TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 018

I. Pen-and-paper [11v]

Given the bivariate observations $\left\{\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}-1\\1\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}\right\}$,
and the multivariate Gaussian mixture

$$\mathbf{u}_1 = \begin{pmatrix}2\\2\end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix}0\\0\end{pmatrix}, \mathbf{\Sigma}_1 = \begin{pmatrix}2 & 1\\1 & 2\end{pmatrix}, \mathbf{\Sigma}_2 = \begin{pmatrix}2 & 0\\0 & 2\end{pmatrix}, \pi_1 = 0.5, \pi_2 = 0.5.$$

1) [7v] Perform one epoch of the EM clustering algorithm and determine the new parameters. Indicate all calculus step by step (you can use a computer, however, disclose intermediary steps).

$$k = 1 : \left\{u_1 = \begin{pmatrix}2\\2\end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix}2 & 1\\1 & 2\end{pmatrix}\right\}, \qquad k = 2 : \left\{u_2 = \begin{pmatrix}0\\0\end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix}2 & 0\\0 & 2\end{pmatrix}\right\};$$

Para começar, temos que:

$$\det(\Sigma_1) = (2 \times 2) - (1 \times 1) = 3, \qquad \det(\Sigma_2) = (2 \times 2) - (0 \times 0) = 4;$$

$$(\Sigma_1)^{-1} = \begin{bmatrix}\frac{2}{3} & -\frac{1}{3}\\-\frac{1}{3} & \frac{2}{3}\end{bmatrix}, \qquad (\Sigma_2)^{-1} = \begin{bmatrix}\frac{1}{2} & 0\\0 & \frac{1}{2}\end{bmatrix}, \qquad x_1 = \begin{bmatrix}1\\2\end{bmatrix}, \; x_2 = \begin{bmatrix}-1\\1\end{bmatrix}, \; x_3 = \begin{bmatrix}1\\0\end{bmatrix}.$$

– **Passo E**: $\gamma_{ik} = p(c_k|x_i) = \frac{p(x_i|c_k)\,p(c_k)}{p(x_i)} = \frac{N(x_i|u_k,\Sigma_k)\,\pi_k}{\sum_k \pi_k\, N(x_i|u_k,\Sigma_k)}$

$$p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}} \times \frac{1}{\det(\Sigma)^{1/2}} \times e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}, \qquad com\ d = 2.$$

Para x_1:

$$p(x_1 \mid \mu_1, \Sigma_1) = \frac{1}{2\pi} \times \frac{1}{\sqrt{3}} \times \exp\left(-\frac{1}{2}\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}2\\2\end{bmatrix}\right)^T \begin{bmatrix}\frac{2}{3} & -\frac{1}{3}\\-\frac{1}{3} & \frac{2}{3}\end{bmatrix}\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}2\\2\end{bmatrix}\right)\right) = 0.06584$$

$$p(x_1 \mid \mu_2, \Sigma_2) = \frac{1}{2\pi} \times \frac{1}{\sqrt{4}} \times \exp\left(-\frac{1}{2}\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0\\0\end{bmatrix}\right)^T \begin{bmatrix}\frac{1}{2} & 0\\0 & \frac{1}{2}\end{bmatrix}\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0\\0\end{bmatrix}\right)\right) = 0.02280$$

$$p(k=1 \mid x_1) = \frac{p(x_1 \mid k=1)p(k=1)}{p(x_1)} = \frac{p(x_1 \mid \mu_1, \sum_1)\pi_1}{p(x_1)} = \frac{0.06584 \times 0.5}{p(x_1)} = \frac{0.03292}{p(x_1)}$$

$$p(k=2 \mid x_1) = \frac{p(x_1 \mid k=2)p(k=2)}{p(x_1)} = \frac{p(x_1 \mid \mu_2, \sum_2)\pi_2}{p(x_1)} = \frac{0.02280 \times 0.5}{p(x_1)} = \frac{0.01140}{p(x_1)}$$

