Aprendizagem 2024/25

Homework IV – Group 066

(ist106794, ist1107301)

I. Pen-and-paper

1)

Homewor IV

1.

$x_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad x_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix} \quad x_3 = \begin{bmatrix} 3 \\ -1 \end{bmatrix} \quad \pi_1 = 0{,}5 \quad \pi_2 = 0{,}5$

$\mu_1 = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

$$p(x_1 | c_1) = \frac{1}{2\pi \times \left|\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}\right|^{1/2}} e^{-1/2\left(\left([1\ 0] - [2\ -1]\right)\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}^{-1}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 2 \\ -1 \end{bmatrix}\right)\right)}$$

$$= \frac{1}{2\pi \times \sqrt{15}} e^{-1/30\left([-1\ 1]\begin{bmatrix} 4 & -1 \\ -1 & 4 \end{bmatrix}\begin{bmatrix} -1 \\ 1 \end{bmatrix}\right)}$$

$$= \frac{1}{2\pi\sqrt{15}} e^{-1/3} = 0{,}029$$

$$p(x_1 | c_2) = \frac{1}{2\pi \times \left|\begin{matrix} 2 & 0 \\ 0 & 2 \end{matrix}\right|^{1/2}} e^{-1/2\left(\left([1\ 0] - [1\ 1]\right)\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}^{-1}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)\right)}$$

$$= \frac{1}{2\pi \times 2} e^{-2/8} = 0{,}062$$

$$\gamma_{11} = \frac{0{,}029 \times 0{,}5}{0{,}5 \times 0{,}029 + 0{,}5 \times 0{,}062} = 0{,}319$$

$$\gamma_{21} = \frac{0{,}062 \times 0{,}5}{0{,}5 \times 0{,}062 + 0{,}5 \times 0{,}029} = 0{,}681$$

$$p(x_2 | c_1) = \frac{1}{2\pi \times \left|\begin{matrix} 4 & 1 \\ 1 & 4 \end{matrix}\right|^{1/2}} e^{-1/2\left(\left([0\ 2] - [2\ -1]\right)\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}^{-1}\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix} - \begin{bmatrix} 2 \\ -1 \end{bmatrix}\right)\right)}$$

$$= \frac{1}{2\pi\sqrt{15}} e^{-1/30 \times 64} = 0{,}005$$

Homework IV – Group 066

(ist106794, ist1107301)

$$P(x_2 | C_2) = \frac{1}{2\pi \times 2} e^{-1/2\left(\left([0\ 2]-[1\ 1]\right)\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}^{-1}\left(\begin{bmatrix}0\\2\end{bmatrix}-\begin{bmatrix}1\\1\end{bmatrix}\right)\right)}$$

$$= \frac{1}{4\pi} e^{-1/8 \times 4} = 0,048$$

$$\gamma_{12} = \frac{0,005 \times 0,5}{0,5 \times (0,005 + 0,048)} = 0,094$$

$$\gamma_{22} = \frac{0,048 \times 0,5}{0,5 (0,005 + 0,048)} = 0,906$$

— // —

$$P(x_3 | C_1) = \frac{1}{2\pi\sqrt{15}} e^{-1/2\left(\left([3\ -1]-[2\ -1]\right)\begin{bmatrix}4 & 1\\ 1 & 4\end{bmatrix}^{-1}\left(\begin{bmatrix}3\\-1\end{bmatrix}-\begin{bmatrix}2\\-1\end{bmatrix}\right)\right)}$$

$$= 0,035$$

$$P(x_3 | C_2) = \frac{1}{2\pi \times 2} e^{-1/2\left(\left([3\ -1]-[1\ 1]\right)\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}^{-1}\left(\begin{bmatrix}3\\-1\end{bmatrix}-\begin{bmatrix}1\\1\end{bmatrix}\right)\right)}$$

$$= 0,011$$

$$\gamma_{13} = \frac{0,035 \times 0,5}{0,5 \times (0,035 + 0,011)} = 0,766$$

$$\gamma_{23} = \frac{0,011 \times 0,5}{0,5 \times (0,011 + 0,035)} = 0,234$$

— // —

$$N_1 = 0,319 + 0,094 + 0,766 = 1,179$$

$$N_2 = 0,681 + 0,906 + 0,234 = 1,821$$

$$\mu_1 = \frac{1}{1,179} \times \left(0,319 \times \begin{bmatrix}1\\0\end{bmatrix} + 0,094 \times \begin{bmatrix}0\\2\end{bmatrix} + 0,766 \times \begin{bmatrix}3\\-1\end{bmatrix}\right)$$

