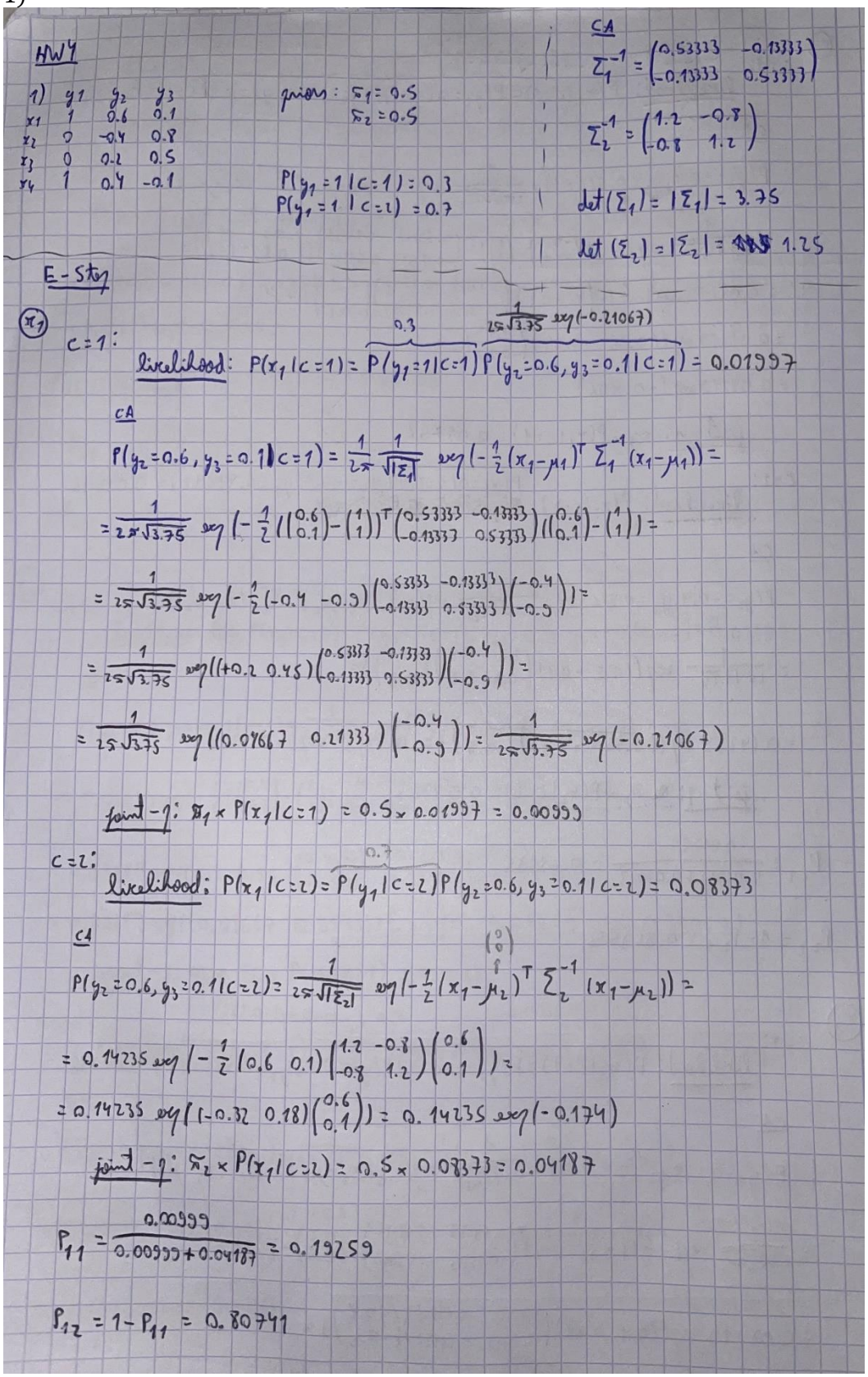
Aprendizagem 2022/23

Homework IV – Group 107

(ist1103811, ist1103479)

I. Pen-and-paper

1)



HW4

1)

	y_1	y_2	y_3
x_1	1	0.6	0.1
x_2	0	-0.4	0.8
x_3	0	0.2	0.5
x_4	1	0.4	-0.1

priors: $\pi_1 = 0.5$, $\pi_2 = 0.5$

$P(y_1 = 1 \mid C = 1) = 0.3$

$P(y_1 = 1 \mid C = 2) = 0.7$

CA

$\Sigma_1^{-1} = \begin{pmatrix} 0.53333 & -0.13333 \\ -0.13333 & 0.53333 \end{pmatrix}$

$\Sigma_2^{-1} = \begin{pmatrix} 1.2 & -0.8 \\ -0.8 & 1.2 \end{pmatrix}$

$\det(\Sigma_1) = |\Sigma_1| = 3.75$

$\det(\Sigma_2) = |\Sigma_2| = 1.25$

E-Step

(x_1)

$C = 1$:

likelihood: $P(x_1 \mid C=1) = \overbrace{P(y_1 = 1 \mid C=1)}^{0.3} \overbrace{P(y_2 = 0.6, y_3 = 0.1 \mid C=1)}^{\frac{1}{2\pi\sqrt{3.75}} \exp(-0.21067)} = 0.01997$

CA

$P(y_2 = 0.6, y_3 = 0.1 \mid C=1) = \frac{1}{2\pi} \frac{1}{\sqrt{|\Sigma_1|}} \exp\left(-\frac{1}{2}(x_1 - \mu_1)^T \Sigma_1^{-1} (x_1 - \mu_1)\right) =$

$= \frac{1}{2\pi\sqrt{3.75}} \exp\left(-\frac{1}{2}\left(\begin{pmatrix} 0.6 \\ 0.1 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right)^T \begin{pmatrix} 0.53333 & -0.13333 \\ -0.13333 & 0.53333 \end{pmatrix} \left(\begin{pmatrix} 0.6 \\ 0.1 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right)\right) =$

$= \frac{1}{2\pi\sqrt{3.75}} \exp\left(-\frac{1}{2}(-0.4 \;\; -0.9) \begin{pmatrix} 0.53333 & -0.13333 \\ -0.13333 & 0.53333 \end{pmatrix} \begin{pmatrix} -0.4 \\ -0.9 \end{pmatrix}\right) =$

$= \frac{1}{2\pi\sqrt{3.75}} \exp\left((+0.2 \;\; 0.45) \begin{pmatrix} 0.53333 & -0.13333 \\ -0.13333 & 0.53333 \end{pmatrix} \begin{pmatrix} -0.4 \\ -0.9 \end{pmatrix}\right) =$