Normalizando:

$$p(k=1 \mid x_1) = \frac{0.03292}{0.03292 + 0.01140} = 0.7428$$

$$p(k=2 \mid x_1) = \frac{0.01140}{0.03292 + 0.01140} = 0.2572$$

Para x_2:

$$p(x_2 \mid \mu_1, \sum_1) = \frac{1}{2\pi} \times \frac{1}{\sqrt{3}} \times \exp\left(-\frac{1}{2}\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}2\\2\end{bmatrix}\right)^T \begin{bmatrix}\frac{2}{3} & -\frac{1}{3}\\ -\frac{1}{3} & \frac{2}{3}\end{bmatrix} \left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}2\\2\end{bmatrix}\right)\right) = 0.008911$$

$$p(x_2 \mid \mu_2, \sum_2) = \frac{1}{2\pi} \times \frac{1}{\sqrt{4}} \times \exp\left(-\frac{1}{2}\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0\\0\end{bmatrix}\right)^T \begin{bmatrix}\frac{1}{2} & 0\\ 0 & \frac{1}{2}\end{bmatrix} \left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0\\0\end{bmatrix}\right)\right) = 0.04827$$

$$p(k=1 \mid x_2) = \frac{p(x_2 \mid k=1)p(k=1)}{p(x_2)} = \frac{p(x_2 \mid \mu_1, \sum_1)\pi_1}{p(x_2)} = \frac{0.008911 \times 0.5}{p(x_2)} = \frac{0.004456}{p(x_2)}$$

$$p(k=2 \mid x_2) = \frac{p(x_2 \mid k=2)p(k=2)}{p(x_2)} = \frac{p(x_2 \mid \mu_2, \sum_2)\pi_2}{p(x_2)} = \frac{0.04827 \times 0.5}{p(x_2)} = \frac{0.02414}{p(x_2)}$$

Normalizando:

$$p(k=1 \mid x_2) = \frac{0.004456}{0.004456 + 0.02414} = 0.1558$$

$$p(k=2 \mid x_2) = \frac{0.02414}{0.004456 + 0.02414} = 0.8442$$

TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 018

Para x_3:

$$p(x_3 \mid \mu_1, \sum_1) = \frac{1}{2\pi} \times \frac{1}{\sqrt{3}} \times \exp\left(-\frac{1}{2}\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}2\\2\end{bmatrix}\right)^T \begin{bmatrix}\frac{2}{3} & -\frac{1}{3}\\ -\frac{1}{3} & \frac{2}{3}\end{bmatrix} \left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}2\\2\end{bmatrix}\right)\right) = 0.03380$$

$$p(x_3 \mid \mu_2, \sum_2) = \frac{1}{2\pi} \times \frac{1}{\sqrt{4}} \times \exp\left(-\frac{1}{2}\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0\\0\end{bmatrix}\right)^T \begin{bmatrix}\frac{1}{2} & 0\\ 0 & \frac{1}{2}\end{bmatrix} \left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0\\0\end{bmatrix}\right)\right) = 0.06197$$

$$p(k=1 \mid x_3) = \frac{p(x_3 \mid k=1)p(k=1)}{p(x_3)} = \frac{p(x_3 \mid \mu_1, \sum_1)\pi_1}{p(x_3)} = \frac{0.03380 \times 0.5}{p(x_3)} = \frac{0.01690}{p(x_3)}$$

$$p(k=2 \mid x_3) = \frac{p(x_3 \mid k=2)p(k=2)}{p(x_3)} = \frac{p(x_3 \mid \mu_2, \sum_2)\pi_2}{p(x_3)} = \frac{0.06197 \times 0.5}{p(x_3)} = \frac{0.03099}{p(x_3)}$$

Normalizando:

$$p(k=1 \mid x_3) = \frac{0.01690}{0.01690 + 0.03099} = 0.3529$$

$$p(k=2 \mid x_3) = \frac{0.03099}{0.01690 + 0.03099} = 0.6471$$

Temos então que:

$$\begin{array}{ll} p(k=1 \mid x_1) = 0.7428\,, & p(k=2 \mid x_1) = 0.2572 \\ p(k=1 \mid x_2) = 0.1558\,, & p(k=2 \mid x_2) = 0.8442 \\ p(k=1 \mid x_3) = 0.3529\,, & p(k=2 \mid x_3) = 0.6471 \end{array}$$