$$= \begin{bmatrix}2,220\\-0,450\end{bmatrix}$$

$$\mu_2 = \frac{1}{1,821} \times \left(0,681 \times \begin{bmatrix}1\\0\end{bmatrix} + 0,906 \times \begin{bmatrix}0\\2\end{bmatrix} + 0,234 \times \begin{bmatrix}3\\-1\end{bmatrix}\right) = \begin{bmatrix}0,759\\0,866\end{bmatrix}$$

$$\Sigma_1 = \frac{1}{1,179}\left(0,319 \times \left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}2,220\\-0,490\end{bmatrix}\right)\left(\begin{bmatrix}1 & 0\end{bmatrix} - \begin{bmatrix}2,220 & -0,490\end{bmatrix}\right) + 0,094 \times \left(\begin{bmatrix}0\\2\end{bmatrix} - \begin{bmatrix}2,220\\-0,490\end{bmatrix}\right)\left(\begin{bmatrix}0 & 2\end{bmatrix} - \begin{bmatrix}2,220 & -0,490\end{bmatrix}\right) + 0,766 \times \left(\begin{bmatrix}3\\-1\end{bmatrix} - \begin{bmatrix}2,220\\-0,490\end{bmatrix}\right)\left(\begin{bmatrix}3 & -1\end{bmatrix} - \begin{bmatrix}2,220 & -0,490\end{bmatrix}\right)\right)$$

$$= \frac{1}{1,179} \times \left(0,319 \times \begin{bmatrix}-1,220\\0,490\end{bmatrix}\begin{bmatrix}-1,220 & 0,490\end{bmatrix} + 0,094 \times \begin{bmatrix}-2,220\\2,490\end{bmatrix}\begin{bmatrix}-2,220 & 2,490\end{bmatrix} + 0,766 \times \begin{bmatrix}0,780\\-0,510\end{bmatrix}\begin{bmatrix}0,780 & -0,510\end{bmatrix}\right)$$

$$= \frac{1}{1,179}\left(0,319 \times \begin{bmatrix}1,488 & -0,598\\-0,598 & 0,240\end{bmatrix} + 0,094 \times \begin{bmatrix}4,928 & -5,528\\-5,528 & 6,200\end{bmatrix} + 0,766 \times \begin{bmatrix}0,608 & -0,358\\-0,398 & 0,260\end{bmatrix}\right)$$

$$= \boxed{\begin{bmatrix}1,190 & -0,861\\-0,861 & 0,728\end{bmatrix}}$$

$$\Sigma_2 = \frac{1}{1,821} \times \left(0,681 \times \begin{bmatrix}0,241\\-0,866\end{bmatrix}\begin{bmatrix}0,241 & -0,866\end{bmatrix} + 0,906 \times \begin{bmatrix}-0,759\\1,134\end{bmatrix}\begin{bmatrix}-0,759 & 1,134\end{bmatrix} + 0,234 \times \begin{bmatrix}2,241\\-1,866\end{bmatrix}\begin{bmatrix}2,241 & -1,866\end{bmatrix}\right)$$

$$= \frac{1}{1,821} \times \left(0,681 \times \begin{bmatrix}0,058 & -0,209\\-0,209 & 0,750\end{bmatrix} + 0,906 \times \begin{bmatrix}0,576 & -0,861\\-0,861 & 1,286\end{bmatrix} + 0,234 \times \begin{bmatrix}5,022 & -4,182\\-4,182 & 3,482\end{bmatrix}\right)$$

$$= \boxed{\begin{bmatrix}0,954 & -1,044\\-1,044 & 1,368\end{bmatrix}}$$

Aprendizagem 2024/25

Homework IV – Group 066

(ist106794, ist1107301)

$$\boxed{\begin{aligned}\pi_1 &= \frac{1,179}{1,179+1,821} = 0,393 \\ \pi_2 &= \frac{1,821}{1,821+1,179} = 0,607\end{aligned}}$$