$= \frac{1}{2\pi\sqrt{3.75}} \exp\left((0.07667 \;\; 0.21333) \begin{pmatrix} -0.4 \\ -0.9 \end{pmatrix}\right) = \frac{1}{2\pi\sqrt{3.75}} \exp(-0.21067)$

joint-1: $\pi_1 \times P(x_1 \mid C=1) = 0.5 \times 0.01997 = 0.00999$

$C = 2$:

likelihood: $P(x_1 \mid C=2) = \overbrace{P(y_1 \mid C=2)}^{0.7} P(y_2 = 0.6, y_3 = 0.1 \mid C=2) = 0.08373$

CA

$P(y_2 = 0.6, y_3 = 0.1 \mid C=2) = \frac{1}{2\pi\sqrt{|\Sigma_2|}} \exp\left(-\frac{1}{2}(x_1 - \overbrace{\mu_2}^{\begin{pmatrix} 0 \\ 0 \end{pmatrix}})^T \Sigma_2^{-1} (x_1 - \mu_2)\right) =$

$= 0.14235 \exp\left(-\frac{1}{2}(0.6 \;\; 0.1) \begin{pmatrix} 1.2 & -0.8 \\ -0.8 & 1.2 \end{pmatrix} \begin{pmatrix} 0.6 \\ 0.1 \end{pmatrix}\right) =$

$= 0.14235 \exp\left((-0.32 \;\; 0.18) \begin{pmatrix} 0.6 \\ 0.1 \end{pmatrix}\right) = 0.14235 \exp(-0.174)$

joint-1: $\pi_2 \times P(x_1 \mid C=2) = 0.5 \times 0.08373 = 0.04187$

$P_{11} = \frac{0.00999}{0.00999 + 0.04187} = 0.19259$

$P_{12} = 1 - P_{11} = 0.80741$

x_2

$c=1$:

likelihood: $P(x_2|c=1) = \underbrace{P(y_1=0|c=1)}_{1-0.3=0.7} \underbrace{P(y_2=-0.4, y_3=0.8|c=1)}_{0.08219\exp(+0.496)} = 0.03503$

CA

$$P(y_2=-0.4, y_3=0.8|c=1) = \frac{1}{2\pi\sqrt{|\Sigma_1|}} \exp\left(-\frac{1}{2}(x_2-\mu_1)^T \Sigma_1^{-1}(x_2-\mu_1)\right) =$$

$$= \frac{1}{2\pi\sqrt{3.75}} \exp\left(-\frac{1}{2}\left(\begin{pmatrix}-0.4\\0.8\end{pmatrix}-\begin{pmatrix}1\\1\end{pmatrix}\right)^T \begin{pmatrix}0.53333 & -0.13333\\-0.13333 & 0.53333\end{pmatrix}\left(\begin{pmatrix}-0.4\\0.8\end{pmatrix}-\begin{pmatrix}1\\1\end{pmatrix}\right)\right) =$$

$$= 0.08219 \exp\left(-\frac{1}{2}\begin{pmatrix}-1.4 & -0.2\end{pmatrix}\begin{pmatrix}0.53333 & -0.13333\\-0.13333 & 0.53333\end{pmatrix}\begin{pmatrix}-1.4\\-0.2\end{pmatrix}\right) =$$

$$= 0.08219 \exp\left(\begin{pmatrix}0.36 & -0.04\end{pmatrix}\begin{pmatrix}-1.4\\-0.2\end{pmatrix}\right) =$$

$$= 0.08219 \exp(-0.496)$$

joint-p: $\pi_1 \times P(x_2|c=1) = 0.01752$

$c=2$:

likelihood: $P(x_2|c=2) = \underbrace{P(y_1=0|c=2)}_{1-0.7=0.3} \underbrace{P(y_2=-0.4, y_3=0.8|c=2)}_{0.14235\exp(-0.736)} = 0.02046$

CA

$$P(y_2=-0.4, y_3=0.8|c=2) = \frac{1}{2\pi\sqrt{|\Sigma_2|}} \exp\left(-\frac{1}{2}(x_2-\mu_2)^T \Sigma_2^{-1}(x_2-\mu_2)\right) =$$

$$= \frac{1}{2\pi\sqrt{1.75}} \exp\left(\begin{pmatrix}0.2 & -0.4\end{pmatrix}\begin{pmatrix}1.2 & -0.8\\-0.8 & 1.2\end{pmatrix}\begin{pmatrix}-0.4\\0.8\end{pmatrix}\right) =$$

$$= 0.14235 \exp\left(\begin{pmatrix}0.56 & -0.64\end{pmatrix}\begin{pmatrix}-0.4\\0.8\end{pmatrix}\right) = 0.14235 \exp(-0.736)$$

joint-p: $\pi_2 \times P(x_2|c=2) = 0.01023$

$$P_{21} = \frac{0.01752}{0.01752+0.01023} = 0.63135$$

$$P_{22} = 1 - P_{21} = 0.36865$$

x_3

$c=1$:

likelihood: $P(x_3|c=1) = \underbrace{P(y_1=0|c=1)}_{0.7} \underbrace{P(y_2=0.2, y_3=0.5|c=1)}_{0.08219\exp(-0.184)} = 0.04786$

CA

$$P(y_2=0.2, y_3=0.5|c=1) = \frac{1}{2\pi\sqrt{|\Sigma_1|}} \exp\left(-\frac{1}{2}(x_3-\mu_1)^T \Sigma_1^{-1}(x_3-\mu_1)\right) =$$

$$= \frac{1}{2\pi\sqrt{3.75}} \exp\left(-\frac{1}{2}\left(\begin{pmatrix}0.2\\0.5\end{pmatrix}-\begin{pmatrix}1\\1\end{pmatrix}\right)^T \begin{pmatrix}0.53333 & -0.13333\\-0.13333 & 0.53333\end{pmatrix}\left(\begin{pmatrix}0.2\\0.5\end{pmatrix}-\begin{pmatrix}1\\1\end{pmatrix}\right)\right) =$$

$$= 0.08219 \exp\left(\begin{pmatrix}0.4 & 0.25\end{pmatrix}\begin{pmatrix}0.53333 & -0.13333\\-0.13333 & 0.53333\end{pmatrix}\begin{pmatrix}-0.8\\-0.5\end{pmatrix}\right) =$$

$$= 0.08219 \exp\left(\begin{pmatrix}0.18 & 0.08\end{pmatrix}\begin{pmatrix}-0.8\\-0.5\end{pmatrix}\right) = 0.08219 \exp(-0.184)$$