- **Passo M**:

$$N_k = \sum_{i=1}^{n} \gamma_{ik}\,; \quad u_k = \frac{1}{N_k}\sum_{i=1}^{n} \gamma_{ik} x_i\,; \quad \sum_k = \frac{1}{N_k}\sum_{i=1}^{n} \gamma_{ik}(x_i - u_k)(x_i - u_k)^T\,; \quad \pi_k = p(c_k) = \frac{N_k}{n};$$

Para k=1:

$$N_1 = \sum_{i=1}^{3} \gamma_{i1} = p(k=1 \mid x_1) + p(k=1 \mid x_2) + p(k=1 \mid x_3) = 0.7428 + 0.1558 + 0.3529 = 1.2515\,;$$

$$u_1 = \frac{1}{N_1}\sum_{i=1}^{3} \gamma_{i1} x_i = \frac{1}{N_1}(p(k=1 \mid x_1)\,x_1 + p(k=1 \mid x_2)\,x_2 + p(k=1 \mid x_3)\,x_3) =$$
$$= \frac{1}{1.2515}(0.7428\begin{bmatrix}1\\2\end{bmatrix} + 0.1558\begin{bmatrix}-1\\1\end{bmatrix} + 0.3529\begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}0.7510\\1.3115\end{bmatrix};$$

$$\Sigma_1 = \frac{1}{N_1}\sum_{i=1}^{3} \gamma_{i1}(x_i - u_1)(x_i - u_1)^T =$$
$$\frac{1}{1.2515}(0.7428\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0.7510\\1.3115\end{bmatrix}\right)\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0.7510\\1.3115\end{bmatrix}\right)^T + 0.1558\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.7510\\1.3115\end{bmatrix}\right)\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.7510\\1.3115\end{bmatrix}\right)^T +$$
$$0.3529\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.7510\\1.3115\end{bmatrix}\right)\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.7510\\1.3115\end{bmatrix}\right)^T =$$
$$\frac{1}{1.2515}\left(\begin{bmatrix}0.04605 & 0.1273\\0.1273 & 0.3521\end{bmatrix} + \begin{bmatrix}0.4777 & 0.08498\\0.08498 & 0.01511\end{bmatrix} + \begin{bmatrix}0.02188 & -0.1152\\-0.1152 & 0.6070\end{bmatrix}\right) = \begin{bmatrix}0.4360 & 0.07756\\0.07756 & 0.07784\end{bmatrix};$$

$$\pi_1 = p(k=1) = \frac{N_1}{3} = \frac{1.2515}{3} = 0.4172\,;$$

Para k=2:

$$N_2 = \sum_{i=1}^{3} \gamma_{i2} = p(k=2 \mid x_1) + p(k=2 \mid x_2) + p(k=2 \mid x_3) = 0.2572 + 0.8442 + 0.6471 = 1.7485\,;$$

$$u_2 = \frac{1}{N_2}\sum_{i=1}^{3} \gamma_{i2} x_i = \frac{1}{N_2}(p(k=2 \mid x_1)\,x_1 + p(k=2 \mid x_2)\,x_2 + p(k=2 \mid x_3)\,x_3) =$$
$$= \frac{1}{1.7485}(0.2572\begin{bmatrix}1\\2\end{bmatrix} + 0.8442\begin{bmatrix}-1\\1\end{bmatrix} + 0.6471\begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}0.03437\\0.7770\end{bmatrix};$$

$$\Sigma_2 = \frac{1}{N_2}\sum_{i=1}^{3}\gamma_{i2}(x_i - u_2)(x_i - u_2)^T =$$

$$\frac{1}{1.7485}(0.2572\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0.03437\\0.7770\end{bmatrix}\right)\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0.03437\\0.7770\end{bmatrix}\right)^T + 0.8442\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.03437\\0.7770\end{bmatrix}\right)\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.03437\\0.7770\end{bmatrix}\right)^T +$$

