TÉCNICO LISBOA

Aprendizagem 2023/24

Homework III

I. Pen-and-paper [12v]

Given the following observations, $\left\{\begin{pmatrix}1\\0.6\\0.1\end{pmatrix}, \begin{pmatrix}0\\-0.4\\0.8\end{pmatrix}, \begin{pmatrix}0\\0.2\\0.5\end{pmatrix}, \begin{pmatrix}1\\0.4\\-0.1\end{pmatrix}\right\}$.

Consider a Bayesian clustering that assumes $\{y_1\} \perp\!\!\!\perp \{y_2, y_3\}$, two clusters following a Bernoulli distribution on y_1 (p_1 and p_2), a multivariate Gaussian on $\{y_2, y_3\}$ (N_1 and N_2), and the following initial mixture:

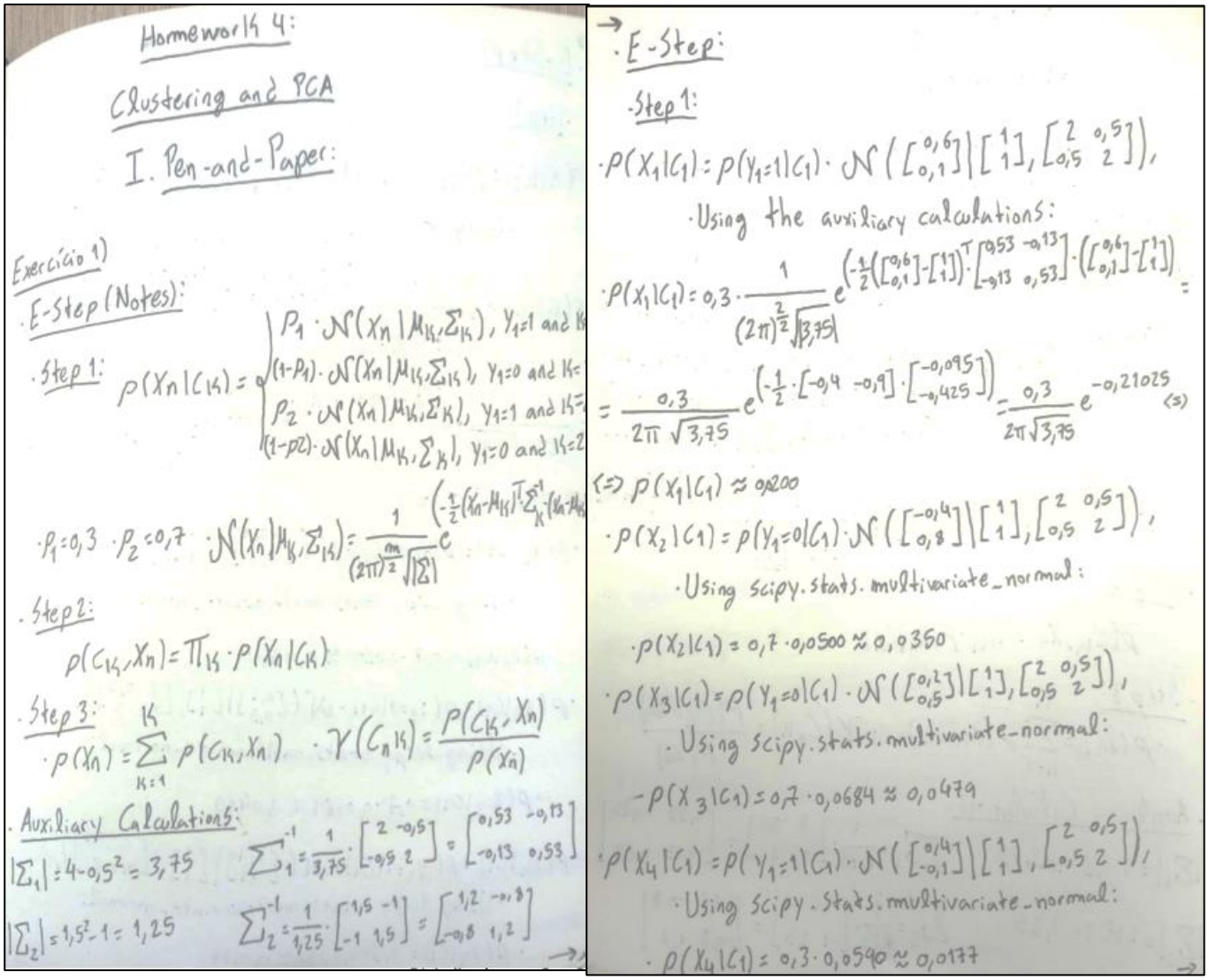
$$\pi_1 = 0.5, \pi_2 = 0.5$$

$$p_1 = P(y_1 = 1) = 0.3, \quad p_2 = P(y_1 = 1) = 0.7$$

$$N_1\left(\mathbf{u}_1 = \begin{pmatrix}1\\1\end{pmatrix}, \mathbf{\Sigma}_1 = \begin{pmatrix}2 & 0.5\\0.5 & 2\end{pmatrix}\right), \quad N_2\left(\mathbf{u}_2 = \begin{pmatrix}0\\0\end{pmatrix}, \mathbf{\Sigma}_2 = \begin{pmatrix}1.5 & 1\\1 & 1.5\end{pmatrix}\right).$$

Note: you can solve this exercise by neglecting y_1 and still scoring up to 70% of its grade.