joint-η: $\pi_1 \times P(x_3|c_1) = 0.5 \times 0.04786 = 0.02393$

$c=2$:

likelihood: $P(x_3|c=2) = \overbrace{P(y_1=0|c=2)}^{0.3} \overbrace{P(y_2=0.2, y_3=0.5|c=2)}^{0.14235 \exp(-0.094)} = 0.03887$

c1

$$P(y_2=0.2, y_3=0.5|c=2) = \frac{1}{2\pi\sqrt{1.25}} \exp\left(-\frac{1}{2}(x_3-\mu_2)^T \Sigma_2^{-1} (x_3-\mu_2)\right) =$$

$$= 0.14235 \exp\left((-0.1 \ -0.25)\begin{pmatrix}1.2 & -0.8\\ -0.8 & 1.2\end{pmatrix}\begin{pmatrix}0.2\\ 0.5\end{pmatrix}\right) = 0.14235 \exp\left((0.08 \ -0.22)\begin{pmatrix}0.2\\ 0.5\end{pmatrix}\right) =$$

$$= 0.14235 \exp(-0.094)$$

joint-η: $\pi_2 \times P(x_3|c=2) = 0.01944$

$$P_{31} = \frac{0.02393}{0.02393+0.01944} = 0.55181$$

$$P_{32} = 1 - P_{31} = 0.44819$$

(x_4)

$c=1$:

likelihood: $P(x_4|c=1) = \overbrace{P(y_1=1|c=1)}^{0.3} \overbrace{P(y_2=0.4, y_3=-0.1|c=1)}^{0.08219 \exp(-0.33067)} = 0.01771$

c1

$$P(y_2=0.4, y_3=-0.1|c=1) = \frac{1}{2\pi\sqrt{3.75}} \exp\left(-\frac{1}{2}(x_4-\mu_1)^T \Sigma_1^{-1} (x_4-\mu_1)\right) =$$

$$= 0.08219 \exp\left(-\frac{1}{2}\left(\begin{pmatrix}0.4\\ -0.1\end{pmatrix}-\begin{pmatrix}1\\ 1\end{pmatrix}\right)^T \begin{pmatrix}0.53333 & -0.13333\\ -0.13333 & 0.53333\end{pmatrix}\left(\begin{pmatrix}0.4\\ -0.1\end{pmatrix}-\begin{pmatrix}1\\ 1\end{pmatrix}\right)\right) =$$

$$= 0.08219 \exp\left((0.3 \ \ 0.55)\begin{pmatrix}0.53333 & -0.13333\\ -0.13333 & 0.53333\end{pmatrix}\begin{pmatrix}-0.6\\ -1.1\end{pmatrix}\right) =$$

$$= 0.08219 \exp\left((0.08667 \ \ 0.25333)\begin{pmatrix}-0.6\\ -1.1\end{pmatrix}\right) = 0.08219 \exp(-0.33067)$$

joint-η: $\pi_1 \times P(x_4|c=1) = 0.00886$

$c=2$:

likelihood: $P(x_4|c=2) = \overbrace{P(y_1=1|c=2)}^{0.7} \overbrace{P(y_2=0.4, y_3=-0.1|c=2)}^{0.14235 \exp(-0.134)} = 0.08715$

c1

$$P(y_2=0.4, y_3=-0.1|c=1) = \frac{1}{2\pi\sqrt{1.25}} \exp\left(-\frac{1}{2}(x_4-\mu_2)^T \Sigma_2^{-1} (x_4-\mu_2)\right) =$$

$$0.14235 \exp\left((-0.2 \ \ 0.05)\begin{pmatrix}1.2 & -0.8\\ -0.8 & 1.2\end{pmatrix}\begin{pmatrix}0.4\\ -0.1\end{pmatrix}\right) = 0.14235 \exp\left((-0.28 \ \ 0.22)\begin{pmatrix}0.4\\ -0.1\end{pmatrix}\right) =$$

$$= 0.14235 \exp(-0.134)$$

joint-η: $\pi_2 \times P(x_4|c=2) = 0.04358$

Homework IV – Group 107

(ist1103811, ist1103479)

$$P_{41} = \frac{0.00886}{0.00886+0.04358} = 0.16892$$

$$P_{42} = 1 - P_{41} = 0.83108$$

	x_1	x_2	x_3	x_4
c=1	0.19259	0.63135	0.55181	0.16892
c=2	0.30741	0.36865	0.44819	0.83108

CA

$$A = 0.19259 + 0.63135 + 0.55181 + 0.16892 = 1.54467$$

$$B = 0.30741 + 0.36865 + 0.44819 + 0.83108 = 2.45533$$

M-Step

$$P(c=1) = \frac{A}{A+B} = 0.38617$$

$$P(c=2) = \frac{B}{A+B} = 0.61383$$

$$P(y_1=1|c=1) = \frac{(0.19259\times 1)+(0.63135\times 0)+(0.55181\times 0)+(0.16892\times 1)}{A} = 0.23404$$

$$P(y_1=1|c=2) = \frac{(0.30741\times 1)+(0.36865\times 0)+(0.44819\times 0)+(0.83108\times 1)}{B} = 0.66732$$

$$\mu_1 = \frac{(0.19259\times x_1)+(0.63135 x_2)+(0.55181 x_3)+(0.16892 x_4)}{A} =$$

($x_1 = \begin{pmatrix}0.6\\0.1\end{pmatrix}$, $x_2 = \begin{pmatrix}-0.4\\0.3\end{pmatrix}$, $x_3 = \begin{pmatrix}0.2\\0.5\end{pmatrix}$, $x_4 = \begin{pmatrix}0.4\\-0.1\end{pmatrix}$)

$$= \frac{\begin{pmatrix}0.11555\\0.01926\end{pmatrix}+\begin{pmatrix}-0.25262\\0.50524\end{pmatrix}+\begin{pmatrix}0.11036\\0.27591\end{pmatrix}+\begin{pmatrix}0.06757\\-0.01689\end{pmatrix}}{1.54467} = \frac{\begin{pmatrix}0.04086\\0.78352\end{pmatrix}}{1.54467} = \begin{pmatrix}0.02651\\0.50713\end{pmatrix}$$

$$\mu_2 = \frac{(0.30741 x_1)+(0.36865 x_2)+(0.44819 x_3)+(0.83108 x_4)}{B} =$$

$$= \frac{\begin{pmatrix}0.48445\\0.03074\end{pmatrix}+\begin{pmatrix}-0.14746\\0.29492\end{pmatrix}+\begin{pmatrix}0.08964\\0.2241\end{pmatrix}+\begin{pmatrix}0.33243\\-0.08311\end{pmatrix}}{2.45533} = \frac{\begin{pmatrix}0.75906\\0.51665\end{pmatrix}}{2.45533} = \begin{pmatrix}0.30914\\0.21042\end{pmatrix}$$