— // —

2nd Epoch

$$P(x_1|c_1) = \frac{1}{2\pi \times \sqrt{0,125}} e^{-1/2\left(\left([1\ 0]-[2,22\ \ -0,49]\right)\begin{bmatrix}1,190 & -0,861\\ -0,861 & 0,728\end{bmatrix}^{-1}\left(\begin{bmatrix}1\\0\end{bmatrix}-\begin{bmatrix}1,2\\-0,\end{bmatrix}\right.\right.}$$

$$= 0,116$$

$$P(x_1|c_2) = \frac{1}{2\pi \times \sqrt{0,215}} e^{-1/2\left(\left([1\ 0]-[0,755\ \ 0,866]\right)\begin{bmatrix}0,654 & -1,044\\ -1,044 & 1,1368\end{bmatrix}^{-1}\left(\begin{bmatrix}1\\0\end{bmatrix}-\begin{bmatrix}0,759\\0,866\end{bmatrix}\right.\right.}$$

$$= 0,149$$

$$\gamma_{11} = \frac{0,116 \times 0,393}{0,116 \times 0,393 + 0,149 \times 0,607} = 0,335$$

$$\gamma_{21} = \frac{0,149 \times 0,607}{0,149 \times 0,607 + 0,116 \times 0,393} = 0,665$$

— // —

$$P(x_2|c_1) = \frac{1}{2\pi \times \sqrt{0,125}} e^{-1/2\left(\left([0, 2]-[2,22\ \ -0,49]\right)\begin{bmatrix}1,190 & -0,861\\ -0,861 & 0,728\end{bmatrix}\left(\begin{bmatrix}0\\2\end{bmatrix}-\begin{bmatrix}2,22\\-0,49\end{bmatrix}\right)\right)}$$

$$= 0,001$$

$$P(x_2|c_2) = \frac{1}{2\pi \times \sqrt{0,215}} e^{-1/2\left(\left([0\ 2]-[0,755\ \ 0,866]\right)\begin{bmatrix}0,654 & -1,044\\ -1,044 & 1,368\end{bmatrix}^{-1}\left(\begin{bmatrix}0\\2\end{bmatrix}-\begin{bmatrix}0,759\\0,866\end{bmatrix}\right.\right.}$$

$$= 0,207$$

$$\gamma_{12} = \frac{0,001 \times 0,393}{0,001 \times 0,393 + 0,207 \times 0,607} = 0,003$$

$$\gamma_{22} = \frac{0,207 \times 0,607}{0,207 \times 0,607 + 0,001 \times 0,393} = 0,997$$

1)

$$P(x_3 \mid C_1) = \frac{1}{2\pi\sqrt{0{,}173}}\, e^{-1/2\left(\left([3\ \ -1]-[2{,}21\ \ 0{,}45]\right)\begin{bmatrix}1{,}190 & -0{,}861\\ -0{,}861 & 0{,}928\end{bmatrix}^{-1}\left(\begin{bmatrix}3\\-1\end{bmatrix}-\begin{bmatrix}2{,}21\\0{,}45\end{bmatrix}\right)\right)}$$

$$= 0{,}344$$

$$P(x_3 \mid C_2) = \frac{1}{2\pi\sqrt{0{,}215}}\, e^{-1/2\left(\left([3\ \ -1]-[0{,}759\ \ 0{,}866]\right)\begin{bmatrix}0{,}954 & -1{,}044\\ -1{,}044 & 1{,}368\end{bmatrix}^{-1}\left(\begin{bmatrix}3\\-1\end{bmatrix}-\begin{bmatrix}0{,}759\\0{,}866\end{bmatrix}\right)\right)}$$

$$= 0{,}012$$

$$\gamma_{13} = \frac{0{,}344 \times 0{,}393}{0{,}344 \times 0{,}393 + 0{,}012 \times 0{,}607} = 0{,}949$$

$$\gamma_{23} = \frac{0{,}012 \times 0{,}607}{0{,}012 \times 0{,}607 + 0{,}344 \times 0{,}393} = 0{,}051$$

— // —

$$N_1 = 0{,}335 + 0{,}003 + 0{,}949 = 1{,}287$$

$$N_2 = 0{,}665 + 0{,}997 + 0{,}051 = 1{,}713$$

$$\boxed{\begin{aligned}
\mu_1 &= \frac{1}{1{,}287}\left(0{,}335 \times \begin{bmatrix}1\\0\end{bmatrix} + 0{,}003 \times \begin{bmatrix}0\\2\end{bmatrix} + 0{,}949 \times \begin{bmatrix}3\\-1\end{bmatrix}\right) = \begin{bmatrix}2{,}472\\-0{,}733\end{bmatrix}\\
\mu_2 &= \frac{1}{1{,}713}\left(0{,}665 \times \begin{bmatrix}1\\0\end{bmatrix} + 0{,}997 \begin{bmatrix}0\\2\end{bmatrix} + 0{,}051 \times \begin{bmatrix}3\\-1\end{bmatrix}\right) = \begin{bmatrix}0{,}478\\1{,}134\end{bmatrix}
\end{aligned}}$$