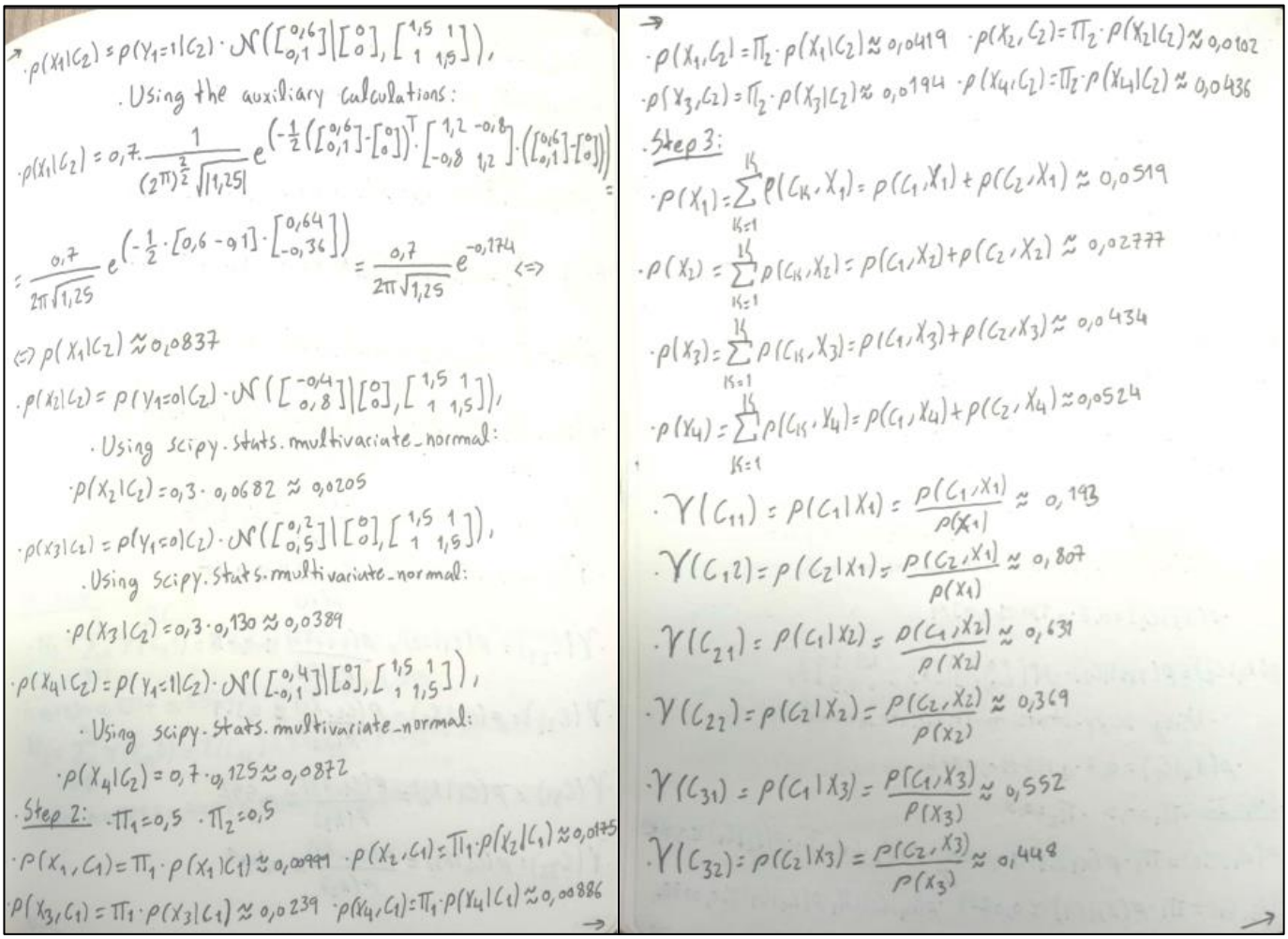
1) [6v] Perform one epoch of the EM clustering algorithm and determine the new parameters. Hint: we suggest you to use numpy and scipy, however disclose the intermediary results step by step.



TÉCNICO LISBOA

Aprendizagem 2023/24

Homework III



$\cdot p(X_1|C_2) = p(Y_1=1|C_2) \cdot \mathcal{N}\left(\begin{bmatrix}0,6\\0,1\end{bmatrix} \middle| \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1,5 & 1\\1 & 1,5\end{bmatrix}\right)$,

. Using the auxiliary calculations:

$\cdot p(X_1|C_2) = 0,7 \cdot \frac{1}{(2\pi)^{\frac{2}{2}}\sqrt{|1,25|}} e^{\left(-\frac{1}{2}\left(\begin{bmatrix}0,6\\0,1\end{bmatrix}-\begin{bmatrix}0\\0\end{bmatrix}\right)^T \cdot \begin{bmatrix}1,2 & -0,8\\-0,8 & 1,2\end{bmatrix} \cdot \left(\begin{bmatrix}0,6\\0,1\end{bmatrix}-\begin{bmatrix}0\\0\end{bmatrix}\right)\right)} =$

$= \frac{0,7}{2\pi\sqrt{1,25}} e^{\left(-\frac{1}{2} \cdot [0,6 \; 0,1] \cdot \begin{bmatrix}0,64\\0,36\end{bmatrix}\right)} = \frac{0,7}{2\pi\sqrt{1,25}} e^{-0,174} \Leftrightarrow$

$\Leftrightarrow p(X_1|C_2) \approx 0,0837$

$\cdot p(X_2|C_2) = p(Y_1=0|C_2) \cdot \mathcal{N}\left(\begin{bmatrix}-0,4\\0,8\end{bmatrix} \middle| \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1,5 & 1\\1 & 1,5\end{bmatrix}\right)$,

. Using scipy.stats.multivariate_normal:

