sqrt{\det(\Sigma_2)}}\cdot\exp\left(-\frac{1}{2}(x_1-\mu_2)^T\Sigma_2^{-1}(x_1-\mu_2)\right)=$

$=\frac{1}{2\pi\cdot\sqrt{0,4205}}\exp\left(-\frac{1}{2}\left(\begin{bmatrix}1\\2\end{bmatrix}-\begin{bmatrix}0,0344\\0,7770\end{bmatrix}\right)^T\cdot\begin{bmatrix}1,1118 & 0,5122\\0,5122 & 2,3753\end{bmatrix}\left(\begin{bmatrix}1\\2\end{bmatrix}-\begin{bmatrix}0,0344\\0,7770\end{bmatrix}\right)\right)=$

$=\frac{1}{2\pi\cdot0,6485}\exp\left(-\frac{1}{2}\times5,7992\right)=0,0135$

$P(x_1,c=1)=P(x_1|c=1)\cdot\pi_1=0,1957\cdot0,4171=0,0816$

$P(x_1,c=2)=P(x_1|c=2)\cdot\pi_2=0,0135\cdot0,5829=0,0079$

$\operatorname{argmax}(P(x_1,c=1);P(x_1,c=2))=P(x_1,c=1)\rightarrow x_1\in$ Cluster 1

x_2:

$P(x_2|c=1)=N(x_2|\mu_1,\Sigma_1)=\frac{1}{2\pi}\cdot\frac{1}{\sqrt{\det(\Sigma_1)}}\cdot\exp\left(-\frac{1}{2}(x_2-\mu_1)^T\Sigma_1^{-1}(x_2-\mu_1)\right)=$

$=\frac{1}{2\pi}\cdot\frac{1}{\sqrt{0,3334}}\cdot\exp\left(-\frac{1}{2}\left(\begin{bmatrix}-1\\1\end{bmatrix}-\begin{bmatrix}0,7510\\1,3112\end{bmatrix}\right)^T\begin{bmatrix}2,3347 & -0,2328\\-0,2328 & 1,3077\end{bmatrix}\left(\begin{bmatrix}-1\\1\end{bmatrix}-\begin{bmatrix}0,7510\\1,3112\end{bmatrix}\right)\right)=$

$=\frac{1}{2\pi\times0,5774}\exp\left(-\frac{1}{2}\times7,0311\right)=0,0082$

$P(x_2|c=2)=N(x_2|\mu_2,\Sigma_2)=\frac{1}{2\pi}\cdot\frac{1}{\sqrt{\det(\Sigma_2)}}\cdot\exp\left(-\frac{1}{2}(x_2-\mu_2)^T\Sigma_2^{-1}(x_2-\mu_2)\right)=$

$=\frac{1}{2\pi}\cdot\frac{1}{\sqrt{0,4205}}\exp\left(-\frac{1}{2}\left(\begin{bmatrix}-1\\1\end{bmatrix}-\begin{bmatrix}0,0344\\0,7770\end{bmatrix}\right)^T\begin{bmatrix}1,1118 & 0,5122\\0,5122 & 2,3753\end{bmatrix}\left(\begin{bmatrix}-1\\1\end{bmatrix}-\begin{bmatrix}0,0344\\0,7770\end{bmatrix}\right)\right)=$

$=\frac{1}{2\pi\cdot0,6485}\exp\left(-\frac{1}{2}\cdot1,0714\right)=0,1436$

$P(x_2,c=1)=P(x_2|c=1)\cdot\pi_1=0,0082\cdot0,4171=0,0034$

$P(x_2,c=2)=P(x_2|c=2)\cdot\pi_2=0,1436\cdot0,5829=0,0837$

$\operatorname{argmax}(P(x_2,c=1);P(x_2,c=2))=P(x_2,c=2)\rightarrow x_2\in$ Cluster 2

x_3:

$P(x_3|c=1)=N(x_3|\mu_1,\Sigma_1)=\frac{1}{2\pi}\cdot\cancel{\frac{1}{\sqrt{\det}}}\frac{1}{\det(\Sigma_1)^{1/2}}\cdot\exp\left(-\frac{1}{2}(x_3-\mu_1)^T\Sigma_1^{-1}(x_3-\mu_1)\right)=$

$=\frac{1}{2\pi}\cdot\frac{1}{\sqrt{0,3334}}\cdot\exp\left(-\frac{1}{2}\left(\begin{bmatrix}1\\0\end{bmatrix}-\begin{bmatrix}0,7510\\1,3112\end{bmatrix}\right)^T\begin{bmatrix}2,3347 & -0,2328\\-0,2328 & 1,3077\end{bmatrix}\left(\begin{bmatrix}1\\0\end{bmatrix}-\begin{bmatrix}0,7510\\1,3112\end{bmatrix}\right)\right)=$

