Aprendizagem 2021/22

Homework IV – Group 020

I. Pen-and-paper

1)

$\mu_1 = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} -1 \\ -4 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad \pi_1 = p(c_1 = 1) = 0.7$

$\pi_2 = p(c_2 = 1) = 0.3$

$\Sigma_1^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \Sigma_2^{-1} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}$

$|\Sigma_1| = 1 \qquad |\Sigma_2| = 4$

$\boxed{\text{E-step}}$

$\boxed{\text{Likelihoods}}$

$$p(x_1 | c_1 = 1) = \mathcal{N}\left(\begin{bmatrix} 2 \\ 4 \end{bmatrix} \middle| \begin{bmatrix} 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = \frac{1}{2\pi\sqrt{1}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} 2 \\ 4 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix}\right)^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\left(\begin{bmatrix} 2 \\ 4 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix}\right)\right) = 0.159155$$

$$p(x_2 | c_1 = 1) = \mathcal{N}\left(\begin{bmatrix} -1 \\ -4 \end{bmatrix} \middle| \begin{bmatrix} 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = \frac{1}{2\pi\sqrt{1}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} -1 \\ -4 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix}\right)^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\left(\begin{bmatrix} -1 \\ -4 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix}\right)\right) = 2.2391 \times 10^{-17}$$

$$p(x_3 | c_1 = 1) = \mathcal{N}\left(\begin{bmatrix} -1 \\ 2 \end{bmatrix} \middle| \begin{bmatrix} 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = \frac{1}{2\pi\sqrt{1}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} -1 \\ 2 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix}\right)^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\left(\begin{bmatrix} -1 \\ 2 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix}\right)\right) = 0.000239$$

$$p(x_4 | c_1 = 1) = \mathcal{N}\left(\begin{bmatrix} 4 \\ 0 \end{bmatrix} \middle| \begin{bmatrix} 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = \frac{1}{2\pi\sqrt{1}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} 4 \\ 0 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix}\right)^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\left(\begin{bmatrix} 4 \\ 0 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix}\right)\right) = 7.2256 \times 10^{-6}$$

$$p(x_1 | c_1 = 2) = \mathcal{N}\left(\begin{bmatrix} 2 \\ 4 \end{bmatrix} \middle| \begin{bmatrix} -1 \\ -4 \end{bmatrix}, \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\right) = \frac{1}{2\pi\sqrt{4}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} 2 \\ 4 \end{bmatrix} - \begin{bmatrix} -1 \\ -4 \end{bmatrix}\right)^T \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\left(\begin{bmatrix} 2 \\ 4 \end{bmatrix} - \begin{bmatrix} -1 \\ -4 \end{bmatrix}\right)\right) = 9.4388 \times 10^{-10}$$

$$p(x_2 | c_1 = 2) = \mathcal{N}\left(\begin{bmatrix} -1 \\ -4 \end{bmatrix} \middle| \begin{bmatrix} -1 \\ -4 \end{bmatrix}, \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\right) = \frac{1}{2\pi\sqrt{4}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} -1 \\ -4 \end{bmatrix} - \begin{bmatrix} -1 \\ -4 \end{bmatrix}\right)^T \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\left(\begin{bmatrix} -1 \\ -4 \end{bmatrix} - \begin{bmatrix} -1 \\ -4 \end{bmatrix}\right)\right) = 0.079577$$

$$p(x_3 | c_1 = 2) = \mathcal{N}\left(\begin{bmatrix} -1 \\ 2 \end{bmatrix} \middle| \begin{bmatrix} -1 \\ -4 \end{bmatrix}, \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\right) = \frac{1}{2\pi\sqrt{4}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} -1 \\ 2 \end{bmatrix} - \begin{bmatrix} -1 \\ -4 \end{bmatrix}\right)^T \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\left(\begin{bmatrix} -1 \\ 2 \end{bmatrix} - \begin{bmatrix} -1 \\ -4 \end{bmatrix}\right)\right) = 9.8206 \times 10^{-6}$$

$$p(x_4 | c_1 = 2) = \mathcal{N}\left(\begin{bmatrix} 4 \\ 0 \end{bmatrix} \middle| \begin{bmatrix} -1 \\ -4 \end{bmatrix}, \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\right) = \frac{1}{2\pi\sqrt{4}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} 4 \\ 0 \end{bmatrix} - \begin{bmatrix} -1 \\ -4 \end{bmatrix}\right)^T \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\left(\begin{bmatrix} 4 \\ 0 \end{bmatrix} - \begin{bmatrix} -1 \\ -4 \end{bmatrix}\right)\right) = 2.8137 \times 10^{-6}$$