$\cdot p(X_2|C_2) = 0,3 \cdot 0,0682 \approx 0,0205$

$\cdot p(X_3|C_2) = p(Y_1=0|C_2) \cdot \mathcal{N}\left(\begin{bmatrix}0,2\\0,5\end{bmatrix} \middle| \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1,5 & 1\\1 & 1,5\end{bmatrix}\right)$,

. Using scipy.stats.multivariate_normal:

$\cdot p(X_3|C_2) = 0,3 \cdot 0,130 \approx 0,0389$

$\cdot p(X_4|C_2) = p(Y_1=1|C_2) \cdot \mathcal{N}\left(\begin{bmatrix}0,4\\-0,1\end{bmatrix} \middle| \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1,5 & 1\\1 & 1,5\end{bmatrix}\right)$,

. Using scipy.stats.multivariate_normal:

$\cdot p(X_4|C_2) = 0,7 \cdot 0,125 \approx 0,0872$

Step 2: $\cdot \pi_1 = 0,5 \quad \cdot \pi_2 = 0,5$

$\cdot p(X_1, C_1) = \pi_1 \cdot p(X_1|C_1) \approx 0,0099 \quad \cdot p(X_2, C_1) = \pi_1 \cdot p(X_2|C_1) \approx 0,0175$

$\cdot p(X_3, C_1) = \pi_1 \cdot p(X_3|C_1) \approx 0,0239 \quad \cdot p(X_4, C_1) = \pi_1 \cdot p(X_4|C_1) \approx 0,0086$

$\cdot p(X_1, C_2) = \pi_2 \cdot p(X_1|C_2) \approx 0,0419 \quad \cdot p(X_2, C_2) = \pi_2 \cdot p(X_2|C_2) \approx 0,0102$

$\cdot p(X_3, C_2) = \pi_2 \cdot p(X_3|C_2) \approx 0,0194 \quad \cdot p(X_4, C_2) = \pi_2 \cdot p(X_4|C_2) \approx 0,0436$

Step 3:

$\cdot p(X_1) = \sum_{k=1}^{K} p(C_k, X_1) = p(C_1, X_1) + p(C_2, X_1) \approx 0,0519$

$\cdot p(X_2) = \sum_{k=1}^{K} p(C_k, X_2) = p(C_1, X_2) + p(C_2, X_2) \approx 0,02777$

$\cdot p(X_3) = \sum_{k=1}^{K} p(C_k, X_3) = p(C_1, X_3) + p(C_2, X_3) \approx 0,0434$

$\cdot p(X_4) = \sum_{k=1}^{K} p(C_k, X_4) = p(C_1, X_4) + p(C_2, X_4) \approx 0,0524$

$\cdot \gamma(C_{11}) = p(C_1|X_1) = \frac{p(C_1, X_1)}{p(X_1)} \approx 0,193$

$\cdot \gamma(C_{12}) = p(C_2|X_1) = \frac{p(C_2, X_1)}{p(X_1)} \approx 0,807$

$\cdot \gamma(C_{21}) = p(C_1|X_2) = \frac{p(C_1, X_2)}{p(X_2)} \approx 0,631$

$\cdot \gamma(C_{22}) = p(C_2|X_2) = \frac{p(C_2, X_2)}{p(X_2)} \approx 0,369$

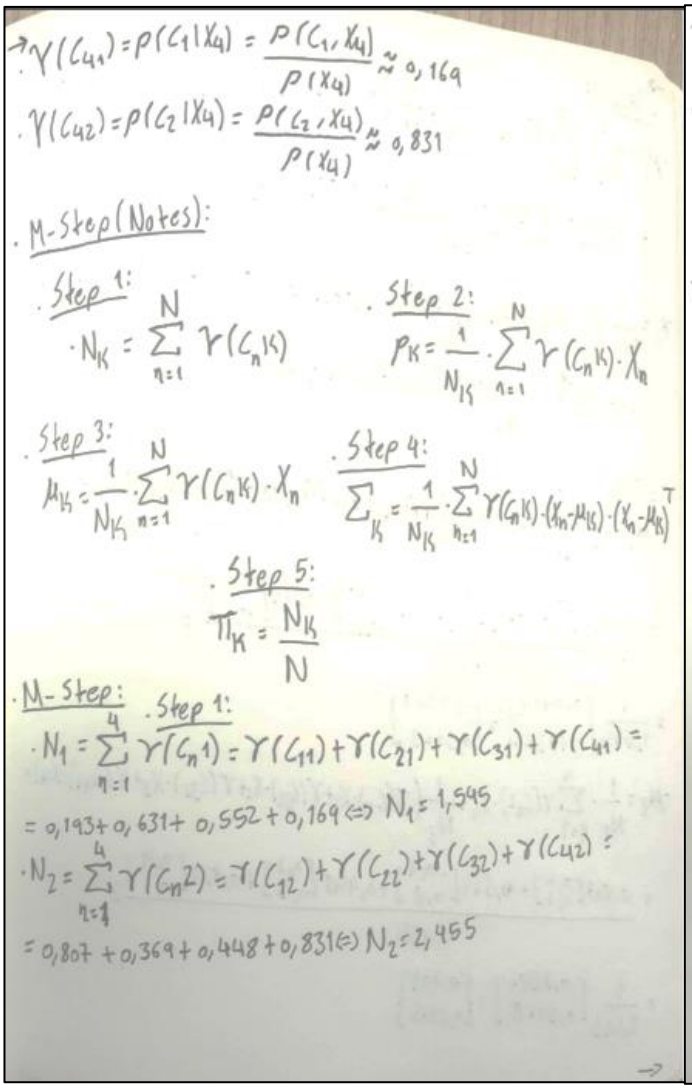
$\cdot \gamma(C_{31}) = p(C_1|X_3) = \frac{p(C_1, X_3)}{p(X_3)} \approx 0,552$