Homework IV – Group 107

(ist1103811, ist1103479)

$$\Sigma_1 = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} 0.14136 & -0.1054 \\ -0.1054 & 0.09605 \end{pmatrix}$$

CA

$$\Sigma_{11} = \frac{0.19259(x_{11}-\mu_{11})^2 + 0.63135(x_{21}-\mu_{11})^2 + 0.55181(x_{31}-\mu_{11})^2 + 0.16892(x_{41}-\mu_{11})^2}{A} =$$

$$= \frac{0.06334 + 0.11485 + 0.01861 + 0.02786}{1.54467} = 0.14136$$

$$\Sigma_{22} = \frac{0.19259(x_{12}-\mu_{12})^2 + 0.63135(x_{22}-\mu_{12})^2 + 0.55181(x_{32}-\mu_{12})^2 + 0.16892(x_{42}-\mu_{12})^2}{A} =$$

$$= \frac{0.03192 + 0.05415 + 2.8\times10^{-5} + 0.06227}{1.54467} = 0.09605$$

$$\Sigma_{12} = \frac{0.19259(x_{11}-\mu_{11})(x_{12}-\mu_{12}) + 0.63135(x_{21}-\mu_{11})(x_{22}-\mu_{12}) + 0.55181(x_{31}-\mu_{11})(x_{32}-\mu_{12}) + 0.16892(x_{41}-\mu_{11})(x_{42}-\mu_{12})}{A} =$$

$$= \frac{-0.04497 - 0.07886 - 0.00068 - 0.0383}{1.54467} = -0.1054$$

$$\Sigma_2 = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} 0.108295 & -0.08865 \\ -0.08865 & 0.10412 \end{pmatrix}$$

CA

$$\Sigma_{11} = \frac{0.80741(x_{11}-\mu_{21})^2 + 0.36865(x_{21}-\mu_{21})^2 + 0.44819(x_{31}-\mu_{21})^2 + 0.83108(x_{41}-\mu_{21})^2}{B} =$$

$$= \frac{0.06831 + 0.17589 + 0.00534 + 0.00681}{2.45533} = 0.108295$$

$$\Sigma_{22} = \frac{0.80741(x_{12}-\mu_{22})^2 + 0.36865(x_{22}-\mu_{22})^2 + 0.44819(x_{32}-\mu_{22})^2 + 0.83108(x_{42}-\mu_{22})^2}{B} =$$

$$= \frac{0.00984 + 0.12814 + 0.03758 + 0.08008}{2.45533} = 0.10412$$

$$\Sigma_{12} = \frac{0.80741(x_{11}-\mu_{21})(x_{12}-\mu_{22}) + 0.36865(x_{21}-\mu_{21})(x_{22}-\mu_{22}) + 0.44819(x_{31}-\mu_{21})(x_{32}-\mu_{22}) + 0.83108(x_{41}-\mu_{21})(x_{42}-\mu_{22})}{B} =$$

$$= \frac{-0.02593 - 0.15413 - 0.01416 - 0.02744}{2.45533} = -0.08865$$

2)

2) $x_{new} = \begin{pmatrix} 1 \\ 0.3 \\ 0.7 \end{pmatrix}$

c=1:

likelihood: $P(x_{new}|c=1) = \overbrace{P(y_1=1|c=1)}^{0.23404} \overbrace{P(y_2=0.3, y_3=0.7|c=1)}^{0.02706} = 0.00633$

cA

$P(y_2=0.3, y_3=0.7|c=1) = \frac{1}{2\pi\sqrt{|\Sigma_1|}} \exp\left(-\frac{1}{2}(x_{new}-\mu_1)^T \Sigma_1^{-1} (x_{new}-\mu_1)\right) =$

$= \frac{1}{2\pi\sqrt{|\Sigma_1|}} \exp\left(-\frac{1}{2}\left(\begin{pmatrix} 0.3 \\ 0.7 \end{pmatrix} - \begin{pmatrix} 0.02651 \\ 0.50713 \end{pmatrix}\right)^T \Sigma_1^{-1} (x_{new}-\mu_1)\right) =$

$= \frac{1}{2\pi\sqrt{|\Sigma_1|}} \exp\left(-\frac{1}{2}(0.27349 \quad 0.19287)\, \Sigma_1^{-1} (x_{new}-\mu_1)\right) =$

$= \frac{1}{2\pi\sqrt{|\Sigma_1|}} \exp\left((-0.13675 \quad -0.09644) \begin{pmatrix} 38.91659 & 42.70599 \\ 42.70599 & 57.27533 \end{pmatrix} (x_{new}-\mu_1)\right) =$

$= \frac{1}{2\pi\sqrt{0.00247}} \exp\left((-9.44041 \quad -11.36368) \begin{pmatrix} 0.27349 \\ 0.19287 \end{pmatrix}\right) =$

$= 3.20237 \exp(-4.77357) = 0.02706$

CA

$|\Sigma_1| = 0.00247$

$|\Sigma_2| = 0.00342$

$\Sigma_1^{-1} = \begin{pmatrix} 38.91659 & 42.70599 \\ 42.70599 & 57.27533 \end{pmatrix}$

$\Sigma_2^{-1} = \begin{pmatrix} 30.4748 & 25.94662 \\ 25.94662 & 31.69525 \end{pmatrix}$

joint-p: $P(c=1)P(x_{new}|c=1) = 0.38617 \times 0.00633 = 0.00245$

c=2:

likelihood: $P(x_{new}|c=2) = \overbrace{P(y_1=1|c=2)}^{0.66732} \overbrace{P(y_2=0.3, y_3=0.7|c=2)}^{0.06839} = 0.04564$

cA

$P(y_2=0.3, y_3=0.7|c=2) = \frac{1}{2\pi\sqrt{|\Sigma_2|}} \exp\left(-\frac{1}{2}(x_{new}-\mu_2)^T \Sigma_2^{-1} (x_{new}-\mu_2)\right) =$

$= \frac{1}{2\pi\sqrt{0.00342}} \exp\left(-\frac{1}{2}\left(\begin{pmatrix} 0.3 \\ 0.7 \end{pmatrix} - \begin{pmatrix} 0.30914 \\ 0.21042 \end{pmatrix}\right)^T \Sigma_2^{-1} (x_{new}-\mu_2)\right) =$