$$\Sigma_1 = \frac{1}{1{,}287}\Bigg(0{,}335 \times \begin{bmatrix}-1{,}472\\0{,}733\end{bmatrix}\begin{bmatrix}-1{,}472 & 0{,}733\end{bmatrix} + 0{,}003 \times \begin{bmatrix}-2{,}472\\2{,}733\end{bmatrix}\begin{bmatrix}-2{,}472 & 2{,}733\end{bmatrix} + 0{,}949 \times \begin{bmatrix}0{,}528\\-0{,}267\end{bmatrix}\begin{bmatrix}0{,}528 & -0{,}267\end{bmatrix}\Bigg)$$

$$= \frac{1}{1{,}287}\Bigg(0{,}335 \times \begin{bmatrix}2{,}167 & -1{,}079\\-1{,}079 & 0{,}537\end{bmatrix} + 0{,}003 \times \begin{bmatrix}6{,}111 & -6{,}756\\-6{,}756 & 7{,}469\end{bmatrix} + 0{,}949 \times \begin{bmatrix}0{,}279 & -0{,}141\\-0{,}141 & 0{,}071\end{bmatrix}\Bigg)$$

$$= \boxed{\begin{bmatrix}0{,}784 & -0{,}400\\-0{,}400 & 0{,}210\end{bmatrix}}$$

$$\Sigma_2 = \frac{1}{1{,}713} \times \Bigg(0{,}665 \begin{bmatrix}0{,}522\\-1{,}134\end{bmatrix}\begin{bmatrix}0{,}522 & -1{,}134\end{bmatrix} + 0{,}557 \begin{bmatrix}-0{,}478\\0{,}866\end{bmatrix}\begin{bmatrix}-0{,}478 & 0{,}866\end{bmatrix} + 0{,}051 \begin{bmatrix}2{,}522\\-2{,}134\end{bmatrix}\begin{bmatrix}2{,}522 & -2{,}134\end{bmatrix}\Bigg)$$

$$= \boxed{\begin{bmatrix}0{,}428 & -0{,}631\\-0{,}631 & 1{,}071\end{bmatrix}}$$

$$\boxed{\begin{aligned}\pi_1 &= \frac{1{,}287}{1{,}713+1{,}287} = 0{,}429\\ \pi_2 &= \frac{1{,}713}{1{,}713+1{,}287} = 0{,}571\end{aligned}}$$

Aprendizagem 2024/25

Homework IV – Group 066

(ist106794, ist1107301)

2)

a)

2.

a)

x_1:

$$P(x_1|c_1) = \frac{1}{2\pi\sqrt{0,005}} e^{-\frac{1}{2}\left(\left([1\ 0] - [2,472\ \ -0,733]\right)\begin{bmatrix}0,784 & -0,4\\ -0,4 & 0,21\end{bmatrix}^{-1}\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}2,472\\-0,733\end{bmatrix}\right)\right)}$$

$$= 0,571$$

$$P(x_1|c_2) = \frac{1}{2\pi\sqrt{0,060}} e^{-\frac{1}{2}\left(\left([1\ 0] - [0,478\ \ 1,134]\right)\begin{bmatrix}0,428 & -0,631\\ -0,631 & 1,071\end{bmatrix}^{-1}\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0,478\\1,134\end{bmatrix}\right)\right)}$$

$$= 0,112$$

$$\gamma_{11} = \frac{0,429 \times 0,571}{0,429 \times 0,571 + 0,571 \times} = 0,793$$

$$\boxed{x_1 \in C_1}$$

$$\gamma_{21} = \frac{0,571 \times 0,112}{0,571 \times 0,112 + 0,429 \times 0,571} = 0,207$$

x_2:

$$P(x_2|c_1) = \frac{1}{2\pi\sqrt{0,005}} e^{-\frac{1}{2}\left(\left([0\ 2] - [2,472\ \ -0,733]\right)\begin{bmatrix}0,784 & -0,4\\ -0,4 & 0,21\end{bmatrix}^{-1}\left(\begin{bmatrix}0\\2\end{bmatrix} - \begin{bmatrix}2,472\\-0,733\end{bmatrix}\right)\right)}$$