$\boxed{\text{Joints}}$

$$p(x_1, C_1 = 1) = p(C_1 = 1)\, p(x_1 | C_1 = 1) = 0.7 \times 0.159155 = 0.111408$$
$$p(x_2, C_1 = 1) = p(C_1 = 1)\, p(x_2 | C_1 = 1) = 0.7 \times 2.2391 \times 10^{-17} = 1.5674 \times 10^{-17}$$
$$p(x_3, C_1 = 1) = p(C_1 = 1)\, p(x_3 | C_1 = 1) = 0.7 \times 0.000239 = 0.000167$$
$$p(x_4, C_1 = 1) = p(C_1 = 1)\, p(x_4 | C_1 = 1) = 0.7 \times 7.2256 \times 10^{-6} = 5.0579 \times 10^{-6}$$
$$p(x_1, C_2 = 1) = p(C_2 = 1)\, p(x_1 | C_2 = 1) = 0.3 \times 9.4388 \times 10^{-10} = 2.8316 \times 10^{-10}$$
$$p(x_2, C_2 = 1) = p(C_2 = 1)\, p(x_2 | C_2 = 1) = 0.3 \times 0.079577 = 0.023873$$
$$p(x_3, C_2 = 1) = p(C_2 = 1)\, p(x_3 | C_2 = 1) = 0.3 \times 9.8206 \times 10^{-6} = 2.9462 \times 10^{-6}$$
$$p(x_4, C_2 = 1) = p(C_2 = 1)\, p(x_4 | C_2 = 1) = 0.3 \times 2.8137 \times 10^{-6} = 8.4411 \times 10^{-7}$$

$\boxed{\text{Denominators}}$

$$p(x_1) = p(x_1, C_1 = 1) + p(x_1, C_2 = 1) = 0.111408 + 2.8316 \times 10^{-10} = 0.111408$$
$$p(x_2) = p(x_2, C_1 = 1) + p(x_2, C_2 = 1) = 1.5674 \times 10^{-17} + 0.023873 = 0.023873$$
$$p(x_3) = p(x_3, C_1 = 1) + p(x_3, C_2 = 1) = 0.000167 + 2.9462 \times 10^{-6} = 0.000170$$
$$p(x_4) = p(x_4, C_1 = 1) + p(x_4, C_2 = 1) = 5.0579 \times 10^{-6} + 8.4411 \times 10^{-7} = 5.9020 \times 10^{-6}$$

Posteriors

$$p(c_1=1|x_1) = \frac{P(x_1, c_1=1)}{P(x_1)} = \frac{0.111408}{0.111408} = 1$$
$$p(c_1=1|x_2) = \frac{P(x_2, c_1=1)}{P(x_2)} = \frac{1.5674 \times 10^{-17}}{0.023873} = 6.566 \times 10^{-16} \approx 0$$
$$p(c_1=1|x_3) = \frac{P(x_3, c_1=1)}{P(x_3)} = \frac{0.000167}{0.000170} = 0.982$$
$$p(c_1=1|x_4) = \frac{P(x_4, c_1=1)}{P(x_4)} = \frac{5.0579 \times 10^{-6}}{5.9020 \times 10^{-6}} = 0.857$$
$$p(c_2=1|x_1) = \frac{P(x_1, c_2=1)}{P(x_1)} = \frac{2.8316 \times 10^{-10}}{0.111408} = 2.542 \times 10^{-9} \approx 0$$
$$p(c_2=1|x_2) = \frac{P(x_2, c_2=1)}{P(x_2)} = \frac{0.023873}{0.023873} = 1$$
$$p(c_2=1|x_3) = \frac{P(x_3, c_2=1)}{P(x_3)} = \frac{2.9462 \times 10^{-6}}{0.000170} = 0.018$$
$$p(c_2=1|x_4) = \frac{P(x_4, c_2=1)}{P(x_4)} = \frac{8.4411 \times 10^{-7}}{5.9020 \times 10^{-6}} = 0.143$$