$= 2.72149 \exp\left(-\frac{1}{2}(-0.00914 \quad 0.48958) \begin{pmatrix} 30.4748 & 25.94662 \\ 25.94662 & 31.69525 \end{pmatrix} \begin{pmatrix} -0.00914 \\ 0.48958 \end{pmatrix}\right) =$

$= 2.72149 \exp\left((-6.2122 \quad -7.6401) \begin{pmatrix} -0.00914 \\ 0.48958 \end{pmatrix}\right) =$

$= 2.72149 \exp(-3.68303) = 0.06839$

joint-p: $P(c=2)P(x_{new}|c=2) = 0.61383 \times 0.04564 = 0.02802$

$P_1 = \frac{0.00245}{0.00245 + 0.02802} = 0.08041$

$P_2 = 1 - P_1 = 0.91959$

3)

3)

x_1

$C=1$:

$$\text{likelihood}: P(x_1|C=1) = \overbrace{P(y_1=1|C=1)}^{0.23404}\overbrace{P(y_2=0.6, y_3=0.1|C=1)}^{3.20237\exp(-1.17545)} = 0.23135$$

CA

$$P(y_2=0.6, y_3=0.1|C=1) = \frac{1}{2\pi\sqrt{|\Sigma_1|}}\exp\left(-\frac{1}{2}(x_1-\mu_1)^T\Sigma_1^{-1}(x_1-\mu_1)\right) =$$

$$= \frac{1}{2\pi\sqrt{0.00247}}\exp\left(-\frac{1}{2}\left(\begin{pmatrix}0.6\\0.1\end{pmatrix}-\begin{pmatrix}0.02651\\0.50713\end{pmatrix}\right)^T\Sigma_1^{-1}(x_1-\mu_1)\right) =$$

$$= 3.20237\exp\left(\begin{pmatrix}-0.28675 & 0.20357\end{pmatrix}\begin{pmatrix}38.91659 & 42.70599\\42.70599 & 57.27533\end{pmatrix}\begin{pmatrix}0.57349\\-0.40713\end{pmatrix}\right) =$$

$$= 3.20237\exp\left(\begin{pmatrix}-2.46684 & -0.58768\end{pmatrix}\begin{pmatrix}0.57349\\-0.40713\end{pmatrix}\right) = 3.20237\exp(-1.17545)$$

$C=2$:

$$\text{prior}: P(C=2) = 0.61383$$

$$\text{likelihood}: P(x_1|C=2) = \overbrace{P(y_1=1|C=2)}^{0.66732}\overbrace{P(y_2=0.6, y_3=0.1|C=2)}^{2.72149\exp(-0.64898)} = \boxed{0.94906}$$

CA

$$P(y_2=0.6, y_3=0.1|C=2) = \frac{1}{2\pi\sqrt{|\Sigma_2|}}\exp\left(-\frac{1}{2}(x_1-\mu_2)^T\Sigma_2^{-1}(x_1-\mu_2)\right) =$$

$$= 2.72149\exp\left(-\frac{1}{2}\begin{pmatrix}0.29086 & -0.11042\end{pmatrix}\begin{pmatrix}30.4748 & 25.94662\\25.94662 & 37.69525\end{pmatrix}\begin{pmatrix}0.29086\\-0.11042\end{pmatrix}\right) =$$

$$= 2.72149\exp\left(\begin{pmatrix}-2.99944 & -2.02352\end{pmatrix}\begin{pmatrix}0.29086\\-0.11042\end{pmatrix}\right) = 2.72149\exp(-0.64898)$$

x_2

$C=1$:

$$\text{likelihood}: P(x_2|C=1) = \overbrace{P(y_1=0|C=1)}^{1-0.23404}P(y_2=-0.4, y_3=0.8|C=1) = \boxed{1.11719}$$

CA

$$P(y_2=-0.4, y_3=0.8|C=1) = 3.20237\exp\left(-\frac{1}{2}\left(\begin{pmatrix}-0.4\\0.8\end{pmatrix}-\begin{pmatrix}0.02651\\0.50713\end{pmatrix}\right)^T\Sigma_1^{-1}(x_2-\mu_1)\right) =$$

$$= 3.20237\exp\left(\begin{pmatrix}0.21326 & -0.14644\end{pmatrix}\begin{pmatrix}38.91659 & 42.70599\\40.70599 & 57.27533\end{pmatrix}\begin{pmatrix}-0.42651\\0.29287\end{pmatrix}\right) =$$

$$= 3.20237\exp\left(\begin{pmatrix}2.33837 & 0.72008\end{pmatrix}\begin{pmatrix}-0.42651\\0.29287\end{pmatrix}\right) = 3.20237\exp(-0.78645)$$

$c=2$:

likelihood: $P(x_2|c=2) = \overbrace{1-0.66732}\, P(y_1=0|c=2)\, P(y_2=-0.4, y_3=0.8|c=2) = 0.088696$

CA

$$P(y_2=-0.4, y_3=0.8|c=2) = \frac{1}{2\pi\sqrt{|\Sigma_2|}} \exp\left(-\tfrac{1}{2}(x_2-\mu_2)^T \Sigma_2^{-1} (x_2-\mu_2)\right) =$$

$$= 2.72149 \exp\left(-\tfrac{1}{2}\left(\begin{pmatrix}-0.4\\0.8\end{pmatrix}-\begin{pmatrix}0.30914\\0.21042\end{pmatrix}\right)^T \Sigma_2^{-1} (x_2-\mu_2)\right) =$$

$$= 2.72149 \exp\left((0.35457 \quad -0.29479)\begin{pmatrix}30.4748 & 25.94662\\25.94662 & 31.69525\end{pmatrix}\begin{pmatrix}-0.70914\\0.58958\end{pmatrix}\right) =$$

$$= 2.72149 \exp\left((3.15665 \quad -0.14355)\begin{pmatrix}-0.70914\\0.58958\end{pmatrix}\right) = 2.72149 \exp(-2.32314)$$

x_3

$c=1$:

likelihood: $P(x_3|c=1) = \overbrace{1-0.23404}\, P(y_1=0|c=1)\, P(y_2=0.2, y_3=0.5|c=1) = 1.43404$