$=\frac{1}{2\pi\cdot0,5774}\cdot\exp\left(-\frac{1}{2}\cdot2,545\right)=0,0772$

$P(x_3|c=2)=N(x_3|\mu_2,\Sigma_2)=\frac{1}{2\pi}\cdot\frac{1}{\sqrt{\det(\Sigma_2)}}\cdot\exp\left(-\frac{1}{2}(x_3-\mu_2)^T\Sigma_2^{-1}(x_3-\mu_2)\right)=$

$=\frac{1}{2\pi}\cdot\frac{1}{\sqrt{0,4205}}\cdot\exp\left(-\frac{1}{2}\left(\begin{bmatrix}1\\0\end{bmatrix}-\begin{bmatrix}0,0344\\0,7770\end{bmatrix}\right)^T\begin{bmatrix}1,1118 & 0,5122\\0,5122 & 2,3753\end{bmatrix}\left(\begin{bmatrix}1\\0\end{bmatrix}-\begin{bmatrix}0,0344\\0,7770\end{bmatrix}\right)\right)=$

$=\frac{1}{2\pi\cdot0,6485}\exp\left(-\frac{1}{2}\times1,7021\right)=0,1048$

$P(x_3,c=1)=P(x_3|c=1)\cdot\pi_1=0,0772\cdot0,4171=0,0322$

$P(x_3,c=2)=P(x_3|c=2)\cdot\pi_2=0,1048\cdot0,5829=0,06109$

$\operatorname{argmax}(P(x_3,c=1);P(x_3,c=2))=P(x_3,c=2)\rightarrow x_3\in$ ~~Bz~~ Cluster 2 b)

2b) O maior cluster é C_2, pois possui mais elementos.

$x_1=\begin{bmatrix}1\\2\end{bmatrix}$; $x_2=\begin{bmatrix}-1\\1\end{bmatrix}$; $x_3=\begin{bmatrix}1\\0\end{bmatrix}$; $\mu_2=\begin{bmatrix}0,0344\\0,7770\end{bmatrix}$

$x_1 \in C_1$; $x_2, x_3 \in C_2$

x_2:

$$d(x_2,x_3)=\sqrt{|x_{21}-x_{31}|^2+|x_{22}-x_{32}|^2}=\sqrt{|-1-1|^2+|1-0|^2}=\sqrt{5}$$

$$d(x_2,x_1)=\sqrt{|x_{21}-x_{11}|^2+|x_{22}-x_{12}|^2}=\sqrt{|-1-1|^2+|1-1|^2}=\sqrt{5}$$

$$s(x_2)=\frac{d(x_2,x_1)}{d(x_1,x_3)}-1=0$$

x_3:

$$d(x_2,x_3)=\sqrt{|x_{21}-x_{31})^2+|x_{22}-x_{32}|^2}=\sqrt{5}$$

$$d(x_3,x_1)=\sqrt{|x_{31}-x_{11}|^2+|x_{32}-x_{12}|^2}=\sqrt{|1-1|^2+|0-2|^2}=\sqrt{4}=2$$

$$s(x_3)=\frac{d(x_3,x_1)}{d(x_2,x_3)}-1=\frac{2}{\sqrt{5}}-1=-0,1056$$

$$s(C_2)=\frac{s(x_2)+s(x_3)}{2}=\frac{0-0,1056}{2}=-0,0528$$

II. Programming and critical analysis

3)

Silhouette score of Solution 1: 0.1136;

Silhouette score of Solution 2: 0.1140;

Silhouette score of Solution 3: 0.1136;

Purity of Solution 1: 0.7672;

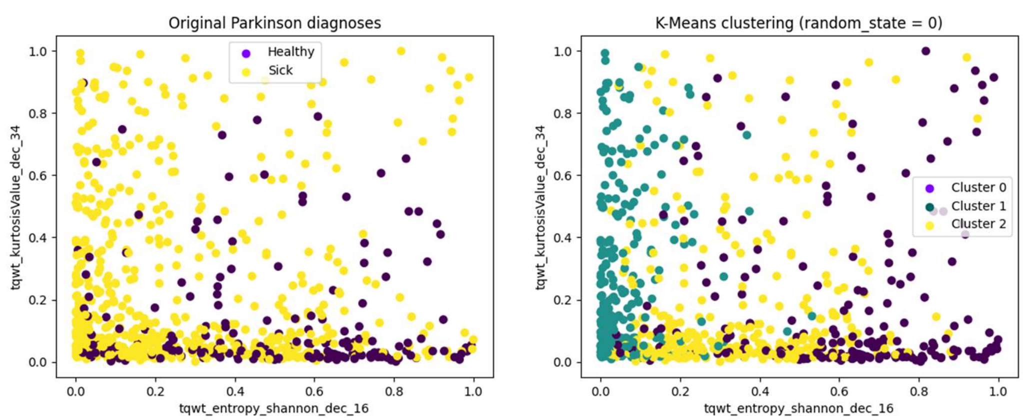
Purity of Solution 2: 0.7632;

Purity of Solution 3: 0.7672.