→ Resultados: $P_1 = \begin{bmatrix} 1 \\ 0 \\ 0.982 \\ 0.857 \end{bmatrix}$ $P_2 = \begin{bmatrix} 0 \\ 1 \\ 0.018 \\ 0.143 \end{bmatrix}$ $\Rightarrow$ $x_1, x_3, x_4 \in c_1$, $x_2 \in c_2$

$w_1 = 1 + 0.982 + 0.857 = 2.839$ $\quad$ $w_2 = 1 + 0.018 + 0.143 = 1.161$

M-step

Estimate priors

$$p(c_1=1) = \frac{w_1}{w_1+w_2} = \frac{2.839}{4} = 0.710 \qquad p(c_2=1) = \frac{w_2}{w_1+w_2} = \frac{1.161}{4} = 0.290$$

Estimate μ_1, μ_2

$$\mu_1 = \frac{1 \times \begin{bmatrix} 2 \\ 4 \end{bmatrix} + 0 \times \begin{bmatrix} -1 \\ -4 \end{bmatrix} + 0.982 \begin{bmatrix} -1 \\ 2 \end{bmatrix} + 0.857 \begin{bmatrix} 4 \\ 0 \end{bmatrix}}{2.839} = \begin{bmatrix} 1.566 \\ 2.101 \end{bmatrix}$$

$$\mu_2 = \frac{0 \times \begin{bmatrix} 2 \\ 4 \end{bmatrix} + 1 \times \begin{bmatrix} -1 \\ -4 \end{bmatrix} + 0.018 \begin{bmatrix} -1 \\ 2 \end{bmatrix} + 0.143 \begin{bmatrix} 4 \\ 0 \end{bmatrix}}{1.161} = \begin{bmatrix} -0.384 \\ -3.414 \end{bmatrix}$$

Estimate Σ_1, Σ_2

$$\Sigma_1 = \frac{1}{2.839}\left(1 \times \begin{bmatrix} 2-1.566 \\ 4-2.101 \end{bmatrix}\begin{bmatrix} 2-1.566 & 4-2.101 \end{bmatrix} + 0 \times \begin{bmatrix} -1-1.566 \\ -4-2.101 \end{bmatrix}\begin{bmatrix} -1-1.566 & -4-2.101 \end{bmatrix} + \right.$$
$$\left. + 0.982 \begin{bmatrix} -1-1.566 \\ 2-2.101 \end{bmatrix}\begin{bmatrix} -1-1.566 & 2-2.101 \end{bmatrix} + 0.857 \begin{bmatrix} 4-1.566 \\ 0-2.101 \end{bmatrix}\begin{bmatrix} 4-1.566 & 0-2.101 \end{bmatrix}\right) = \begin{bmatrix} 4.132 & -1.164 \\ -1.164 & 2.606 \end{bmatrix}$$

$$\Sigma_2 = \frac{1}{1.161}\left(0 \times \begin{bmatrix} 2+0.384 \\ 4+3.414 \end{bmatrix}\begin{bmatrix} 2+0.384 & 4+3.414 \end{bmatrix} + 1 \times \begin{bmatrix} -1+0.384 \\ -4+3.414 \end{bmatrix}\begin{bmatrix} -1+0.384 & -4+3.414 \end{bmatrix} + \right.$$
$$\left. + 0.018 \begin{bmatrix} -1+0.384 \\ 2+3.414 \end{bmatrix}\begin{bmatrix} -1+0.384 & 2+3.414 \end{bmatrix} + 0.143 \begin{bmatrix} 4+0.384 \\ 0+3.414 \end{bmatrix}\begin{bmatrix} 4+0.384 & 0+3.414 \end{bmatrix}\right) = \begin{bmatrix} 2.700 & 2.103 \\ 2.103 & 2.186 \end{bmatrix}$$

E-step após o update

$$\mu_1 = \begin{bmatrix} 1.566 \\ 2.101 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} -0.384 \\ -3.414 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 4.132 & -1.164 \\ -1.164 & 2.606 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 2.700 & 2.103 \\ 2.103 & 2.186 \end{bmatrix} \quad \pi_1 = p(c_1=1) = 0.7 \quad \pi_2 = p(c_2=1) = 0.3$$

$$\Sigma_1^{-1} = \begin{bmatrix} 0.277 & 0.124 \\ 0.124 & 0.439 \end{bmatrix} \quad \Sigma_2^{-1} = \begin{bmatrix} 1.477 & -1.421 \\ -1.421 & 1.825 \end{bmatrix}$$