CA

$$P(y_2=0.2, y_3=0.5|c=1) = 3.20237 \exp\left(-\tfrac{1}{2}\left(\begin{pmatrix}0.2\\0.5\end{pmatrix}-\begin{pmatrix}0.02651\\0.50713\end{pmatrix}\right)^T \Sigma_1^{-1} (x_3-\mu_1)\right) =$$

$$= 3.20237 \exp\left((-0.0875 \quad 0.00357)\begin{pmatrix}38.91659 & 40.70599\\40.70599 & 57.27533\end{pmatrix}(x_3-\mu_1)\right) =$$

$$= 3.20237 \exp\left((-3.23069 \quad -3.32677)\begin{pmatrix}0.17349\\-0.00713\end{pmatrix}\right) =$$

$$= 3.20237 \exp(-0.53677)$$

$c=2$:

likelihood: $P(x_3|c=2) = (1-0.66732) \times 2.72149 \exp(-0.69039) = 0.45394$

CA

$$P(y_2=0.2, y_3=0.5|c=2) = 2.72149 \exp\left(-\tfrac{1}{2}\left(\begin{pmatrix}0.2\\0.5\end{pmatrix}-\begin{pmatrix}0.30914\\0.21042\end{pmatrix}\right)^T \Sigma_2^{-1} (x_3-\mu_2)\right) =$$

$$= 2.72149 \exp\left((0.05457 \quad -0.14479)\begin{pmatrix}30.4748 & 25.94662\\25.94662 & 31.69525\end{pmatrix}\begin{pmatrix}-0.10914\\0.28958\end{pmatrix}\right) =$$

$$= 2.72149 \exp\left((-2.0938 \quad -3.17325)\begin{pmatrix}-0.10914\\0.28958\end{pmatrix}\right) = 2.72149 \exp(-0.69039)$$

x_4

$c=1$:

likelihood: $P(x_4|c=1) = 0.23404 \times 3.20237 \exp(-3.58653) = 0.02076$

CA

$$P(y_2=0.4, y_3=-0.1|c=1) = 3.20237 \exp\left(-\tfrac{1}{2}\left(\begin{pmatrix}0.4\\-0.1\end{pmatrix}-\begin{pmatrix}0.02651\\0.50713\end{pmatrix}\right)^T \Sigma_1^{-1} (x_4-\mu_1)\right) =$$

$$= 3.20237 \exp\left((-0.18675 \quad 0.30357)\begin{pmatrix}38.91659 & 42.70599\\42.70599 & 57.27533\end{pmatrix}(x_4-\mu_1)\right) =$$

$$= 3.20237 \exp\left((5.69658 \quad 9.41173)\begin{pmatrix}0.37349\\-0.60713\end{pmatrix}\right) = 3.20237 \exp(-3.58653)$$

Homework IV – Group 107

(ist1103811, ist1103479)

$c=2$:

likelihood: $P(x_4|c=2) = 0.66732 \times 2.72149 \exp(-0.92106) = 0.72299$

C1

$$P(y_2=0.4, y_3=-0.1|c=2) = 2.72149 \exp\left(-\frac{1}{2}\left(\begin{pmatrix}0.4\\-0.1\end{pmatrix}-\begin{pmatrix}0.30914\\0.21042\end{pmatrix}\right)^T \Sigma_2^{-1}(x_4-\mu_2)\right) =$$

$$= 2.72149 \exp\left((-0.04543 \quad 0.15521)\begin{pmatrix}30.4748 & 25.94662\\25.94662 & 31.69525\end{pmatrix}(x_4-\mu_2)\right) =$$

$$= 2.72149 \exp\left((2.6427 \quad 3.74066)\begin{pmatrix}0.09086\\-0.21042\end{pmatrix}\right) = 2.72149 \exp(-0.92106)$$

Assignments: $x_1 \to c_2$, $x_2 \to c_1$, $x_3 \to c_1$, $x_4 \to c_2$

$$S(x_1) = 1 - \frac{\|x_1-x_4\|_1}{\frac{1}{2}(\|x_1-x_2\|_1+\|x_1-x_3\|_1)} = 1-\frac{0.4}{\frac{1}{2}(2.7+1.8)} = 1-\frac{0.4}{2.25} = 0.82222$$

$$S(x_2) = 1 - \frac{\|x_2-x_3\|_1}{\frac{1}{2}(\|x_2-x_1\|_1+\|x_2-x_4\|_1)} = 1-\frac{0.9}{\frac{1}{2}(2.7+2.7)} = 1-\frac{0.9}{2.7} = 0.66667$$

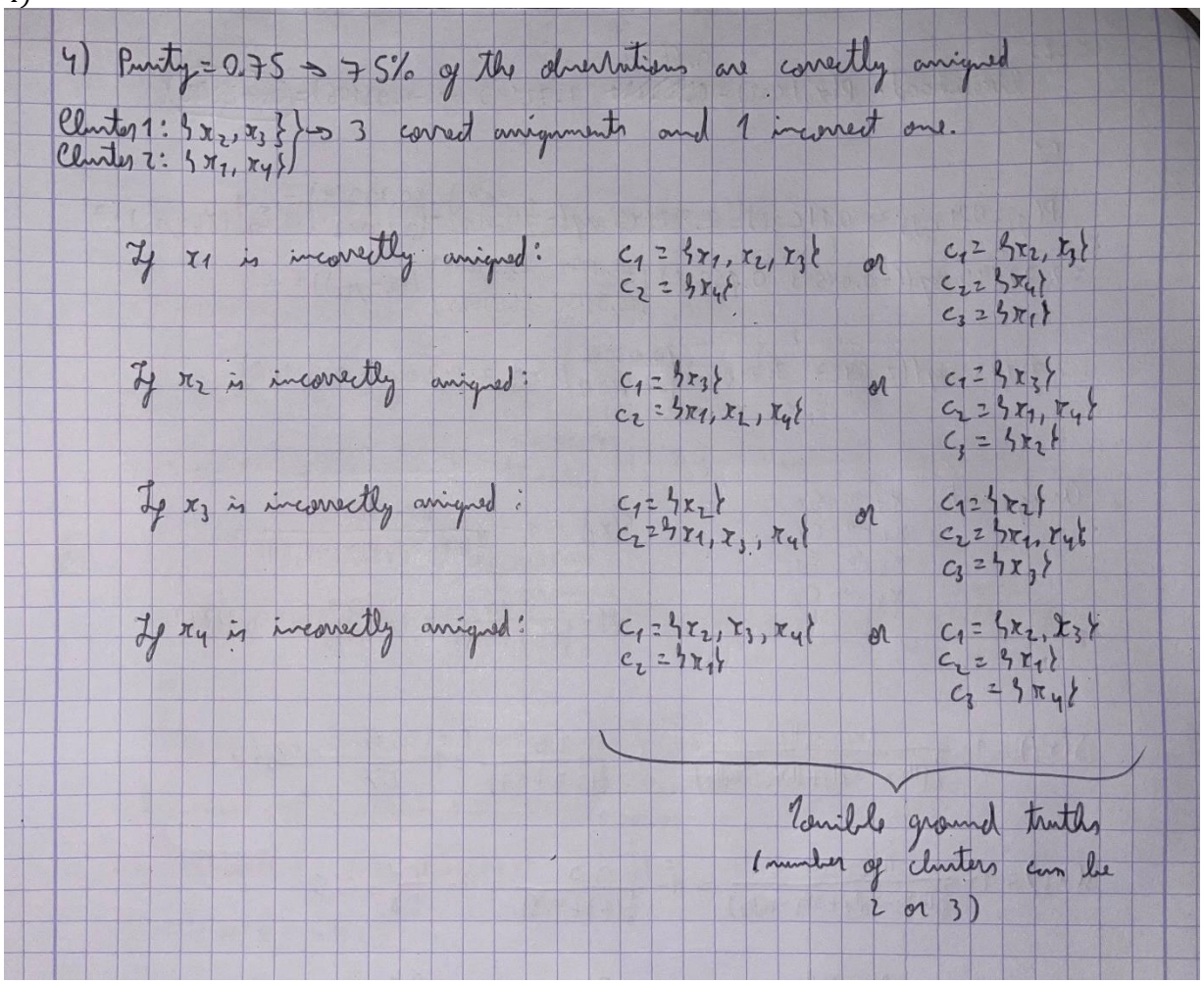
$$S(x_3) = 1 - \frac{\|x_3-x_2\|_1}{\frac{1}{2}(\|x_3-x_1\|_1+\|x_3-x_4\|_1)} = 1-\frac{0.9}{\frac{1}{2}(1.8+1.8)} = 1-\frac{0.9}{1.8} = 0.5$$