4)

A inicialização dos centróides é a principal causa de não determinismo presente neste método, daí a utilização de random_state=n, que tem como objetivo aplicar uma determinação transformação a n de forma a atribuir pseudo-aleatoriamente determinados valores como sendo os centros de cada cluster. Como o resto do algoritmo (k-means) pode ser considerado determinista, concluimos que esta é a única causa para o aparente não-determinismo.

5)



6)

Number of principal components: 31.

III. APPENDIX

```
#######         Importing required libraries          #######
import pandas as pd
from scipy.io.arff import loadarff
from sklearn.preprocessing import MinMaxScaler
from sklearn import metrics, cluster
import matplotlib.pyplot as plt
import numpy as np
from sklearn.decomposition import PCA


#######           Defining Purity Function            #######
def purity_score(y_true, y_pred):
    # compute contingency/confusion matrix
    confusion_matrix = metrics.cluster.contingency_matrix(y_true, y_pred)
    return round(np.sum(np.amax(confusion_matrix, axis=0)) /
np.sum(confusion_matrix), 4)


#######             Reading the ARFF file             #######
data = loadarff('pd_speech.arff')
df = pd.DataFrame(data[0])
df['class'] = df['class'].str.decode('utf-8')
X, y = df.drop('class', axis=1), df['class']


#######                  Normalizing                  #######
scaler = MinMaxScaler()
X_normalized = scaler.fit_transform(X)


#######                    K-Means                    #######
kmeans_algo0 = cluster.KMeans(n_clusters=3, random_state=0)
kmeans_algo1 = cluster.KMeans(n_clusters=3, random_state=1)
kmeans_algo2 = cluster.KMeans(n_clusters=3, random_state=2)

# learning the model
kmeans_model0 = kmeans_algo0.fit(X_normalized)
kmeans_model1 = kmeans_algo1.fit(X_normalized)
kmeans_model2 = kmeans_algo2.fit(X_normalized)

# getting the predicted labels
y_pred0 = kmeans_model0.labels_
y_pred1 = kmeans_model1.labels_
y_pred2 = kmeans_model2.labels_


#######                     Ex 1                      #######
print("Silhouette score of Solution 1:", round(metrics.silhouette_score(X_normalized,
y_pred0), 4))
print("Purity of Solution 1:", purity_score(y, y_pred0))
```

```python
print("Silhouette score of Solution 2:", round(metrics.silhouette_score(X_normalized,
y_pred1), 4))
print("Purity of Solution 2:", purity_score(y, y_pred1))
print("Silhouette score of Solution 3:", round(metrics.silhouette_score(X_normalized,
y_pred2), 4))
print("Purity of Solution 3:", purity_score(y, y_pred2))


#######                    Ex 3                     #######
# variance by feature of normalized data
variance = X_normalized.var(axis=0)
# get the two highest features based on variance
two_highest_variance = variance.argsort()[-2:][::-1]

# plotting
plt.figure(figsize=(14, 5))

plt.subplot(121)
y_values = np.array([int(i) for i in y.values])
plt.scatter(X_normalized[:, two_highest_variance[0]], X_normalized[:,
two_highest_variance[1]], c=y_values)
plt.legend(handles=[plt.scatter([], [], label='Healthy', c='#7F00FF'),
                    plt.scatter([], [], label='Sick', c='#FFF333')])
plt.title("Original Parkinson diagnoses")
plt.xlabel(X.columns[two_highest_variance[0]])
plt.ylabel(X.columns[two_highest_variance[1]])

plt.subplot(122)
plt.scatter(X_normalized[:, two_highest_variance[0]], X_normalized[:,
two_highest_variance[1]], c=y_pred0)
plt.legend(handles=[plt.scatter([], [], label='Cluster 0', c='#7F00FF'),
                    plt.scatter([], [], label='Cluster 1', c='#006666'),
                    plt.scatter([], [], label='Cluster 2', c='#FFF333')])
plt.title("K-Means clustering (random_state = 0)")
plt.xlabel(X.columns[two_highest_variance[0]])
plt.ylabel(X.columns[two_highest_variance[1]])

plt.show()


#######                    Ex 4                     #######
pca = PCA()
pca.fit(X_normalized)
i = 1
for var in np.cumsum(pca.explained_variance_ratio_):
    if var > 0.8:
        break
    i += 1
print("Number of principal components:", i)
```

END