$$= 0,000$$

$$P(x_2|c_2) = \frac{1}{2\pi\sqrt{0,060}} e^{-\frac{1}{2}\left(\left([0\ 2] - [0,478\ \ 1,134]\right)\begin{bmatrix}0,428 & -0,631\\ -0,631 & 1,071\end{bmatrix}^{-1}\left(\begin{bmatrix}0\\2\end{bmatrix} - \begin{bmatrix}0,478\\1,134\end{bmatrix}\right)\right)}$$

$$= 0,004$$

$$\gamma_{12} = \frac{0,000 \times 0,429}{0,000 \times 0,429 + 0,004 \times 0,571} = 0$$

$$\boxed{x_2 \in C_2}$$

$$\gamma_{22} = \frac{0,004 \times 0,571}{0,000 \times 0,429 + 0,004 \times 0,571} = 1$$

$$P(x_3 \mid C_1) = \frac{1}{2\pi\sqrt{0{,}005}}\, e^{-1/2\left(\left([3\ \ -1] - [2{,}172\ \ -0{,}783]\right)\begin{bmatrix} 1{,}781 & -0{,}4 \\ -0{,}4 & 0{,}21 \end{bmatrix}\left(\begin{bmatrix} 3 \\ -1 \end{bmatrix} - \begin{bmatrix} 2{,}172 \\ -0{,}783 \end{bmatrix}\right)\right)}$$

$$= 1{,}955$$

$$P(x_3 \mid C_2) = \frac{1}{2\pi\sqrt{0{,}060}}\, e^{-1/2\left(\left([3\ \ -1] - [0{,}418\ \ 1{,}134]\right)\begin{bmatrix} 0{,}418 & -0{,}671 \\ -0{,}671 & 1{,}071 \end{bmatrix}^{-1}\left(\begin{bmatrix} 3 \\ -1 \end{bmatrix} - \begin{bmatrix} 0{,}478 \\ 1{,}134 \end{bmatrix}\right)\right)}$$

$$= 0{,}000$$

$$\gamma_{13} = \frac{0{,}429 \times 1{,}955}{0{,}429 \times 1{,}955 + 0{,}571 \times 0} = 1 \quad \boxed{x_3 \in C_1}$$

$$\gamma_{23} = \frac{0{,}571 \times 0}{0{,}571 \times 0 + 0{,}429 \times 1{,}955} = 0$$

b)

b)

$$\|x_1, x_3\|_2 = \sqrt{(1-3)^2 + (0-(-1))^2} = \sqrt{4+1} = \sqrt{5}$$

$$\|x_1, x_2\|_2 = \sqrt{(1-0)^2 + (0-2)^2} = \sqrt{1+4} = \sqrt{5}$$

$$s(x_1) = 1 - \frac{a(x_1)}{b(x_1)} = 1 - \frac{\|x_1, x_3\|_2}{\|x_1, x_2\|_2} = 1 - \frac{\sqrt{5}}{\sqrt{5}} = 0$$

$$\|x_2, x_3\|_2 = \sqrt{(0-3)^2 + (2-(-1))^2} = \sqrt{9+9} = \sqrt{18} = 3\sqrt{2}$$

$$s(x_3) = 1 - \frac{a(x_3)}{b(x_3)} = 1 - \frac{\sqrt{5}}{3\sqrt{2}} = 1 - 0{,}527 = 0{,}473$$

$$s(C_1) = \frac{s(x_1) + s(x_3)}{2} = \frac{0 + 0{,}473}{2} \boxed{= 0{,}236}$$

Aprendizagem 2024/25

Homework IV – Group 066

(ist106794, ist1107301)

II. Programming and critical analysis

Here we have an overview of all the imports used and how we imported the dataset, which are common to all questions. Therefore, we will be omitting this code from the beginning of each task to avoid redundancy

```
import pandas as pd, matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA

# Load the data
df = pd.read_csv('accounts.csv')
X = df.drop('deposit', axis=1)
X = pd.get_dummies(X, drop_first=True)
X = X.iloc[:, :8].drop_duplicates().dropna()
Y = pd.get_dummies(df['deposit'], drop_first=True)
```

1)