$$0.6471\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.03437\\0.7770\end{bmatrix}\right)\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.03437\\0.7770\end{bmatrix}\right)^T =$$

$$\frac{1}{1.7485}\left(\begin{bmatrix}0.2398 & 0.3037\\0.3037 & 0.3847\end{bmatrix} + \begin{bmatrix}0.9032 & -0.1947\\-0.1947 & 0.04198\end{bmatrix} + \begin{bmatrix}0.6034 & -0.4855\\-0.4855 & 0.3907\end{bmatrix}\right) = \begin{bmatrix}0.9988 & -0.2153\\-0.2153 & 0.4674\end{bmatrix};$$

$$\pi_2 = p(k=2) = \frac{N_2}{3} = \frac{1.7485}{3} = 0.5828\,;$$

E então, temos que:

$$u_1 = \begin{bmatrix}0.7510\\1.3115\end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix}0.4360 & 0.07756\\0.07756 & 0.7784\end{bmatrix}, \quad \pi_1 = 0.4172 \text{ ❒}$$

$$u_2 = \begin{bmatrix}0.03437\\0.7770\end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix}0.9988 & -0.2153\\-0.2153 & 0.4674\end{bmatrix}, \quad \pi_2 = 0.5828 \text{ ❒}$$

TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 018

2) Given the updated parameters computed in previous question:

   a. [1.5v] perform a hard assignment of observations to clusters under a MAP assumption.

   b. [2.5v] compute the silhouette of the larger cluster using the Euclidean distance.

---

a. $MAP = argmax_c \;\; p(c = P|x) \quad \rightarrow \quad argmax_c \;\; p(x|c = P)p(c = P)$

$$\det(\sum_1) = (0.4360 \times 0.7784) - (0.07756 \times 0.07756) = 0.3334\,;$$
$$\det(\sum_2) = (0.9988 \times 0.4674) - (-0.2153 \times -0.2153) = 0.4205\,;$$

$$(\sum_1)^{-1} = \begin{bmatrix} 2.3350 & -0.2327 \\ -0.2327 & 1.3079 \end{bmatrix}, \quad (\sum_2)^{-1} = \begin{bmatrix} 1.1116 & 0.5120 \\ 0.5120 & 2.3754 \end{bmatrix};$$

$$u_1 = \begin{bmatrix} 0.7510 \\ 1.3115 \end{bmatrix}, \quad u_2 = \begin{bmatrix} 0.03437 \\ 0.7770 \end{bmatrix}, \quad \pi_1 = 0.4172\,, \quad \pi_2 = 0.5828\,.$$

Para x_1:

$$p(x_1 \mid \mu_1, \sum_1) = \frac{1}{2\pi} \times \frac{1}{\sqrt{0.3334}} \times \exp\left(-\frac{1}{2}\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0.7510\\1.3115\end{bmatrix}\right)^T \begin{bmatrix} 2.3350 & -0.2327 \\ -0.2327 & 1.3079 \end{bmatrix}\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0.7510\\1.3115\end{bmatrix}\right)\right) = 0.1957$$

$$p(x_1 \mid \mu_2, \sum_2) = \frac{1}{2\pi} \times \frac{1}{\sqrt{0.4205}} \times \exp\left(-\frac{1}{2}\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0.03437\\0.7770\end{bmatrix}\right)^T \begin{bmatrix} 1.1116 & 0.5120 \\ 0.5120 & 2.3754 \end{bmatrix}\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0.03437\\0.7770\end{bmatrix}\right)\right) = 0.01351$$

$$p(k = 1 \mid x_1) = p(x_1 \mid k = 1)p(k = 1) = 0.1957 \times 0.4172 = 0.08165$$
$$p(k = 2 \mid x_1) = p(x_1 \mid k = 2)p(k = 2) = 0.01351 \times 0.5828 = 0.007873$$

Normalizando:

$$p(k = 1 \mid x_1) = \frac{0.08165}{0.08165 + 0.007873} = 0.9121$$
$$p(k = 2 \mid x_1) = \frac{0.007873}{0.08165 + 0.007873} = 0.08794$$