Aprendizagem 2021/22

Homework IV – Group 020

$$|\Sigma_1| = 9.413 \qquad |\Sigma_2| = 1.480$$

$\boxed{\text{Likelihoods}}$

$$p(x_1|c_1=1) = \mathcal{N}\left(\begin{bmatrix}2\\4\end{bmatrix}\middle|\begin{bmatrix}1.566\\2.101\end{bmatrix}, \begin{bmatrix}4.132 & -1.164\\-1.164 & 2.606\end{bmatrix}\right) = 0.000835$$

$$p(x_2|c_1=1) = \mathcal{N}\left(\begin{bmatrix}-1\\-4\end{bmatrix}\middle|\begin{bmatrix}1.566\\2.101\end{bmatrix}, \begin{bmatrix}4.132 & -1.164\\-1.164 & 2.606\end{bmatrix}\right) = 4.5456 \times 10^{-21}$$

$$p(x_3|c_1=1) = \mathcal{N}\left(\begin{bmatrix}-1\\2\end{bmatrix}\middle|\begin{bmatrix}1.566\\2.101\end{bmatrix}, \begin{bmatrix}4.132 & -1.164\\-1.164 & 2.606\end{bmatrix}\right) = 8.5579 \times 10^{-8}$$

$$p(x_4|c_1=1) = \mathcal{N}\left(\begin{bmatrix}4\\0\end{bmatrix}\middle|\begin{bmatrix}1.566\\2.101\end{bmatrix}, \begin{bmatrix}4.132 & -1.164\\-1.164 & 2.606\end{bmatrix}\right) = 2.0713 \times 10^{-12}$$

$$p(x_1|c_1=2) = \mathcal{N}\left(\begin{bmatrix}2\\4\end{bmatrix}\middle|\begin{bmatrix}-0.384\\-3.414\end{bmatrix}, \begin{bmatrix}2.700 & 2.103\\2.103 & 2.186\end{bmatrix}\right) = 3.5434 \times 10^{-47}$$

$$p(x_2|c_1=2) = \mathcal{N}\left(\begin{bmatrix}-1\\-4\end{bmatrix}\middle|\begin{bmatrix}-0.384\\-3.414\end{bmatrix}, \begin{bmatrix}2.700 & 2.103\\2.103 & 2.186\end{bmatrix}\right) = 0.025207$$

$$p(x_3|c_1=2) = \mathcal{N}\left(\begin{bmatrix}-1\\2\end{bmatrix}\middle|\begin{bmatrix}-0.384\\-3.414\end{bmatrix}, \begin{bmatrix}2.700 & 2.103\\2.103 & 2.186\end{bmatrix}\right) = 1.0630 \times 10^{-12}$$

$$p(x_4|c_1=2) = \mathcal{N}\left(\begin{bmatrix}4\\0\end{bmatrix}\middle|\begin{bmatrix}-0.384\\-3.414\end{bmatrix}, \begin{bmatrix}2.700 & 2.103\\2.103 & 2.186\end{bmatrix}\right) = 4.4267 \times 10^{-32}$$

$\boxed{\text{Joints}}$

$$p(x_1, C_1=1) = p(C_1=1)\,p(x_1|C_1=1) = 0.000585$$
$$p(x_2, C_1=1) = p(C_1=1)\,p(x_2|C_1=1) = 3.1820 \times 10^{-21}$$
$$p(x_3, C_1=1) = p(C_1=1)\,p(x_3|C_1=1) = 5.9905 \times 10^{-8}$$
$$p(x_4, C_1=1) = p(C_1=1)\,p(x_4|C_1=1) = 1.4499 \times 10^{-12}$$
$$p(x_1, C_2=1) = p(C_2=1)\,p(x_1|C_2=1) = 1.0630 \times 10^{-47}$$
$$p(x_2, C_2=1) = p(C_2=1)\,p(x_2|C_2=1) = 0.007562$$
$$p(x_3, C_2=1) = p(C_2=1)\,p(x_3|C_2=1) = 3.1889 \times 10^{-13}$$
$$p(x_4, C_2=1) = p(C_2=1)\,p(x_4|C_2=1) = 1.3280 \times 10^{-32}$$