$$S(x_4) = 1 - \frac{\|x_4-x_1\|_1}{\frac{1}{2}(\|x_4-x_2\|_1+\|x_4-x_3\|_1)} = 1-\frac{0.4}{\frac{1}{2}(2.7+1.8)} = 1-\frac{0.4}{2.25} = 0.82222$$

$$S(c_1) = \frac{S(x_2)+S(x_3)}{2} = \frac{0.66667+0.5}{2} = 0.58333$$

$$S(c_2) = \frac{S(x_1)+S(x_4)}{2} = 0.82222$$

4)


4) Purity = 0.75 → 75% of the observations are correctly assigned
Cluster 1: {x2, x3} → 3 correct assignments and 1 incorrect one.
Cluster 2: {x1, x4}
If x1 is incorrectly assigned: C1 = {x1, x2, x3} C2 = {x4} or C1 = {x2, x3} C2 = {x4} C3 = {x1}
If x2 is incorrectly assigned: C1 = {x3} C2 = {x1, x2, x4} or C1 = {x3} C2 = {x1, x4} C3 = {x2}
If x3 is incorrectly assigned: C1 = {x2} C2 = {x1, x3, x4} or C1 = {x2} C2 = {x1, x4} C3 = {x3}
If x4 is incorrectly assigned: C1 = {x2, x3, x4} C2 = {x1} or C1 = {x2, x3} C2 = {x1} C3 = {x4}
7 possible ground truths
(number of clusters can be 2 or 3)

TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 107

(ist1103811, ist1103479)

II. Programming and critical analysis

1)

```
from scipy.io import arff
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import pairwise_distances_argmin_min

# Load the ARFF dataset
data, meta = arff.loadarff("column_diagnosis.arff")
df = pd.DataFrame(data)

# Convert the "class" column to numeric using LabelEncoder
label_encoder = LabelEncoder()
df['class'] = label_encoder.fit_transform(df['class'])

# Normalize the data using MinMaxScaler
scaler = MinMaxScaler()
normalized_data = scaler.fit_transform(df.drop(columns=['class']))

# Values of k to be tested
k_values = [2, 3, 4, 5]

# Initialize lists to store results
silhouette_scores = []
purity_scores = []

for k in k_values:
    # Apply K-means clustering
    kmeans = KMeans(n_clusters=k, random_state=0)
    cluster_labels = kmeans.fit_predict(normalized_data)

    # Compute silhouette score
    silhouette = silhouette_score(normalized_data, cluster_labels)
    # Calculate purity
    closest_to_centers = pairwise_distances_argmin_min(kmeans.cluster_centers_,
                                                       normalized_data)[0]
    cluster_purity = 0
    for i in range(k):
        cluster_mask = (cluster_labels == i)
        majority_class = df['class'][cluster_mask].value_counts().max()
        cluster_purity += majority_class
```

```
    cluster_purity /= len(cluster_labels)

    silhouette_scores.append(silhouette)
    purity_scores.append(cluster_purity)

# Print the results
for k, silhouette, purity in zip(k_values, silhouette_scores, purity_scores):
    print(f'K={k}: Silhouette Score = {silhouette:.15f}, Purity = {purity:.15f}')
```

K=2: Silhouette Score = 0.360441243404411, Purity = 0.632258064516129
K=3: Silhouette Score = 0.295790557300023, Purity = 0.667741935483871
K=4: Silhouette Score = 0.274424021223402, Purity = 0.661290322580645
K=5: Silhouette Score = 0.238239283978448, Purity = 0.677419354838710

2)

```
#Ex2
from scipy.io import arff
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
from sklearn.decomposition import PCA

# Load the ARFF dataset
data, meta = arff.loadarff("column_diagnosis.arff")
df = pd.DataFrame(data)

# Convert the "class" column to numeric using LabelEncoder
df['class'] = df['class'].str.decode('utf-8')

# Normalize the data using MinMaxScaler
scaler = MinMaxScaler()
normalized_data = scaler.fit_transform(df.drop(columns=['class']))

# Perform PCA to extract the top two principal components
pca = PCA(n_components=2)
principal_components = pca.fit_transform(normalized_data)

# i. Explained variance by the top two principal components
explained_variance = pca.explained_variance_ratio_
print(f"Explained Variance by the Top Two Principal Components:
{explained_variance}")
```