$\cdot \gamma(C_{32}) = p(C_2|X_3) = \frac{p(C_2, X_3)}{p(X_3)} \approx 0,448$

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework III



$\gamma(C_{41}) = p(C_1|X_4) = \frac{p(C_1, X_4)}{p(X_4)} \approx 0,169$

$\gamma(C_{42}) = p(C_2|X_4) = \frac{p(C_2, X_4)}{p(X_4)} \approx 0,831$

M-Step (Notes):

Step 1:

$N_K = \sum_{n=1}^{N} \gamma(C_n K)$

Step 2:

$p_K = \frac{1}{N_K} \cdot \sum_{n=1}^{N} \gamma(C_n K) \cdot X_n$

Step 3:

$\mu_K = \frac{1}{N_K} \cdot \sum_{n=1}^{N} \gamma(C_n K) \cdot X_n$

Step 4:

$\sum_K = \frac{1}{N_K} \cdot \sum_{n=1}^{N} \gamma(C_n K) \cdot (X_n - \mu_K) \cdot (X_n - \mu_K)^T$

Step 5:

$\pi_K = \frac{N_K}{N}$

M-Step:

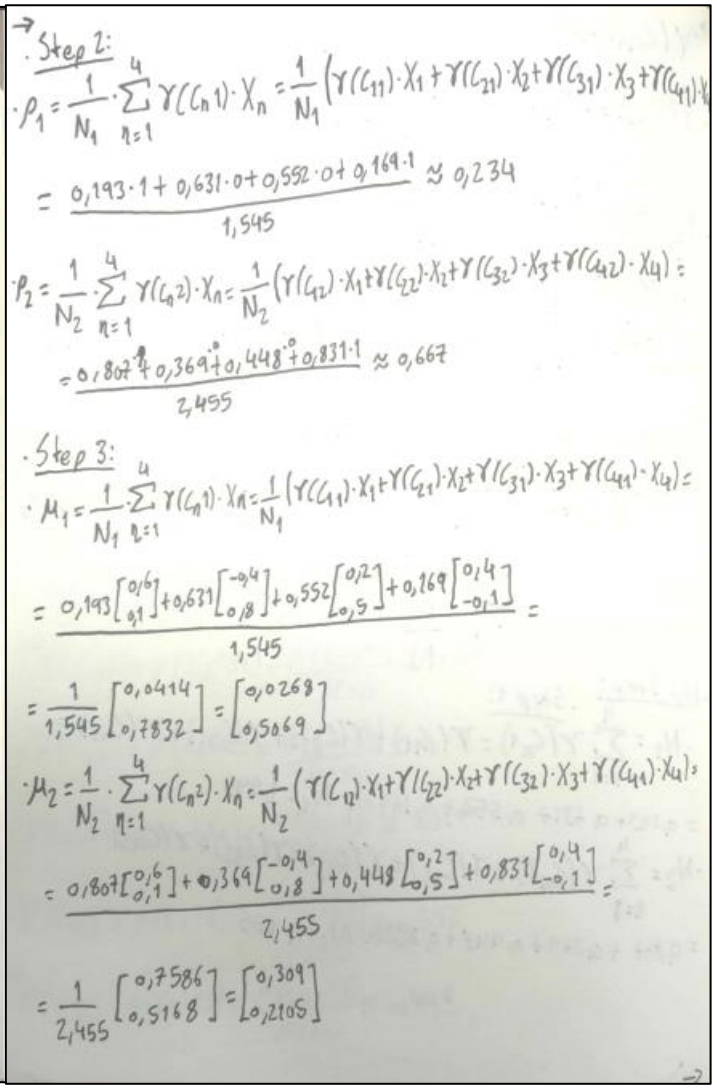
Step 1:

$N_1 = \sum_{n=1}^{4} \gamma(C_n 1) = \gamma(C_{11}) + \gamma(C_{21}) + \gamma(C_{31}) + \gamma(C_{41}) =$

$= 0,193 + 0,631 + 0,552 + 0,169 \Leftrightarrow N_1 = 1,545$

$N_2 = \sum_{n=1}^{4} \gamma(C_n 2) = \gamma(C_{12}) + \gamma(C_{22}) + \gamma(C_{32}) + \gamma(C_{42}) =$

$= 0,807 + 0,369 + 0,448 + 0,831 \Leftrightarrow N_2 = 2,455$



Step 2:

$p_1 = \frac{1}{N_1} \cdot \sum_{n=1}^{4} \gamma(C_n 1) \cdot X_n = \frac{1}{N_1} \left(\gamma(C_{11}) \cdot X_1 + \gamma(C_{21}) \cdot X_2 + \gamma(C_{31}) \cdot X_3 + \gamma(C_{41}) \cdot X_4\right) =$

$= \frac{0,193 \cdot 1 + 0,631 \cdot 0 + 0,552 \cdot 0 + 0,169 \cdot 1}{1,545} \approx 0,234$

$p_2 = \frac{1}{N_2} \cdot \sum_{n=1}^{4} \gamma(C_n 2) \cdot X_n = \frac{1}{N_2} \left(\gamma(C_{12}) \cdot X_1 + \gamma(C_{22}) \cdot X_2 + \gamma(C_{32}) \cdot X_3 + \gamma(C_{42}) \cdot X_4\right) =$

$= \frac{0,807 \cdot 1 + 0,369 \cdot 0 + 0,448 \cdot 0 + 0,831 \cdot 1}{2,455} \approx 0,667$

Step 3:

$\mu_1 = \frac{1}{N_1} \cdot \sum_{n=1}^{4} \gamma(C_n 1) \cdot X_n = \frac{1}{N_1} \left(\gamma(C_{11}) \cdot X_1 + \gamma(C_{21}) \cdot X_2 + \gamma(C_{31}) \cdot X_3 + \gamma(C_{41}) \cdot X_4\right) =$

$= \frac{0,193 \begin{bmatrix} 0,6 \\ 0,1 \end{bmatrix} + 0,631 \begin{bmatrix} -0,4 \\ 0,8 \end{bmatrix} + 0,552 \begin{bmatrix} 0,2 \\ 0,5 \end{bmatrix} + 0,169 \begin{bmatrix} 0,4 \\ -0,1 \end{bmatrix}}{1,545} =$

$= \frac{1}{1,545} \begin{bmatrix} 0,0414 \\ 0,7832 \end{bmatrix} = \begin{bmatrix} 0,0268 \\ 0,5069 \end{bmatrix}$

$\mu_2 = \frac{1}{N_2} \cdot \sum_{n=1}^{4} \gamma(C_n 2) \cdot X_n = \frac{1}{N_2} \left(\gamma(C_{12}) \cdot X_1 + \gamma(C_{22}) \cdot X_2 + \gamma(C_{32}) \cdot X_3 + \gamma(C_{42}) \cdot X_4\right) =$

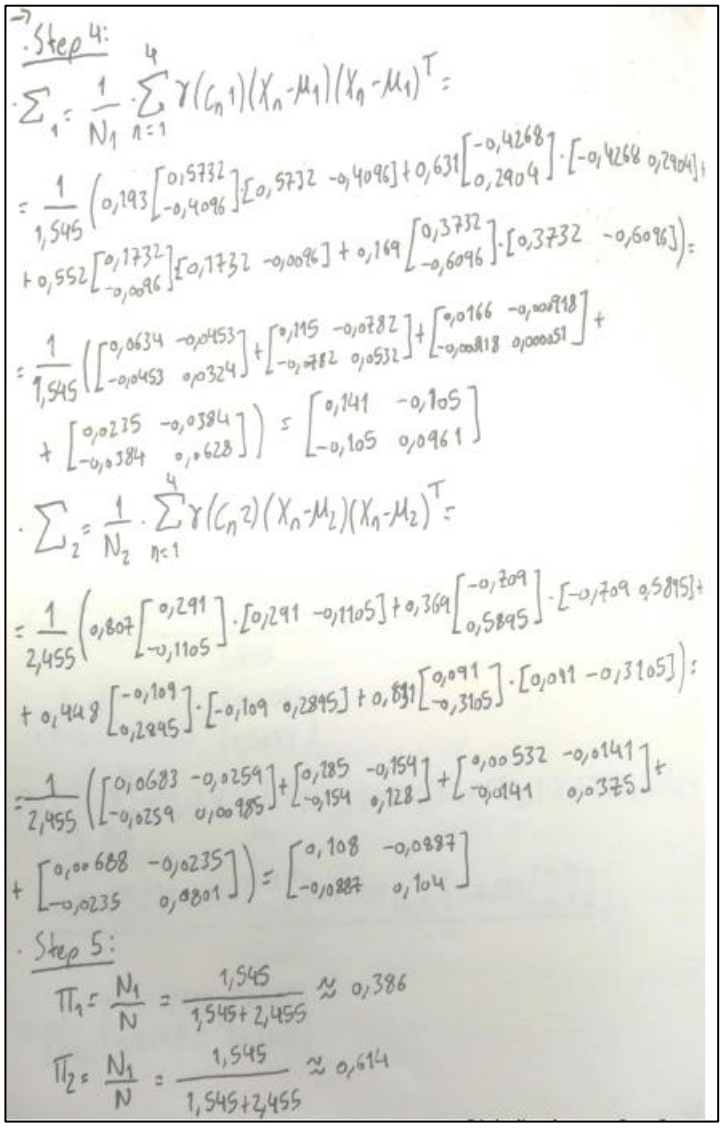
$= \frac{0,807 \begin{bmatrix} 0,6 \\ 0,1 \end{bmatrix} + 0,369 \begin{bmatrix} -0,4 \\ 0,8 \end{bmatrix} + 0,448 \begin{bmatrix} 0,2 \\ 0,5 \end{bmatrix} + 0,831 \begin{bmatrix} 0,4 \\ -0,1 \end{bmatrix}}{2,455} =$

$= \frac{1}{2,455} \begin{bmatrix} 0,7586 \\ 0,5168 \end{bmatrix} = \begin{bmatrix} 0,309 \\ 0,2105 \end{bmatrix}$

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework III



· Step 4:

$$\Sigma_1 = \frac{1}{N_1}\cdot\sum_{n=1}^{4}\gamma(C_n1)(X_n-\mu_1)(X_n-\mu_1)^T =$$

$$= \frac{1}{1,545}\left(0,193\begin{bmatrix}0,5732\\-0,4096\end{bmatrix}\cdot[0,5732\ \ -0,4096]+0,631\begin{bmatrix}-0,4268\\0,2904\end{bmatrix}\cdot[-0,4268\ \ 0,2904]+\right.$$

$$\left.+0,552\begin{bmatrix}0,1732\\-0,0096\end{bmatrix}[0,1732\ \ -0,0096]+0,169\begin{bmatrix}0,3732\\-0,6096\end{bmatrix}\cdot[0,3732\ \ -0,6096]\right) =$$

$$= \frac{1}{1,545}\left(\begin{bmatrix}0,0634 & -0,0453\\-0,0453 & 0,0324\end{bmatrix}+\begin{bmatrix}0,195 & -0,0782\\-0,0782 & 0,0532\end{bmatrix}+\begin{bmatrix}0,0166 & -0,00918\\-0,00918 & 0,000851\end{bmatrix}+\right.$$

