TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 015 - Hugo Silva, Pedro Florindo

I. Pen-and-paper

1) Answer 1

Given the bivariate observations $\{\binom{1}{2}, \binom{-1}{1}, \binom{1}{0}\}$,

and the multivariate Gaussian mixture

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

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).

$\hat{\mu}$ → média

$\hat{\Sigma}$ → Matriz de covariância

Fórmula da Normal / Gaussiana

$$\frac{1}{2\pi\sqrt{\det(\Sigma)}} \times \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right]$$

$$X_1 = \binom{1}{2} \qquad X_2 = \binom{-1}{1} \qquad X_3 = \binom{1}{0}$$

X_1

$$P(X_1 \mid k=1) \sim \mathcal{N}\left(\mu_1 = \binom{2}{2}, \Sigma_1 \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\right) =$$

$$= \frac{1}{2\pi\sqrt{\det(\Sigma_1)}} \times e^{-\frac{1}{2}(X_1-\mu_1)^T \Sigma_1^{-1} (X_1-\mu_1)}$$

$$= \frac{1}{2\pi \times \sqrt{3}} \times e^{-\frac{1}{2}\left(\begin{bmatrix} -1 & 0 \end{bmatrix} \cdot \frac{1}{3} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} -1 \\ 0 \end{bmatrix} \right)}$$

$$= \frac{1}{2\pi\sqrt{3}} \times e^{-\frac{1}{2}\left(\begin{bmatrix} -\frac{2}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} -1 \\ 0 \end{bmatrix} \right)}$$

$$= \frac{1}{2\pi\sqrt{3}} \times e^{-\frac{1}{2}\cdot\frac{8}{3}} = \frac{1}{2\pi\sqrt{3}} e^{-\frac{1}{3}} \approx 0.065841$$

$$P(X_1 \mid K=2) \sim \mathcal{N}\left(\mu_2 = \begin{bmatrix}0\\0\end{bmatrix}, \Sigma_2 = \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}\right) =$$

$$= \frac{1}{2\pi \times \sqrt{4}} \; e^{-\frac{1}{2}\left(\begin{bmatrix}1 & 2\end{bmatrix} \cdot \frac{1}{4}\begin{bmatrix}2 & 0\\0 & 2\end{bmatrix} \cdot \begin{bmatrix}1\\2\end{bmatrix}\right)}$$

$$= \frac{1}{4\pi} \; e^{-\frac{1}{2}\left(\begin{bmatrix}\frac{1}{2} & 1\end{bmatrix} \begin{bmatrix}1\\2\end{bmatrix}\right)} = \frac{1}{4\pi} e^{-\frac{1}{2}\times\frac{5}{2}}$$

$$\approx 0.022799$$

Posteriors

posterior $(K=1)$ $= 0.065841 \times 0.5 = 0.0329205$

posterior $(K=2)$ $= 0.022799 \times 0.5 = 0.0113995$

Normalization

$$P(K=1 \mid X_1) = \frac{0.0329205}{0.0329205 + 0.0113995} = 0.742789$$

$$P(K=2 \mid X_1) = 1 - P(K=1 \mid X_1) = 0.257211$$

TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 015 - Hugo Silva, Pedro Florindo

X_2

$$X_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$P(X_2 \mid K=1) \sim \mathcal{N}\left(\mu_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \Sigma_1 \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\right) =$$

$$= \frac{1}{2\pi \times \sqrt{3}} \times e^{-\frac{1}{2}\left(\begin{bmatrix} -3 & -1 \end{bmatrix} \cdot \frac{1}{3} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} -3 \\ -1 \end{bmatrix} \right)}$$

$$= \frac{1}{2\pi\sqrt{3}} \times e^{-\frac{1}{2}\left(\begin{bmatrix} -\frac{5}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} -3 \\ -1 \end{bmatrix} \right)}$$

$$= \frac{1}{2\pi\sqrt{3}} e^{-\frac{1}{2} \times \frac{14}{3}} \approx 0.008911$$

$$P(X_2 \mid K=2) \sim \mathcal{N}\left(\mu_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}\right) =$$

$$= \frac{1}{2\pi \times \sqrt{4}} e^{-\frac{1}{2}\left(\begin{bmatrix} -1 & 1 \end{bmatrix} \cdot \frac{1}{4} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right)}$$

$$= \frac{1}{4\pi} e^{-\frac{1}{2}\left(\begin{bmatrix} -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right)}$$

$$= \frac{1}{4\pi} e^{-\frac{1}{2} \cdot 1} \approx 0.048266$$

TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 015 - Hugo Silva, Pedro Florindo