$\boxed{\text{Denominators}}$

$$p(x_1) = p(x_1, C_1=1) + p(x_1, C_2=1) = 0.000585$$
$$p(x_2) = p(x_2, C_1=1) + p(x_2, C_2=1) = 0.007562$$
$$p(x_3) = p(x_3, C_1=1) + p(x_3, C_2=1) = 5.9906 \times 10^{-8}$$
$$p(x_4) = p(x_4, C_1=1) + p(x_4, C_2=1) = 1.4499 \times 10^{-12}$$

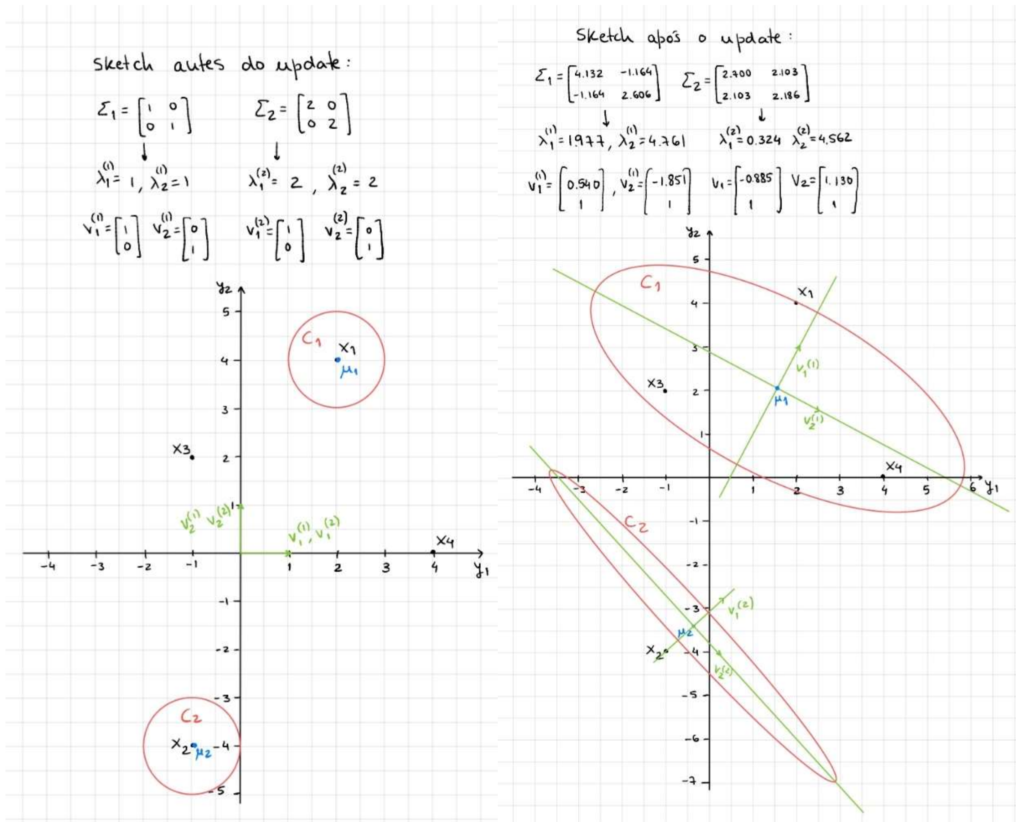
$\boxed{\text{Posteriors}}$

$$p(c_1=1|x_1) = \frac{P(x_1, C_1=1)}{P(x_1)} = 1$$
$$p(c_1=1|x_2) = \frac{P(x_2, C_1=1)}{P(x_2)} = 4.2077 \times 10^{-19} \approx 0$$
$$p(c_1=1|x_3) = \frac{P(x_3, C_1=1)}{P(x_3)} = 0.9999995$$
$$p(c_1=1|x_4) = \frac{P(x_4, C_1=1)}{P(x_4)} = 1$$
$$p(c_2=1|x_1) = \frac{P(x_1, C_2=1)}{P(x_1)} = 1.8179 \times 10^{-44} \approx 0$$
$$p(c_2=1|x_2) = \frac{P(x_2, C_2=1)}{P(x_2)} = 1$$
$$p(c_2=1|x_3) = \frac{P(x_3, C_2=1)}{P(x_3)} = 5.3232 \times 10^{-6}$$
$$p(c_2=1|x_4) = \frac{P(x_4, C_2=1)}{P(x_4)} = 9.1594 \times 10^{-21} \approx 0$$

$$\Rightarrow \quad x_1, x_3, x_4 \in c_1 \qquad x_2 \in c_2$$

Aprendizagem 2021/22

Homework IV – Group 020



2)