Para x_2:

$$p(x_2 \mid \mu_1, \sum_1) = \frac{1}{2\pi} \times \frac{1}{\sqrt{0.3334}} \times \exp\left(-\frac{1}{2}\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.7510\\1.3115\end{bmatrix}\right)^T \begin{bmatrix} 2.3350 & -0.2327 \\ -0.2327 & 1.3079 \end{bmatrix}\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.7510\\1.3115\end{bmatrix}\right)\right) = 0.008190$$

$$p(x_2 \mid \mu_2, \sum_2) = \frac{1}{2\pi} \times \frac{1}{\sqrt{0.4205}} \times \exp\left(-\frac{1}{2}\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.03437\\0.7770\end{bmatrix}\right)^T \begin{bmatrix} 1.1116 & 0.5120 \\ 0.5120 & 2.3754 \end{bmatrix}\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.03437\\0.7770\end{bmatrix}\right)\right) = 0.1437$$

TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 018

$$p(k=1 \mid x_2) = p(x_2 \mid k=1)p(k=1) = 0.008190 \times 0.4172 = 0.003417$$
$$p(k=2 \mid x_2) = p(x_2 \mid k=2)p(k=2) = 0.1437 \times 0.5828 = 0.08375$$

Normalizando:

$$p(k=1 \mid x_2) = \frac{0.003417}{0.003417 + 0.08375} = 0.03920$$
$$p(k=2 \mid x_2) = \frac{0.08375}{0.003417 + 0.08373} = 0{,}9608$$

Para x_3:

$$p(x_3 \mid \mu_1, \sum_1) = \frac{1}{2\pi} \times \frac{1}{\sqrt{0.3334}} \times \exp\left(-\frac{1}{2}\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.7510\\1.3115\end{bmatrix}\right)^T \begin{bmatrix}2.3350 & -0.2327\\-0.2327 & 1.3079\end{bmatrix}\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.7510\\1.3115\end{bmatrix}\right)\right) = 0.07716$$

$$p(x_3 \mid \mu_2, \sum_2) = \frac{1}{2\pi} \times \frac{1}{\sqrt{0.4205}} \times \exp\left(-\frac{1}{2}\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.03437\\0.7770\end{bmatrix}\right)^T \begin{bmatrix}1.1116 & 0.5120\\0.5120 & 2.3754\end{bmatrix}\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.03437\\0.7770\end{bmatrix}\right)\right) = 0.1048$$

$$p(k=1 \mid x_1) = p(x_1 \mid k=1)p(k=1) = 0.07716 \times 0.4172 = 0.03219$$
$$p(k=2 \mid x_1) = p(x_1 \mid k=2)p(k=2) = 0.1048 \times 0.5828 = 0.06108$$

Normalizando:

$$p(k=1 \mid x_3) = \frac{0.03219}{0.03219 + 0.06108} = 0.3451$$
$$p(k=2 \mid x_3) = \frac{0.06108}{0.03219 + 0.06108} = 0.6549$$

Temos então que:

$$p(k=1 \mid x_1) = 0.9121\,, \quad p(k=2 \mid x_1) = 0.08794$$
$$p(k=1 \mid x_2) = 0.03920, \quad p(k=2 \mid x_2) = 0.9608$$
$$p(k=1 \mid x_3) = 0.3451\,, \quad p(k=2 \mid x_3) = 0.6549$$

Logo, podemos atribuir as observações aos clusters da seguinte maneira:

$$k=1 : \{x_1\}\,, \quad k=2 : \{x_2\,, x_3\}\ \square$$

TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 018

b. $Silhouette \rightarrow S(x_i) = 1 - \frac{a(x_i)}{b(x_i)} \quad (se\ a < b)\,, \quad S(x_i) = \frac{b(x_i)}{a(x_i)} - 1 \quad c.c$

$a(x_i) = média\ das\ distâncias\ de\ x_i\ aos\ pontos\ do\ cluster$
$b(x_i) = \min(média\ das\ distâncias\ de\ x_i\ aos\ pontos\ de\ outro\ cluster)$