Homework IV – Group 107

(ist1103811, ist1103479)

```
# ii. Sort input variables by relevance for each of the top two components
component_1_loadings = pca.components_[0]
component_2_loadings = pca.components_[1]

# Sort variables by absolute loadings for component 1
sorted_vars_component_1 = df.columns[abs(component_1_loadings).argsort()[::-1]]
sorted_loadings_component_1 =
component_1_loadings[abs(component_1_loadings).argsort()[::-1]]

# Sort variables by absolute loadings for component 2
sorted_vars_component_2 = df.columns[abs(component_2_loadings).argsort()[::-1]]
sorted_loadings_component_2 =
component_2_loadings[abs(component_2_loadings).argsort()[::-1]]

# Print the sorted variables and their loadings for the top two components
print("\nTop Two Principal Components - Sorted Variables and Loadings:")
print("Principal Component 1:")
for var, loading in zip(sorted_vars_component_1, sorted_loadings_component_1):
    print(f"{var}: {loading:.4f}")

print("\nPrincipal Component 2:")
for var, loading in zip(sorted_vars_component_2, sorted_loadings_component_2):
    print(f"{var}: {loading:.4f}")
```

Top Two Principal Components - Sorted Variables and Loadings:
Principal Component 1:
pelvic_incidence: 0.5916
lumbar_lordosis_angle: 0.5151
pelvic_tilt: 0.4670
sacral_slope: 0.3257
degree_spondylolisthesis: 0.2169
pelvic_radius: -0.1158

Principal Component 2:
pelvic_tilt: -0.6704
pelvic_radius: -0.5811
sacral_slope: 0.4433
pelvic_incidence: 0.1000
lumbar_lordosis_angle: 0.0800
degree_spondylolisthesis: 0.0046

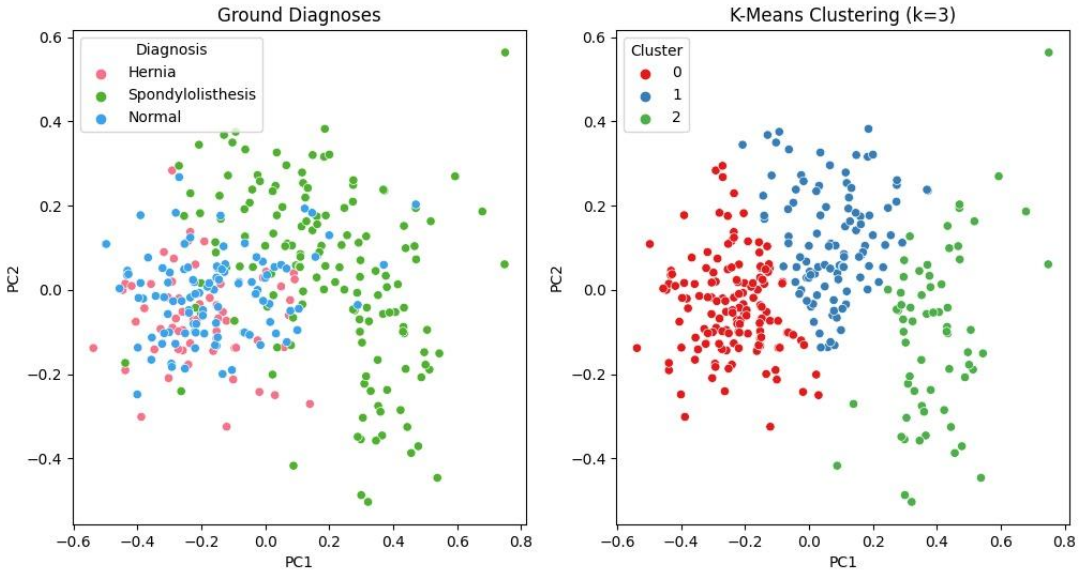
3)

```
#Ex3
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
import seaborn as sns
from scipy.io import arff

# Load the ARFF file
data, meta = arff.loadarff('column_diagnosis.arff')
df = pd.DataFrame(data)

# Normalize the data using MinMaxScaler
scaler = MinMaxScaler()
normalized_data = scaler.fit_transform(df.iloc[:, :-1])

# Perform PCA to reduce the data to 2 dimensions
pca = PCA(n_components=2)
pca_result = pca.fit_transform(normalized_data)

# Create a new DataFrame with the PCA results
pca_df = pd.DataFrame(data=pca_result, columns=['PC1', 'PC2'])

# Add the diagnosis labels to the PCA DataFrame
pca_df['Diagnosis'] = df['class'].str.decode('utf-8')  # Decode bytes to strings

# Create a palette of colors based on unique values in the 'Diagnosis' column
value_palette = sns.color_palette("husl", len(pca_df['Diagnosis'].unique()))

# Map colors to values in the 'Diagnosis' column
value_colors = {value: color for value, color in
zip(pca_df['Diagnosis'].unique(), value_palette)}

# Visualize the data side-by-side
plt.figure(figsize=(12, 6))

plt.subplot(1, 2, 1)
sns.scatterplot(x='PC1', y='PC2', hue='Diagnosis', data=pca_df,
palette=value_colors)
plt.title('Ground Diagnoses')
```

Homework IV – Group 107

(ist1103811, ist1103479)

```
# Perform k-means clustering with k=3
kmeans = KMeans(n_clusters=3, random_state=0)
cluster_labels = kmeans.fit_predict(normalized_data)

# Create a palette of colors based on unique values in the 'cluster_labels' array
cluster_palette = sns.color_palette("Set1", len(np.unique(cluster_labels)))

# Map colors to values in the 'cluster_labels' array
cluster_colors = {cluster: color for cluster, color in
zip(np.unique(cluster_labels), cluster_palette)}

# Add cluster labels to the PCA DataFrame
pca_df['Cluster'] = cluster_labels

plt.subplot(1, 2, 2)
sns.scatterplot(x='PC1', y='PC2', hue='Cluster', data=pca_df,
palette=cluster_colors)
plt.title('K-Means Clustering (k=3)')

plt.savefig("Ex3GRAPH")
```



Homework IV – Group 107

(ist1103811, ist1103479)

4)

Ao usarmos clusters podemos criar "grupos" que representam diferentes condições e grupos de risco. Por exemplo, como no exercício 3, tendo um clustering de k=3, pode originar-se 3 clusters, em que cada um identifica um tipo diferente de condição. Um pode ser para pessoas que sejam saudáveis enquanto que as outras duas categorias são para pessoas com problemas diferentes de saúde. Esta distinção é importante e útil porque facilita na escolha de tratamento de paciente para paciente. Para além dessa vantagem, os clusters podem também associar certas categorias de risco a certos problemas (Exemplo hipotético: Se a pelvic_tilt for superior a 20, o paciente é mais propício a ter uma hérnia) .

Outra vantagem de usarmos clusters é que estes ajudam-nos a encontrar outliers, ou seja, indivíduos com padrões de saúde incomuns e raros. Estes casos podem ser revistos e investigados com mais pormenor de forma a dar o tratamento mais adequado à pessoa.

END