Antes e depois do update obtemos os mesmos resultados: $x_1, x_3, x_4 \in c_1$ e $x_2 \in c_2$ logo vamos obter a mesma silhouette em ambos:

Silhouette de c1

$$s(x_1) = 1 - \frac{a(x_1)}{b(x_1)} = 1 - \frac{\frac{1}{2}(\|x_1 - x_3\|_2 + \|x_1 - x_4\|_2)}{\|x_1 - x_2\|_2} = 0.527289$$

$$s(x_3) = 1 - \frac{a(x_3)}{b(x_3)} = 1 - \frac{\frac{1}{2}(\|x_3 - x_1\|_2 + \|x_3 - x_4\|_2)}{\|x_3 - x_2\|_2} = 0.250774$$

$$s(x_4) = 1 - \frac{a(x_4)}{b(x_4)} = 1 - \frac{\frac{1}{2}(\|x_4 - x_1\|_2 + \|x_4 - x_3\|_2)}{\|x_4 - x_2\|_2} = 0.230274$$

$$\rightarrow s(c_1) = \frac{s(x_1) + s(x_3) + s(x_4)}{3} = \frac{0.527289 + 0.250774 + 0.230274}{3} = 0.336112$$

Silhouette de c2

$$s(x_2) = 1 - \frac{a(x_2)}{b(x_2)} = 1 - \frac{0}{\frac{1}{3}(\|x_2 - x_1\|_2 + \|x_2 - x_3\|_2 + \|x_2 - x_4\|_2)} = 1$$

$$\rightarrow s(c_2) = \frac{s(x_2)}{1} = \frac{1}{1} = 1$$

Silhouette da solução

$$s(c) = \frac{s(c_1) + s(c_2)}{2} = \frac{0.336112 + 1}{2} = 0.668$$

A silhouette é superior a 0.5 logo a qualidade da solução de clusters produzidos é boa, ou seja têm uma boa coesão ($a(x)$ pequeno) e uma boa separação ($b(x)$ elevado).

É de notar que após o update a alocação das observações aos clusters é mais acertiva uma vez que a probabilidade de um ponto pertencer a um certo cluster dá 1 ou muito perto de 1 para todos os pontos.

Aprendizagem 2021/22

Homework IV – Group 020

3)

a. i)

O MLP vai ter a seguinte forma: 5 - 5 - 5 - 5 - 1

input layer; hidden layers; output layer

Escolheu-se usar 1 output layer para a classificação binária, apesar de também ser possível com 2, pois requer menos pesos (w) e bias (b) e assim pode ser mais fácil de treinar.
Assim temos:

$$5 - 5 - 5 - 5 - 1$$

$W^{[1]}_{5\times5}$ $W^{[2]}_{5\times5}$ $W^{[3]}_{5\times5}$ $W^{[4]}_{5\times1}$

$b^{[1]}_{5\times1}$ $b^{[2]}_{5\times1}$ $b^{[3]}_{5\times1}$ $b^{[4]}_{1\times1}$

$\longrightarrow VC = 3\times25 + 3\times5 + 5+1 = 96$

ii)

Independentemente do tamanho do problema podemos sempre criar uma árvore de decisão com um ponto do dataset em cada folha logo a VC dimension é a máxima possível e como as variaveis estão em 3 bins $\rightarrow VC = 2^3 = 8$

iii)

Prior: $P(c) = 2-1 = 1$

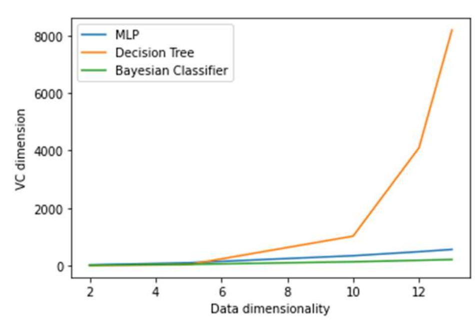
Likelihood:

$P(y_1 y_2 y_3 y_4 y_5 | c=0) = N(\mu_0, \Sigma_0) = 5\times1 + \frac{5\times5-5}{2} + 5 = 20$

$P(y_1 y_2 y_3 y_4 y_5 | c=1) = N(\mu_1, \Sigma_1) = 20$

$\longrightarrow VC = 20 + 20 + 1 = 41$

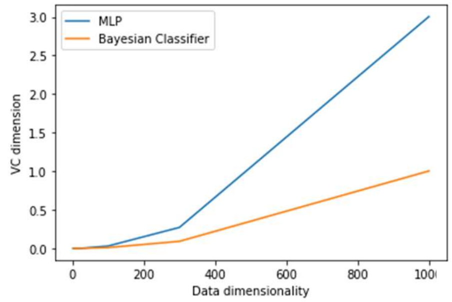
b. Como podemos ver pelo gráfico enquanto que o MLP e o *Bayesian Classifier* crescem linearmente, a *Decision Tree* cresce exponencialmente com o aumento da dimensão da data, tomando valores muito maiores que os outros, ou seja, é o modelo mais complexo e está por isso mais suscetível a *overfitting*.



c. Como podemos ver pelo gráfico abaixo para dimensões de data mais elevadas que em **b.** o MLP e o *Bayesian Classifier* já não crescem linearmente e o MLP cresce mais rapidamente que o *Bayesian Classifier* sendo por isso um modelo mais complexo.

Aprendizagem 2021/22

Homework IV – Group 020



II. Programming and critical analysis

4)

a. Seguem abaixo os valores obtidos para o ECR aplicando o *k-means clustering unsupervised* no *dataset* para k = 2 e k = 3.

	k = 2	k = 3
ECR	13.5	6.66(6)

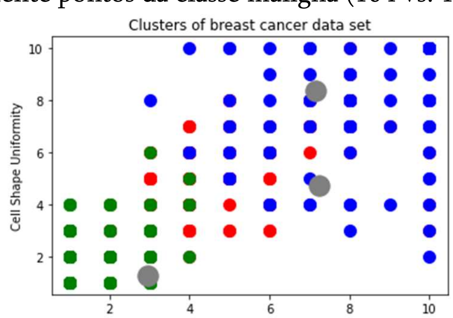
Os valores do ECR são baixos relativamente ao tamanho do *dataset* (≈700) em ambos os casos. Para k = 2, apenas 27 (2*13.5) pontos pertencem a uma classe diferente da classe dominante nos 2 clusters e para k = 3, apenas ≈ 20 (3*6.66(6)) pontos pertencem a uma classe diferente da classe dominante nos 3 clusters.

b. Seguem abaixo os valores obtidos para o *Silhouette coefficient* aplicando o *k-means clustering unsupervised* no *dataset* para k = 2 e k = 3.

	k = 2	k = 3
Silhouette	0.597	0.525

Os valores da *Silhouette* são os dois semelhantes e acima de 0.5 o que indica que a qualidade da solução de clusters produzidos é boa em ambos os casos, sendo um pouco melhor para k = 2.

5) Segue abaixo o gráfico dos clusters obtidos aplicando o *k-means* com k = 3. Os pontos cinzentos representam o centro dos 3 clusters. O cluster verde tem maioritariamente pontos da classe benigna (433 vs. 9) e os clusters vermelho e azul têm maioritariamente pontos da classe maligna (104 vs. 11) e (126 vs. 0) respetivamente.



6) Como se pode verificar no gráfico acima existe uma boa separação entre os clusters verde (classe benigna) e azul (classe maligna) estando o cluster vermelho (classe maligna) misturado entre os dois mas mais com cluster azul, o que faz sentido visto que representam a mesma classe. Assim podemos concluir que no geral, existe uma boa separação entre os clusters que representam classes diferentes e por isso a qualidade da solução produzida é boa.