$$\left.+\begin{bmatrix}0,0235 & -0,0384\\-0,0384 & 0,628\end{bmatrix}\right) = \begin{bmatrix}0,141 & -0,105\\-0,105 & 0,0961\end{bmatrix}$$

$$\cdot\ \Sigma_2 = \frac{1}{N_2}\cdot\sum_{n=1}^{4}\gamma(C_n2)(X_n-\mu_2)(X_n-\mu_2)^T =$$

$$= \frac{1}{2,455}\left(0,807\begin{bmatrix}0,291\\-0,1105\end{bmatrix}\cdot[0,291\ \ -0,1105]+0,369\begin{bmatrix}-0,709\\0,5895\end{bmatrix}\cdot[-0,709\ \ 0,5895]+\right.$$

$$\left.+0,448\begin{bmatrix}-0,109\\0,2895\end{bmatrix}\cdot[-0,109\ \ 0,2895]+0,891\begin{bmatrix}0,091\\-0,3105\end{bmatrix}\cdot[0,091\ \ -0,3105]\right) =$$

$$= \frac{1}{2,455}\left(\begin{bmatrix}0,0683 & -0,0259\\-0,0259 & 0,00985\end{bmatrix}+\begin{bmatrix}0,285 & -0,154\\-0,154 & 0,128\end{bmatrix}+\begin{bmatrix}0,00532 & -0,0141\\-0,0141 & 0,0375\end{bmatrix}+\right.$$

$$\left.+\begin{bmatrix}0,00688 & -0,0235\\-0,0235 & 0,0801\end{bmatrix}\right) = \begin{bmatrix}0,108 & -0,0887\\-0,0887 & 0,104\end{bmatrix}$$

· Step 5:

$$\pi_1 = \frac{N_1}{N} = \frac{1,545}{1,545+2,455} \approx 0,386$$

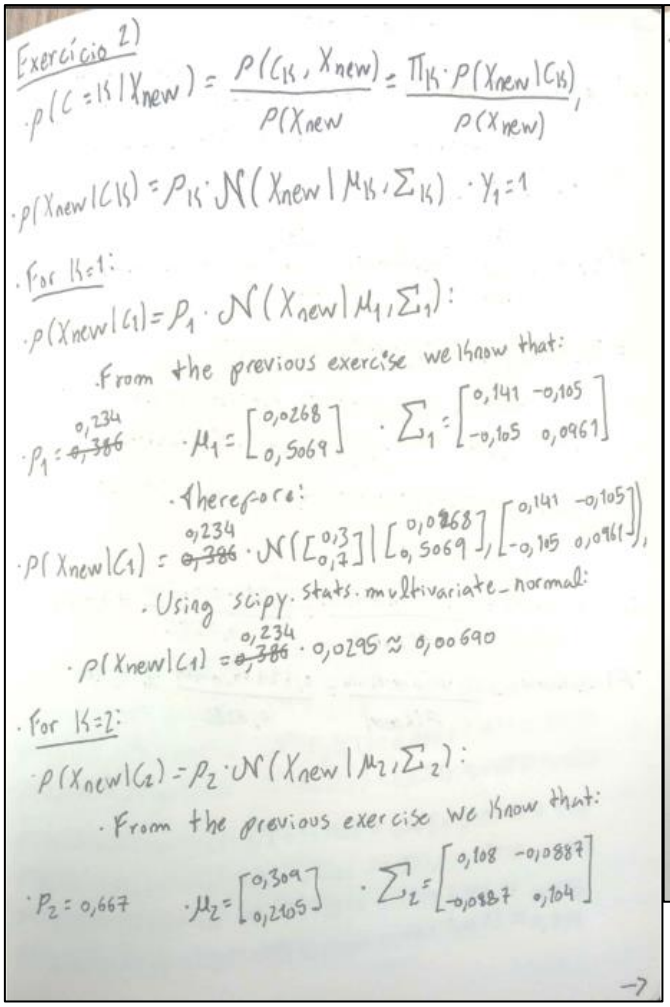
$$\pi_2 = \frac{N_1}{N} = \frac{1,545}{1,545+2,455} \approx 0,614$$

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework III

2) [2v] Given the new observation, $\mathbf{x}_{new} = \begin{pmatrix} 1 \\ 0.3 \\ 0.7 \end{pmatrix}$, determine the cluster memberships (posteriors).



Exercício 2)

$p(C=k|X_{new}) = \frac{p(C_k, X_{new})}{p(X_{new})} = \frac{\pi_k \cdot p(X_{new}|C_k)}{p(X_{new})}$,

$p(X_{new}|C_k) = p_k \cdot \mathcal{N}(X_{new}|\mu_k, \Sigma_k) \quad \cdot y_1 = 1$

. For $k=1$:

$p(X_{new}|C_1) = p_1 \cdot \mathcal{N}(X_{new}|\mu_1, \Sigma_1)$:

. From the previous exercise we know that:

$p_1 = \frac{0,234}{0,386} \quad \mu_1 = \begin{bmatrix} 0,0268 \\ 0,5069 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 0,141 & -0,105 \\ -0,105 & 0,0961 \end{bmatrix}$

. Therefore:

$p(X_{new}|C_1) = \frac{0,234}{0,386} \cdot \mathcal{N}\left(\begin{bmatrix} 0,3 \\ 0,7 \end{bmatrix} \middle| \begin{bmatrix} 0,0268 \\ 0,5069 \end{bmatrix}, \begin{bmatrix} 0,141 & -0,105 \\ -0,105 & 0,0961 \end{bmatrix}\right)$,

. Using scipy.stats.multivariate_normal:

$p(X_{new}|C_1) = \frac{0,234}{0,386} \cdot 0,0295 \approx 0,00690$

. For $k=2$:

$p(X_{new}|C_2) = p_2 \cdot \mathcal{N}(X_{new}|\mu_2, \Sigma_2)$:

. From the previous exercise we know that:

$p_2 = 0,667 \quad \mu_2 = \begin{bmatrix} 0,309 \\ 0,2105 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 0,108 & -0,0887 \\ -0,0887 & 0,104 \end{bmatrix}$

→

. Therefore:

$p(X_{new}|C_2) = 0,667 \cdot \mathcal{N}\left(\begin{bmatrix} 0,3 \\ 0,7 \end{bmatrix} \middle| \begin{bmatrix} 0,309 \\ 0,2105 \end{bmatrix}, \begin{bmatrix} 0,108 & -0,0887 \\ -0,0887 & 0,104 \end{bmatrix}\right)$,

. Using scipy.stats.multivariate_normal:

$p(X_{new}|C_2) = 0,667 \cdot 0,0629 \approx 0,0417$

$p(X_{new}) = \sum_{k=1}^{K} p(C_k, X_{new}) = \sum_{k=1}^{k} \pi_k \cdot p(X_{new}, C_k) =$

$= \pi_1 \cdot p(X_{new}|C_1) + \pi_2 \cdot p(X_{new}|C_2)$:

. From the previous exercise we know that:

$\pi_1 = 0,386 \quad \pi_2 = 0,614$. Therefore:

$p(X_{new}) = 0,386 \cdot 0,00690 + 0,614 \cdot 0,0417 \approx 0,0282$

. Therefore:

$p(C_1|X_{new}) = \frac{\pi_1 \cdot p(X_{new}|C_1)}{p(X_{new})} = \frac{0,386 \cdot 0,00690}{0,0282} \approx 0,0943$

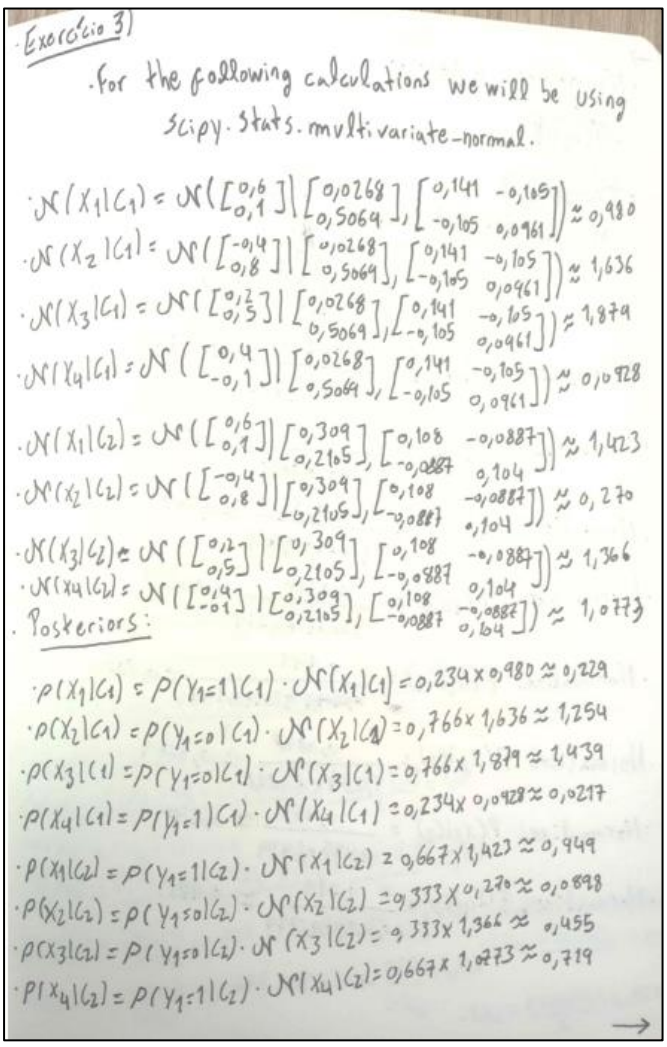
$p(C_2|X_{new}) = \frac{\pi_2 \cdot p(X_{new}|C_2)}{p(X_{new})} = \frac{0,614 \cdot 0,0417}{0,0282} \approx 0,906$