Posteriors

posterior $(k=1) = 0.002911 \times 0.5 = 0.0044555$

posterior $(k=2) = 0.048266 \times 0.5 = 0.024133$

Normalization

$$P(k=1 \mid x_2) = \frac{0.0044555}{0.0044555 + 0.024133} \approx 0.155840$$

$$P(k=2 \mid x_2) = 1 - P(k=1 \mid x_2) \approx 0.844160$$

x_3

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

$$P(x_3 \mid k=1) \sim \mathcal{N}\left(\mu_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \Sigma_1 \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\right)$$

$$= \frac{1}{2\pi \times \sqrt{3}} \times e^{-\frac{1}{2}\left(\begin{bmatrix} -1 & -2 \end{bmatrix} \cdot \frac{1}{3} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} -1 \\ -2 \end{bmatrix} \right)}$$

$$= \frac{1}{2\pi\sqrt{3}} \times e^{-\frac{1}{2}\left(\begin{bmatrix} 0 & -1 \end{bmatrix} \begin{bmatrix} -1 \\ -2 \end{bmatrix} \right)}$$

$$= \frac{1}{2\pi\sqrt{3}} e^{-\frac{1}{2} \times 2} \approx 0.033804$$

Aprendizagem 2022/23

Homework IV – Group 015 - Hugo Silva, Pedro Florindo

$$P(X_3 \mid K=2) \sim \mathcal{N}\left(\mu_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}\right) =$$

$$= \frac{1}{2\pi \times \sqrt{4}} \; e^{-\frac{1}{2}\left(\begin{bmatrix} 1 & 0 \end{bmatrix} \cdot \frac{1}{4} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right)}$$

$$= \frac{1}{4\pi} \; e^{-\frac{1}{2}\left(\begin{bmatrix} \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right)}$$

$$= \frac{1}{4\pi} \; e^{-\frac{1}{2} \times \frac{1}{2}} \approx 0.061975$$

Posteriors

posterior $(K=1) = 0.033804 \times 0.5 = 0.016902$

posterior $(K=2) = 0.061975 \times 0.5 = 0.0309875$

Normalization

$$P(K=1 \mid X_3) = \frac{0.016902}{0.016902 + 0.0309875} \approx 0.352935$$

$$P(K=2 \mid X_3) = 1 - 0.352935 = 0.647065$$

TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 015 - Hugo Silva, Pedro Florindo

M - Step

$$\gamma_{ik} = P(C_k | X_i) = \frac{P(X_i | C_k) \cdot P(C_k)}{P(X_i)}$$

$$N_k = \sum_i \gamma_{ik}$$

$$\rightarrow N_1 = 0.742789 + 0.155840 + 0.352935$$

$$= 1.251564$$

$$\rightarrow N_2 = 0.257211 + 0.844160 + 0.647065$$

$$= 1.748436$$

$$\mu_k = \frac{1}{N_k} \left(\sum_i \gamma_{ik} X_i \right)$$

$$\rightarrow \mu_1 = \frac{1}{1.251564} \left(0.742789\, X_1 + 0.155840\, X_2 + 0.352935\, X_3 \right)$$

$$= \frac{1}{1.251564} \left(0.742789 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 0.155840 \begin{bmatrix} -1 \\ 1 \end{bmatrix} + 0.352935 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right)$$

$$= \frac{1}{1.251564} \begin{bmatrix} 0.939884 \\ 1.644418 \end{bmatrix} = \begin{bmatrix} 0.750968 \\ 1.311493 \end{bmatrix}$$

$$\rightarrow \mu_2 = \frac{1}{1.748436} \left(0.257211 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 0.844160 \begin{bmatrix} -1 \\ 1 \end{bmatrix} + 0.647065 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right)$$

$$= \frac{1}{1.748436} \begin{bmatrix} 0.060116 \\ 1.358582 \end{bmatrix} = \begin{bmatrix} 0.034383 \\ 0.777027 \end{bmatrix}$$

TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 015 - Hugo Silva, Pedro Florindo

$$\Sigma_k = \frac{1}{N_k}\left(\sum_i \gamma_{ik}\,(x_i-\mu_k)(x_i-\mu_k)^T\right)$$

$$\rightarrow \Sigma_1 = \frac{1}{1.251564}\left(0.742789\left(\begin{bmatrix}1\\2\end{bmatrix}-\begin{bmatrix}0.750968\\1.311493\end{bmatrix}\right)\left(\begin{bmatrix}1\\2\end{bmatrix}-\begin{bmatrix}0.750968\\1.311493\end{bmatrix}\right)^T +\right.$$

$$+\ 0.155840\left(\begin{bmatrix}-1\\1\end{bmatrix}-\begin{bmatrix}0.750968\\1.311493\end{bmatrix}\right)\left(\begin{bmatrix}-1\\1\end{bmatrix}-\begin{bmatrix}0.750968\\1.311493\end{bmatrix}\right)^T +$$

$$\left.+\ 0.352935\left(\begin{bmatrix}1\\0\end{bmatrix}-\begin{bmatrix}0.750968\\1.311493\end{bmatrix}\right)\left(\begin{bmatrix}1\\0\end{bmatrix}-\begin{bmatrix}0.750968\\1.311493\end{bmatrix}\right)^T\right)$$

$$\Sigma_1 = \frac{1}{1.251564}\left(0.742789\begin{bmatrix}0.062017 & 0.171460\\ 0.171460 & 0.474042\end{bmatrix} +\right.$$

$$+\ 0.155840\begin{bmatrix}3.065889 & 0.545414\\ 0.545414 & 0.097028\end{bmatrix} +$$

$$\left.+\ 0.352935\begin{bmatrix}0.062017 & -0.326604\\ -0.326604 & 1.719112\end{bmatrix}\right)$$

$$\Sigma_1 = \begin{bmatrix}0.436048 & 0.077572\\ 0.077572 & 0.778201\end{bmatrix}$$

Aprendizagem 2022/23

Homework IV – Group 015 - Hugo Silva, Pedro Florindo

$$\rightarrow \Sigma_2 = \frac{1}{1.748436}\left(0.257211\left(\begin{bmatrix}1\\2\end{bmatrix}-\begin{bmatrix}0.034383\\0.777027\end{bmatrix}\right)\left(\begin{bmatrix}1\\2\end{bmatrix}-\begin{bmatrix}0.034383\\0.777027\end{bmatrix}\right)^T + \right.$$

$$+\ 0.844160\left(\begin{bmatrix}-1\\1\end{bmatrix}-\begin{bmatrix}0.034383\\0.777027\end{bmatrix}\right)\left(\begin{bmatrix}-1\\1\end{bmatrix}-\begin{bmatrix}0.034383\\0.777027\end{bmatrix}\right)^T +$$

$$\left.+\ 0.647065\left(\begin{bmatrix}1\\0\end{bmatrix}-\begin{bmatrix}0.034383\\0.777027\end{bmatrix}\right)\left(\begin{bmatrix}1\\0\end{bmatrix}-\begin{bmatrix}0.034383\\0.777027\end{bmatrix}\right)^T\right)$$

$$\Sigma_2 = \frac{1}{1.748436}\left(0.257211\begin{bmatrix}0.932416 & 1.180924\\ 1.180924 & 1.495663\end{bmatrix} + \right.$$

$$+\ 0.844160\begin{bmatrix}1.069948 & -0.230639\\ -0.230639 & 0.049717\end{bmatrix} +$$

$$\left.+\ 0.647065\begin{bmatrix}0.932416 & -0.750310\\ -0.750310 & 0.603771\end{bmatrix}\right)$$

$$\Sigma_2 = \begin{bmatrix}0.998818 & -0.215306\\ -0.215306 & 0.467474\end{bmatrix}$$

$$\pi_k = \frac{N_k}{N}$$

$$\pi_1 = \frac{1.251564}{3} = 0.417188$$

$$\pi_2 = \frac{1.748436}{3} = 0.582812$$

TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 015 - Hugo Silva, Pedro Florindo