Sendo k=2 o maior cluster (maior número de observações):

$$d(x_1\,, x_2) = \sqrt{(1-(-1))^2 + (2-1)^2} = \sqrt{5}$$
$$d(x_1\,, x_3) = \sqrt{(1-1)^2 + (2-0)^2} = 2$$
$$d(x_2\,, x_3) = \sqrt{(-1-1)^2 + (1-0)^2} = \sqrt{5}$$

$$a(x_2) = \frac{d(x_2\,, x_3)}{1} = \sqrt{5}\,, \quad b(x_2) = \min\left(\frac{d(x_2\,, x_1)}{1}\right) = \sqrt{5}\,;$$

$$a(x_3) = \frac{d(x_3\,, x_2)}{1} = \sqrt{5}\,, \quad b(x_3) = \min\left(\frac{d(x_3\,, x_1)}{1}\right) = 2.$$

Como $a(x_2) \geq b(x_2)$ e $a(x_3) \geq b(x_3)$, temos:

$$S(x_2) = \frac{b(x_2)}{a(x_2)} - 1 = \frac{\sqrt{5}}{\sqrt{5}} - 1 = 0\,;$$

$$S(x_3) = \frac{b(x_3)}{a(x_3)} - 1 = \frac{2}{\sqrt{5}} - 1 = -0.1056\,;$$

$$S(k=2) = \frac{S(x_2) + S(x_3)}{2} = \frac{0 - 0.1056}{2} = -0.05280\ \square$$

Aprendizagem 2022/23

Homework IV – Group 018

II. Programming and critical analysis [9v]

Recall the *pd_speech.arff* dataset from earlier homeworks, centered on the Parkinson diagnosis from speech features. For the following exercises, normalize the data using sklearn's MinMaxScaler.

1) [4.5v] Using sklearn, apply k-means clustering fully unsupervisedly (without targets) on the normalized data with $k = 3$ and three different seeds (using random ε {0,1,2}). Assess the silhouette and purity of the produced solutions.

	Random seed = 0	Random seed = 1	Random seed = 2
Silhouette (Euclidian distance)	0.11362027575179431	0.11403554201377074	0.11362027575179431
Silhouette (Manhattan distance)	0.1440871604300737	0.14400051258512941	0.1440871604300737
Purity	0.7671957671957672	0.7632275132275133	0.7671957671957672

2) [1.5v] What is causing the non-determinism?

O não-determinismo do processo K-means vai depender da inicialização aleatória dos centroides, tendo em atenção o parâmetro *random_seed:*

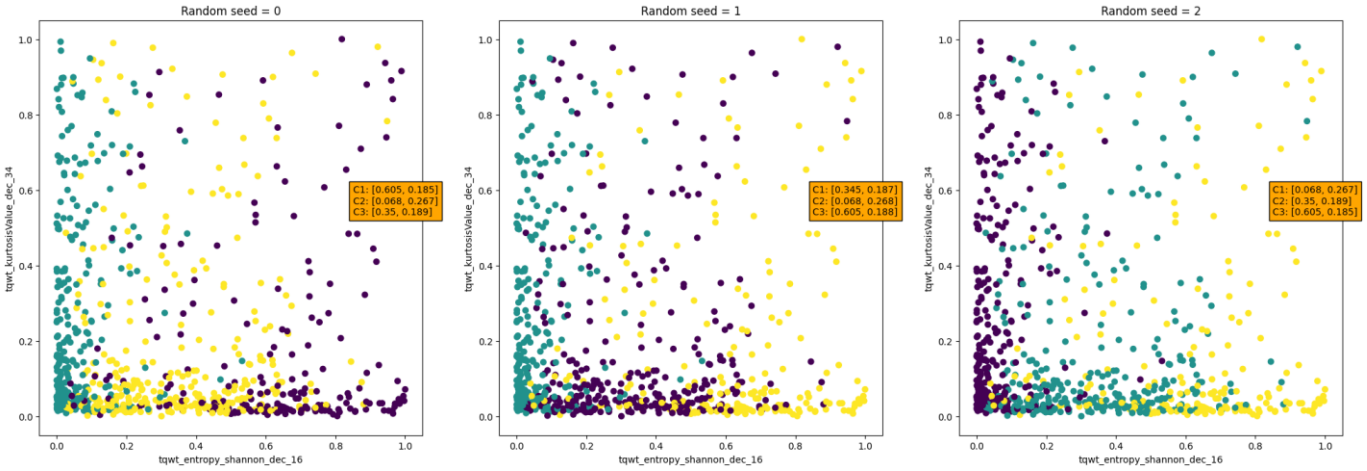
- Se *random_seed* = número inteiro, a aleatoriedade dos pontos gerados para os centroides será determinística;
- Caso contrário, não será determinística.