III. APPENDIX

```
# -------------------- Exercicio 3b) --------------------
import matplotlib.pyplot as plt
import numpy as np

def mlp(i):
    return 3*i*i + 3*i + i + 1

def tree(i):
    return np.power(2, i)

def nb(i):
    l = i*1 + (i*i-i)/2 + i
    return 2*l + 1

x = [2,5,10,12,13]
y_mlp = [0] * 5
y_tree = [0] * 5
y_nb = [0] * 5

j = 0

for i in x:
    y_mlp[j] = mlp(i)
    y_tree[j] = tree(i)
    y_nb[j] = nb(i)
    j += 1

plt.plot(x, y_mlp, label = "MLP")
plt.plot(x, y_tree, label = "Decision Tree")
plt.plot(x, y_nb, label = "Bayesian Classifier")

plt.xlabel('Data dimensionality')
plt.ylabel('VC dimension')
plt.legend()
plt.show()

# -------------------- Exercicio 3c) --------------------
x = [2,5,10,30,100,300,1000]
y_mlp = [0] * 7
y_tree = [0] * 7
y_nb = [0] * 7

j = 0

for i in x:
    y_mlp[j] = mlp(i)
    y_nb[j] = nb(i)
    j += 1

plt.plot(x, y_mlp, label = "MLP")
plt.plot(x, y_nb, label = "Bayesian Classifier")

plt.xlabel('Data dimensionality')
plt.ylabel('VC dimension')
plt.legend()
plt.show()


# -------------------- Exercicio 4 --------------------
from scipy.io import arff
import numpy as np
```

Aprendizagem 2021/22

Homework IV – Group 020

```python
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_samples, silhouette_score
import matplotlib.pyplot as plt
from sklearn.feature_selection import SelectKBest, mutual_info_regression

def calculate_ecr(clusters, y):
    c0, c1, c2 = 0,0,0
    m0, b0, m1, b1, m2, b2 = 0,0,0,0,0,0
    for i,j in zip(clusters, y):
        if (i == 0):
            c0 += 1
            if(j.decode('UTF-8') == 'benign'):
                b0 += 1
            else:
                m0 += 1
        if (i == 1):
            c1 += 1
            if(j.decode('UTF-8') == 'benign'):
                b1 +=1
            else:
                m1 += 1
        if (i == 2):
            c2 += 1
            if(j.decode('UTF-8') == 'benign'):
                b2 += 1
            else:
                m2 += 1

    if(c2 == 0):
        ecr = (1/2)*((c0-max(b0,m0)+(c1-max(b1,m1))))
    else:
        ecr = (1/3)*((c0-max(b0,m0)+(c1-max(b1,m1))+(c2-max(b2,m2))))

    return ecr

data = arff.loadarff(r'/home/sara/apre/tpc-4/breast.w.arff')
df = pd.DataFrame(data[0])

data_array = df.to_numpy()

x = np.empty((0,9))
y = np.empty((0,1))

for row in data_array:
    array_sum = np.sum(row[:-1])
    array_has_nan = np.isnan(array_sum)
    if(not array_has_nan):
        x = np.append(x, [row[:-1]], axis=0)
        y = np.append(y, row[-1])

range_n_clusters = [2,3]

for n_clusters in range_n_clusters:
    kmeans = KMeans(n_clusters=n_clusters, random_state=0).fit(x)
    cluster_labels = kmeans.predict(x)

    # -------------------- Exercicio 4a) ---------------------
    ecr = calculate_ecr(cluster_labels, y)
    print(ecr)

    # -------------------- Exercicio 4b) ---------------------
    silhouette = silhouette_score(x, cluster_labels)
    print(silhouette)
```

Aprendizagem 2021/22

Homework IV – Group 020

```
# -------------------- Exercicio 5 ---------------------
y_aux = np.empty((0,1))

for idx, val in enumerate(y):
    if(val.decode('UTF-8') == 'benign'):
        y_aux = np.append(y_aux, 1)
    else:
        y_aux = np.append(y_aux, 0)

x_new = SelectKBest(score_func=mutual_info_regression, k=2).fit_transform(x, y_aux)

plt.scatter(x_new[cluster_labels==0, 0], x_new[cluster_labels==0, 1], s=100, c='red', label
='Cluster 1')
plt.scatter(x_new[cluster_labels==1, 0], x_new[cluster_labels==1, 1], s=100, c='blue', label
='Cluster 2')
plt.scatter(x_new[cluster_labels==2, 0], x_new[cluster_labels==2, 1], s=100, c='green', label
='Cluster 3')

centers = kmeans.cluster_centers_

plt.scatter(centers[:, 0], centers[:, 1], s=300, c='grey', label = 'Centroids')
plt.title('Clusters of breast cancer data set')
plt.xlabel('Cell Size Uniformity')
plt.ylabel('Cell Shape Uniformity')
plt.show()
```

END