2) Answer 2

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)

$$\Sigma_1 = \begin{bmatrix} 0.436048 & 0.077572 \\ 0.077572 & 0.778201 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 0.998818 & -0.215306 \\ -0.215306 & 0.467474 \end{bmatrix}$$

$$\mu_1 = \begin{bmatrix} 0.750968 \\ 1.311493 \end{bmatrix} \qquad \mu_2 = \begin{bmatrix} 0.034383 \\ 0.777027 \end{bmatrix}$$

$$\pi_1 = \frac{1.251564}{3} = 0.417188 \qquad \pi_2 = \frac{1.748436}{3} = 0.582812$$

$$\Sigma_1^{-1} = \frac{1}{0.333316} \begin{bmatrix} 0.778201 & -0.077572 \\ -0.077572 & 0.436048 \end{bmatrix} \qquad \Sigma_2^{-1} = \frac{1}{0.420565} \begin{bmatrix} 0.467474 & 0.215306 \\ 0.215306 & 0.998818 \end{bmatrix}$$

$$\Sigma_1^{-1} = \begin{bmatrix} 2.334724 & -0.232728 \\ -0.232728 & 1.308212 \end{bmatrix} \qquad \Sigma_2^{-1} = \begin{bmatrix} 1.111538 & 0.511945 \\ 0.511945 & 2.374943 \end{bmatrix}$$

X_1

$$P(X_1 \mid k=1) \sim \mathcal{N}(\mu_1, \Sigma_1) =$$

$$= \frac{1}{2\pi\sqrt{\det(\Sigma_1)}} \times e^{-\frac{1}{2}(X_1-\mu_1)^T \Sigma_1^{-1} (X_1-\mu_1)}$$

$$= \frac{1}{2\pi \times \sqrt{0.333316}} \times e^{-\frac{1}{2}\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0.750968\\1.311493\end{bmatrix}\right)^T \cdot \begin{bmatrix}2.334724 & -0.232728\\-0.232728 & 1.308212\end{bmatrix} \left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0.750968\\1.311493\end{bmatrix}\right)}$$

≈ 0.125712

$$P(X_1 \mid k=2) \sim \mathcal{N}(\mu_2, \Sigma_2) =$$

$$= \frac{1}{2\pi\sqrt{0.420565}} \times e^{-\frac{1}{2}\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0.034383\\0.777027\end{bmatrix}\right)^T \begin{bmatrix}1.111538 & 0.511945\\0.511945 & 2.374943\end{bmatrix}\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}0.034383\\0.777027\end{bmatrix}\right)}$$

≈ 0.013519

X_2

$$P(X_2 \mid k=1) \sim \mathcal{N}(\mu_1, \Sigma_1) =$$

$$= \frac{1}{2\pi \times \sqrt{0.333316}} \times e^{-\frac{1}{2}\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.750968\\1.311493\end{bmatrix}\right)^T \cdot \begin{bmatrix}2.334724 & -0.232728\\-0.232728 & 1.308212\end{bmatrix}\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.750968\\1.311493\end{bmatrix}\right)}$$

≈ 0.008196

$$P(X_2 \mid k=2) \sim \mathcal{N}(\mu_2, \Sigma_2) =$$

$$= \frac{1}{2\pi\sqrt{0.420565}} \times e^{-\frac{1}{2}\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.034383\\0.777027\end{bmatrix}\right)^T \begin{bmatrix}1.111538 & 0.511945\\0.511945 & 2.374943\end{bmatrix}\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.034383\\0.777027\end{bmatrix}\right)}$$

≈ 0.143646

TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 015 - Hugo Silva, Pedro Florindo

X_3

$$P(X_3 \mid k=1) \sim \mathcal{N}(\mu_1, \Sigma_1) =$$

$$\simeq \frac{1}{2\pi \times \sqrt{0.333316}} \times e^{-\frac{1}{2}\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.750968\\1.311493\end{bmatrix}\right)^T \cdot \begin{bmatrix}2.334724 & -0.232728\\-0.232728 & 1.308212\end{bmatrix} \left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.750968\\1.311493\end{bmatrix}\right)}$$

$$\simeq 0.077148$$

$$P(X_3 \mid k=2) \sim \mathcal{N}(\mu_2, \Sigma_2) =$$

$$\simeq \frac{1}{2\pi \sqrt{0.420565}} \times e^{-\frac{1}{2}\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.034383\\0.777027\end{bmatrix}\right)^T \begin{bmatrix}1.111538 & 0.511945\\0.511945 & 2.374943\end{bmatrix} \left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}0.034383\\0.777027\end{bmatrix}\right)}$$