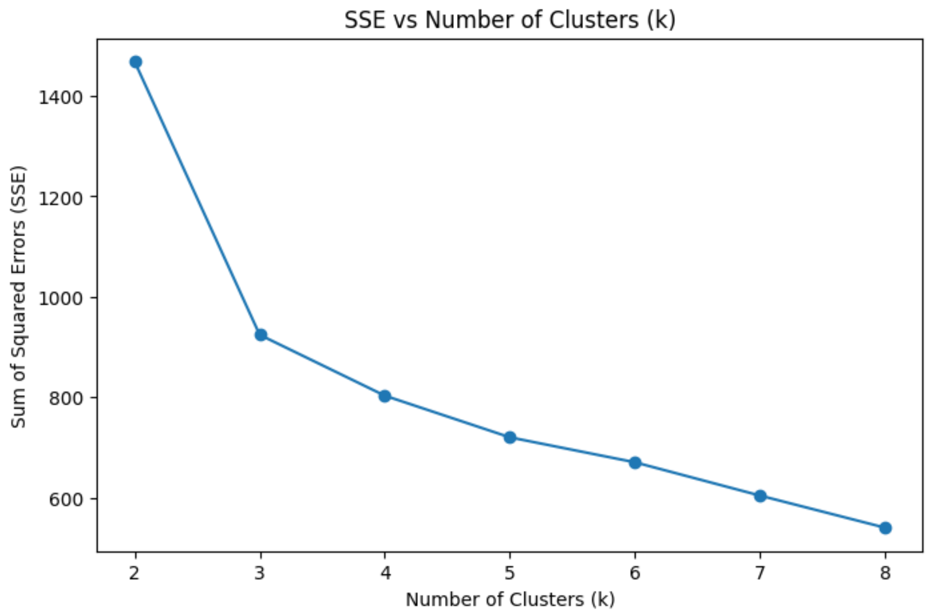
a) Normalizing the data and applying k-means for each k value

```
# Normalize the data using MinMaxScaler
scaler = MinMaxScaler()
X_normalized = scaler.fit_transform(X)

sse = []
k_values = range(2, 9)

# Apply k-means clustering for each k value
for k in k_values:
    kmeans = KMeans(n_clusters=k, max_iter=500, random_state=42)
    kmeans.fit_predict(X_normalized)
    sse.append(kmeans.inertia_)
```

Plotting the result

```
# Plot the SSE for each k
plt.figure(figsize=(8, 5))
plt.plot(k_values, sse, marker='o')
plt.title('SSE vs Number of Clusters (k)')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Sum of Squared Errors (SSE)')
plt.show()
```

Output:



b) According to the plot, and using the knee/elbow finding method we come to the conclusion that the ideal number of clusters is 3. This method consists of finding the point where the decrease in SSE becomes less accentuated with the increase of clusters.

For all values before this point, the model might suffer from underfitting and for values above, the increase in the number of clusters and complexity does not provide substantial improvements to the model. Not only does it not justify, it also contributes to overfitting the data.

c) K-modes is an adaptation of k-means used to better handle categorical features. It uses the mode (most frequent value) instead of means to represent the centroid of the cluster and uses distance based in dissimilarity (like Hamming Distance) to group the data.

Given that our dataset's features are predominantly categorical (10 categorical out of 17), in theory, k-modes would be a better clustering approach, considering the explanation above.

Aprendizagem 2024/25

Homework IV – Group 066

(ist106794, ist1107301)

2)

a) Normalizing the data with StandardScaler, applying the PCA and getting the explained variance ratio for each component

```
scaler = StandardScaler()
X_normalized = scaler.fit_transform(X)

# Apply PCA
pca = PCA(n_components=2)
pca.fit(X_normalized)

variance_explained = pca.explained_variance_ratio_

print("The first component explains", round(variance_explained[0], 5), "variability in the data set.")
print("The second component explains", round(variance_explained[1], 5), "variability in the data set.")
print("The top 2 components explain", round(variance_explained[1] + variance_explained[0], 5), "variability in the data set.")
```

The first component explains 0.19568 variability in the data set.

The second component explains 0.14462 variability in the data set.

The top 2 components explain 0.3403 variability in the dataset.