Como no caso estudado aplicamos K-means ao nosso dataset 3 vezes, cada uma com uma seed diferentes (0, 1 e 2), os centroides de cada execução do algoritmo são inicializados em pontos diferentes, originando, portanto, uma evolução distinta do modelo e, por sua vez, resultados diferentes, explicando por isso o não-determinismo.

Analisando os valores obtidos, reparamos numa inconsistência com a conclusão anterior, visto que os resultados com seed = 0 e seed = 2 são iguais. Uma possível justificação para tal é a diferença entre os pontos escolhidos para os centroides iniciais ser mínima (ou porventura até serem os mesmos), levando assim a resultados iguais, ou com diferenças mínimas muito difíceis de reparar.

No caso de seed = 1, os valores já se encontram de acordo com o não-determinismo quando comparados com os resultados das restantes seeds.

3) [1.5v] Using a scatter plot, visualize side-by-side the labeled data using as labels: i) the original Parkinson diagnoses, and ii) the previously learned $k = 3$ clusters (random= 0). To this end, select the two most informative features as axes and color observations according to their label. For feature selection, select the two input variables with highest variance on the MinMax normalized data.



As duas variáveis escolhidas para os eixos (baseadas no critério de feature selection indicado) foram:

- *tqwt_entropy_shannon_dec_16* (coluna 371);
- *tqwt_kurtosisValue_dec_34* (coluna 749).
-

4) [1.5v] The fraction of variance explained by a principal component is the ratio between the variance of that component (i.e., its eigenvalue) and total variance (i.e., sum of all eigenvalues). How many principal components are necessary to explain more than 80% of variability? Hint: explore the DimReduction notebook to be familiar with PCA in sklearn.

Da execução do código, obtemos o seguinte output:

Not-sorted: explained variance of: 0.8006422402169661 - with 31 PCs
Sorted: explained variance of: 0.8074139593238812 - with 751 PCs

- No caso do array de explained variances estar ordenado de forma crescente (Sorted), consideramos, para atingir os 80% de variância explicada, os vetores próprios que menos variância explicam (de menor valor próprio) e por isso, obtemos um número tão grande de principal componentes (751)

- Caso contrário, obtemos o número mínimo de principal components para atingir 80% de variância explicada (31), sendo que o array está agora ordenado de forma decrescente (Not-sorted).

TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 018

Appendix

```
# Import wall
from scipy.io.arff import loadarff

import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

from sklearn.preprocessing import MinMaxScaler
from sklearn import metrics, cluster
from sklearn.decomposition import PCA

# Question 1

# ------------------------------------------------------------ #

# Load and prepare data
data = loadarff('pd_speech.arff')
df = pd.DataFrame(data[0])

df['class'] = df['class'].str.decode('utf-8')

X_init = df.drop("class", axis=1)

# Normalize data.
scaler = MinMaxScaler()
X = scaler.fit_transform(X_init)

# ------------------------------------------------------------ #

possible_seed_values = [0, 1, 2]
# seed = np.random.RandomState.set_state(possible_seed_values)

cluster_points = {0: [], 1: [], 2: []}
cluster_centroids = {0: [], 1: [], 2: []}
```

```
for seed in possible_seed_values:
    # Parameterize clustering
    kmeans_algo = cluster.KMeans(n_clusters = 3, random_state = seed)


    # Fit the model to our data.
    kmeans_model = kmeans_algo.fit(X)

    # Get the produced clusters and their centroids.
    y_pred = kmeans_model.labels_.tolist()
    y_centroid = kmeans_model.cluster_centers_

    cluster_points[seed] = y_pred
    cluster_centroids[seed] = y_centroid

# ------------------------------------------------------------ #

# Compute purities.
def purity_score(y_true, y_pred):
    # Compute contingency/confusion matrix
    confusion_matrix = metrics.cluster.contingency_matrix(y_true, y_pred)
    return np.sum(np.amax(confusion_matrix, axis=0)) / np.sum(confusion_matrix)

# Get ground truth
y_true = df['class']

# Compute silhouettes and purity.
for i in range(0,3):
    print("Random seed = " + str(i) + ":")
    print(" - Silhouette (Euclidian distance):",
          metrics.silhouette_score(X, cluster_points[i], metric='euclidean'))
    print(" - Silhouette (Manhattan distance):",
          metrics.silhouette_score(X, cluster_points[i], metric='manhattan'))
    print(" - Purity:", purity_score(y_true, cluster_points[i]))
```

TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 018

```
# Question 3

def get_highest_var(input_var, n):
    """Returns the indexes of the n highest variance variables of the
    input array 'input_var'."""

    # Get the variance of every variable.
    variance = np.var(input_var, axis = 0)

    res = []

    # Get index of the n highest variance variables.
    for i in range(0, n):
        max = variance.argmax(axis = 0)

        # Invalidates the entry we just got so the indexes stay consistent.
        variance[max] = -2

        res.append(max)

    return res

# Get the two highest variance variables.
highest_vars = get_highest_var(X, 2)

# Plot multiple scatter plots.
fig, (ax1) = plt.subplots(1,3, figsize = (25, 8))

plot_titles = ["Random seed = 0", "Random seed = 1", "Random seed = 2"]

# Set the plot titles.
for col in range(0,3):
    ax1[col].set_title(plot_titles[col])


# Data used to analyze the plots:
x_data, y_data = highest_vars[0], highest_vars[1]

x, y = X[:, x_data], X[:, y_data]
x_col, y_col = X_init.columns[x_data], X_init.columns[y_data]

print(f"{x_col} (index {x_data}).")
print(f"{y_col} (index {y_data}).")
```

```
for col in range(0,3):
    ax1[col].scatter(x, y, c = cluster_points[col])
    ax1[col].set_xlabel(x_col)
    ax1[col].set_ylabel(y_col)

    c1_coords = "C1: " + str([np.round(cluster_centroids[col][0][x_data], 3),
                          np.round(cluster_centroids[col][0][y_data], 3)])
    c2_coords = "C2: " + str([np.round(cluster_centroids[col][1][x_data], 3),
                          np.round(cluster_centroids[col][1][y_data], 3)])
    c3_coords = "C3: " + str([np.round(cluster_centroids[col][2][x_data], 3),
                          np.round(cluster_centroids[col][2][y_data], 3)])

    centroid_coords = c1_coords + "\n" + c2_coords + "\n" + c3_coords

    ax1[col].text(0.85, 0.535, centroid_coords, fontname = "Sans",
                  bbox = dict(facecolor = "orange", alpha = 1))


# Question 4.
# Learn the transformation and fit the model to out data.
pca = PCA()
pca.fit(X)

# Get the explained variance ratio of each principal component.
explained_variance = np.array(pca.explained_variance_ratio_)

# Sort the array.
sorted_explained_variance = np.sort(explained_variance)

threshold = 0.8

counter = 0
sum = 0


# Loop over the array until we hit the threshold.
for var in explained_variance:
    sum += var
    counter += 1

    if sum >= threshold:
        print(f"Not-sorted: explained variance of: {sum} - with {counter} PCs")
        break
```

```
# Loop over the sorted array until we hit the threshold.
for var in sorted_explained_variance:
    sum += var
    counter += 1

    if sum >= threshold:
        print(f"Sorted: explained variance of: {sum} - with {counter} PCs")
        break
```