$$\simeq 0.104784$$

Posteriors

X_1

$P(k=1 \mid X_1) = 0.195712 \times 0.417188 \simeq \boxed{0.081649}$

$P(k=2 \mid X_1) = 0.013519 \times 0.582812 \simeq 0.007879$

X_1 pertence ao cluster 1

X_2

$P(k=1 \mid X_2) = 0.008196 \times 0.417188 \simeq 0.003419$

$P(k=2 \mid X_2) = 0.143646 \times 0.582812 \simeq \boxed{0.083718}$

X_2 pertence ao cluster 2

X_3

$P(K=1 \mid X_3) = 0.077148 \times 0.417188 \approx 0.032185$

$P(K=2 \mid X_3) = 0.104784 \times 0.582812 \approx 0.061070$

X_3 pertence ao cluster 2.

b)

$a(X_1) = 0$ (X_1 é a única observação que pertence ao K=1)

$$S(X_1) = 1 - \frac{\frac{d(X_1, X_1)}{1}}{\frac{d(X_1, X_2) + d(X_1, X_3)}{2}} = 0$$

$$S(X_2) = 1 - \frac{\frac{d(X_2, X_3)}{1}}{\frac{d(X_2, X_1)}{1}}$$

$$= 1 - \frac{\sqrt{(-1-1)^2 + (1-0)^2}}{\sqrt{(1+1)^2 + (2-1)^2}} = 0$$

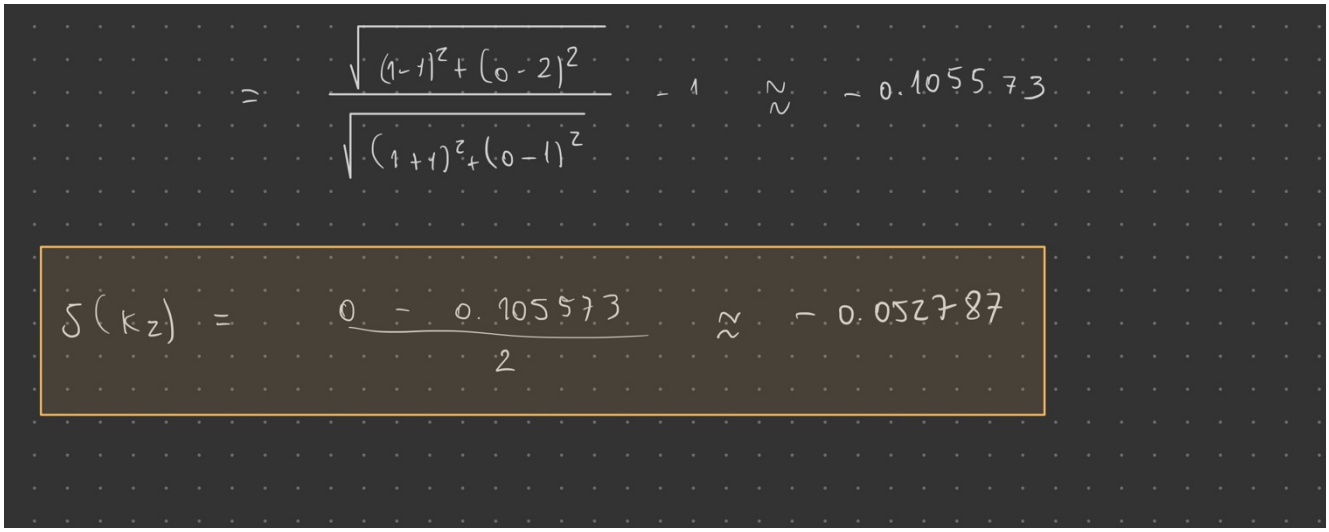
$$S(X_3) = \frac{\frac{d(X_3, X_1)}{1}}{\frac{d(X_3, X_2)}{1}} - 1$$

$\frac{b(X_3)}{a(X_3)} - 1$

pois $a(X_3) > b(X_3)$

Aprendizagem 2022/23

Homework IV – Group 015 - Hugo Silva, Pedro Florindo



II. Programming and critical analysis

1) Answer 1

Silhouettes and purities respectively:

Seed = 0 → 0.11362027575179431; 0.7671957671957672

Seed = 1 → 0.11403554201377074; 0.7632275132275133

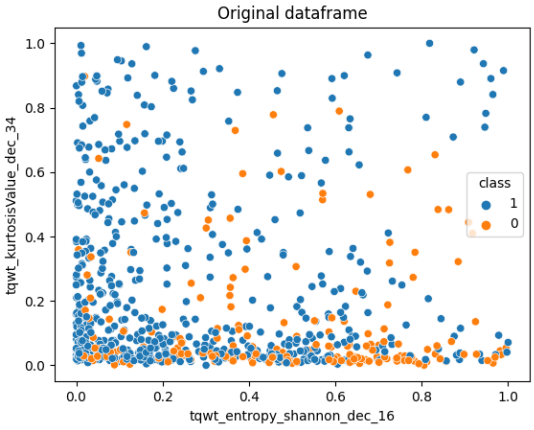
Seed = 2 → 0.11362027575179431; 0.7671957671957672

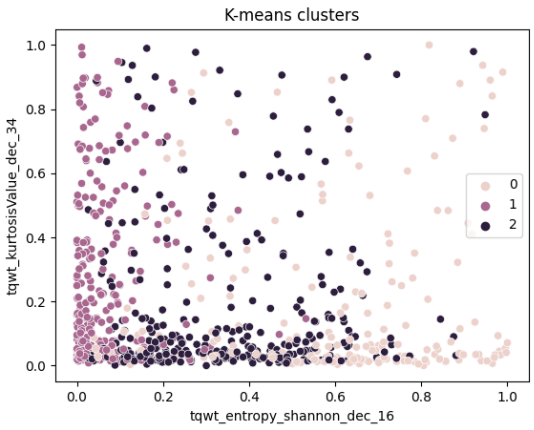
```
Silhouette scores:  [0.11362027575179431, 0.11403554201377074, 0.11362027575179431]
Purity scores:  [0.7671957671957672, 0.7632275132275133, 0.7671957671957672]
```

2) Answer 2

What is causing the non-determinism is the random initialization of the centroids of the clusters. For different seeds, the algorithm converges to different local minimums. Therefore, the obtained results are different, given enough decimal places (for seed = 0 and seed = 2 we obtain approximately the same result for 16 decimal places).

3) Answer 3





4) Answer 4

The number of principal components required to explain more than 80% of variability is 31.

```
Number of principal components: 31
```

III. APPENDIX

```python
import pandas as pd
from scipy.io.arff import loadarff
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import metrics, datasets
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
from sklearn import cluster
import numpy as np
from sklearn.feature_selection import VarianceThreshold
from sklearn.decomposition import PCA


data = loadarff('pd_speech.arff')
df = pd.DataFrame(data[0])
df['class'] = df['class'].str.decode('utf-8')

X = df.drop('class', axis=1)

scaler = MinMaxScaler()
scaledDF = scaler.fit_transform(df)
scaledDF = pd.DataFrame(scaledDF, columns=df.columns)

print("SHAPE:", scaledDF.shape)

scaledDF = scaledDF.drop(columns=['class'])

labels = []
silhouettes = []
purity = []

def purity_score(y_true, y_pred):
    confusion_matrix = metrics.cluster.contingency_matrix(y_true, y_pred)
    return np.sum(np.amax(confusion_matrix, axis=0)) / np.sum(confusion_matrix)

for i in range(3):
    kmeans = KMeans(n_clusters = 3, random_state = i).fit(scaledDF)
    y_pred = kmeans.labels_
    labels.append(y_pred)
    silhouettes.append(metrics.silhouette_score(scaledDF, y_pred, metric='euclidean'))
```

```
    purity.append(purity_score(df["class"], y_pred))

print("Silhouette scores: ", silhouettes)
print("Purity scores: ", purity)

variances = scaledDF.var()

variances = variances.sort_values(ascending=False)
variances = variances[:2]

sns.scatterplot(x = scaledDF[variances.index[0]], y = scaledDF[variances.index[1]], hue =
df["class"])
plt.title("Original dataframe")
plt.savefig("OG_DF.png")
plt.show()

sns.scatterplot(x = scaledDF[variances.index[0]], y = scaledDF[variances.index[1]], hue = labels[0])
plt.title("K-means clusters")
plt.savefig("kmeans.png")
plt.show()

#print(variances)

#print(len(df_scaled["tqwt_kurtosisValue_dec_34"]))

#print(len(df_scaled["tqwt_entropy_shannon_dec_16"]))

pca = PCA(n_components=0.8)
pca.fit(scaledDF)
print("Number of principal components:", pca.n_components_)
```

END