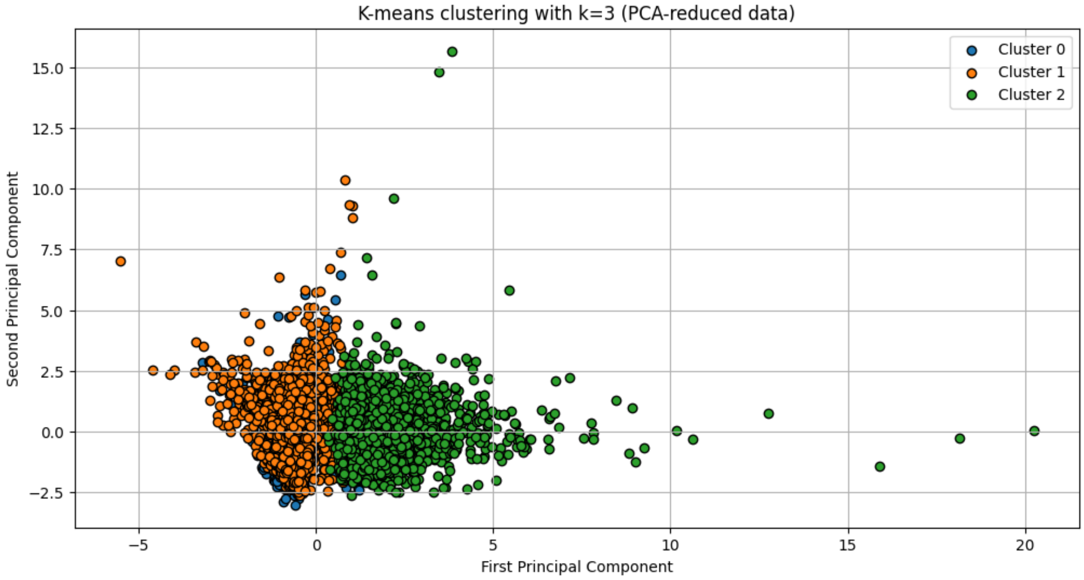
b) Applying k-means clustering and PCA

```
kmeans = KMeans(n_clusters=3, random_state=42)
labels = kmeans.fit_predict(X_normalized)

X_projected = pca.fit_transform(X_normalized)
```

Plotting the results:

```
plt.figure(figsize=(12, 6))
for cluster in range(3):
    plt.scatter(X_projected[labels == cluster, 0], X_projected[labels == cluster, 1], label=f'Cluster {cluster}', edgecolors='k')

plt.title('K-means clustering with k=3 (PCA-reduced data)')
plt.xlabel('First Principal Component')
plt.ylabel('Second Principal Component')
plt.grid()
plt.legend()
plt.show()
```

O utput:



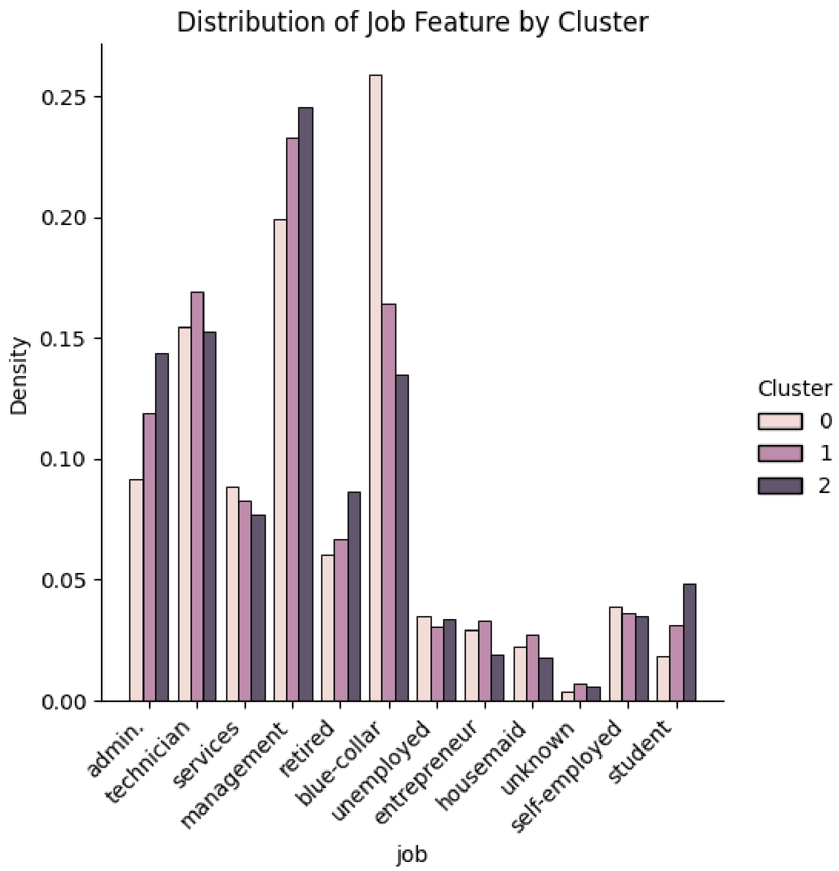
We can only clearly separate two of the three clusters (the yellow one and the magenta one). The third one (dark blue) is very hard to distinguish from the other two as it underlies both of them.

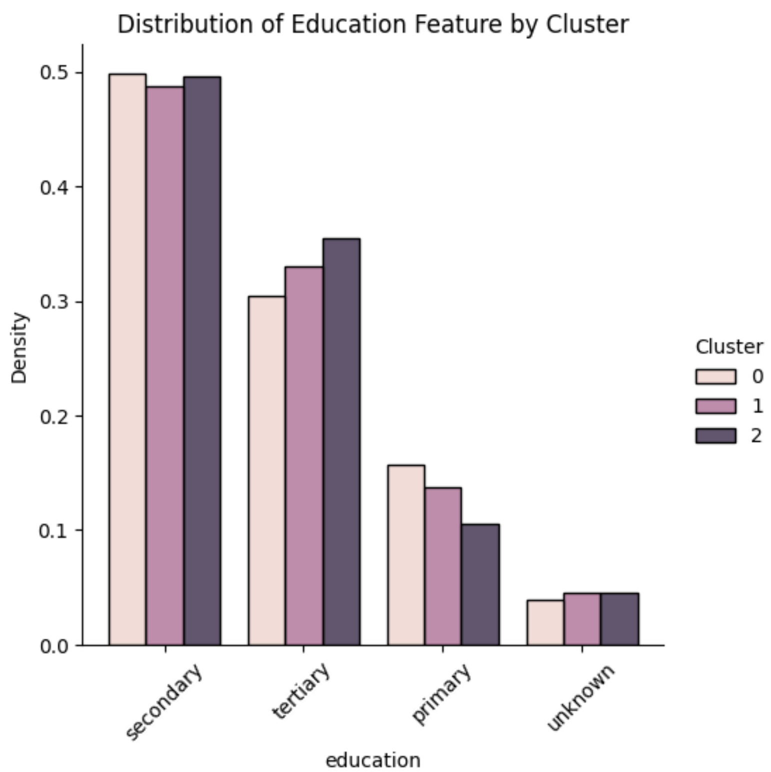
This happens because we are losing a lot of information on the data variance by only considering the top two components to draw the plot. The fact that there is significant overlap between some clusters suggests that using only these two components in a 2D space isn't enough to represent a lot of the variance between the different observations of the dataset, that might become more apparent in higher dimensional spaces.

c) Plotting both graphs

```
df['Cluster'] = labels

# Plot the 'job' feature distribution across clusters
sns.displot(df, x="job", hue="Cluster", multiple="dodge", stat='density', shrink=0.8, common_norm=False)
plt.title('Distribution of Job Feature by Cluster')
plt.xticks(rotation=45, ha = 'right')
plt.show()

# Plot the 'education' feature distribution across clusters
sns.displot(df, x="education", hue="Cluster", multiple="dodge", stat='density', shrink=0.8, common_norm=False)
plt.title('Distribution of Education Feature by Cluster')
plt.xticks(rotation=45)
plt.show()
```

Output:


Distribution of Job Feature by Cluster
Density
0.00
0.05
0.10
0.15
0.20
0.25
admin.
technician
services
management
retired
blue-collar
unemployed
entrepreneur
housemaid
unknown
self-employed
student
job
Cluster
0
1
2


Distribution of Education Feature by Cluster
Density
0.0
0.1
0.2
0.3
0.4
0.5
secondary
tertiary
primary
unknown
education
Cluster
0
1
2

The main differences in the obtained plots can be described as follows:

- Jobs
  - Cluster 0: Jobs like "blue-collar", "services", "unemployed" and "self-employed" appear mostly in this cluster. They can be classified as less qualified positions and working-class jobs.
  - Cluster 1: Jobs like "technician", "entrepreneur", "housemaid" and "unknown" appear mostly in this cluster. We can't find a significant relationship between these jobs, which makes us believe this cluster is better described by other features.
  - Cluster 2: Jobs like "admin", "management", "retired" and "student" appear mostly in this cluster. They describe either people who are not working or highly qualified jobs

- Education
  - Cluster 0: Mostly represented by "primary" and "secondary" levels of education, which translates well to the conclusions we made for the job distribution.
  - Cluster 1: Mostly represented by "unknown" level of education, which can also be a good indicator that this cluster might be better represented by other features.
  - Cluster 2: Mostly represented by "tertiary" level of education, which also translates well to the conclusions we made for the job distribution.

END