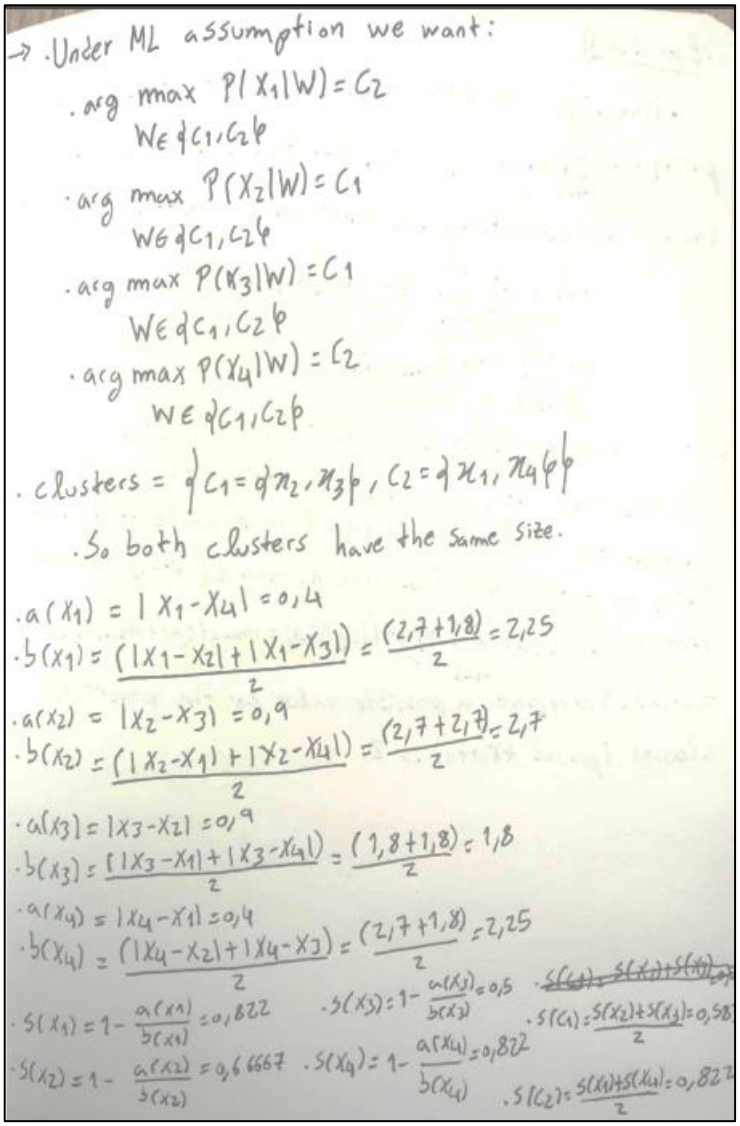
TÉCNICO LISBOA

Aprendizagem 2023/24

Homework III

3) [2.5v] Performing a hard assignment of observations to clusters under a ML assumption, identify the silhouette of the larger cluster under a Manhattan distance.



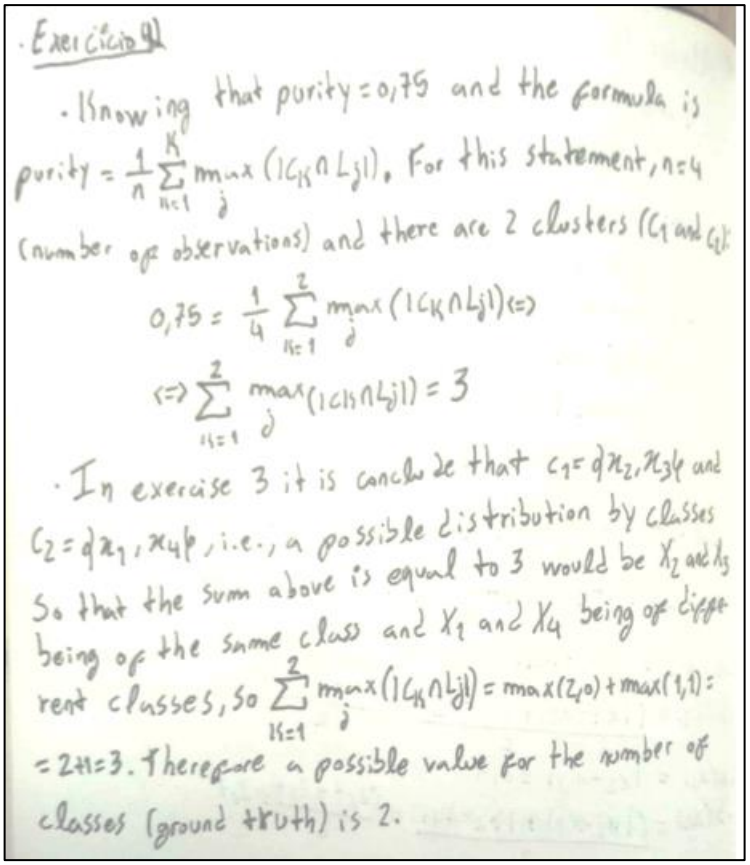


TÉCNICO LISBOA

Aprendizagem 2023/24

Homework III

4) [0.5v] Knowing the purity of the clustering solution is 0.75, identify the number of possible classes (ground truth).



· Exercício 4)

· Knowing that purity = 0,75 and the formula is purity $= \frac{1}{n}\sum_{k=1}^{K}\max_{j}(|C_k \cap L_j|)$. For this statement, n=4 (number of observations) and there are 2 clusters (C_1 and C_2):

$$0{,}75 = \frac{1}{4}\sum_{k=1}^{2}\max_{j}(|C_k \cap L_j|) \Leftrightarrow$$

$$\Leftrightarrow \sum_{k=1}^{2}\max_{j}(|C_k \cap L_j|) = 3$$

· In exercise 3 it is concluded that $c_1 = \{x_2, x_3\}$ and $c_2 = \{x_1, x_4\}$, i.e., a possible distribution by classes so that the sum above is equal to 3 would be x_2 and x_3 being of the same class and x_1 and x_4 being of different classes, so $\sum_{k=1}^{2}\max_{j}(|C_k \cap L_j|) = \max(2,0) + \max(1,1) = 2+1=3$. Therefore a possible value for the number of classes (ground truth) is 2.

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework III

Programming and critical analysis [8v]

Recall the `column_diagnosis.arff` dataset from previous homework's. For the following exercises, normalize the data using sklearn's MinMaxScaler.

1) [4v] Using `sklearn`, apply k-means clustering fully unsupervised Ly on the normalized data with $k \in \{2,3,4,5\}$ (`random=0` and remaining parameters as default). Assess the silhouette and purity of the produced solutions.

```
import pandas as pd
import numpy as np
from scipy.io.arff import loadarff
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
from sklearn.metrics.cluster import contingency_matrix
from sklearn.metrics import silhouette_score

# Read ARFF file
data = loadarff('column_diagnosis.arff')
df = pd.DataFrame(data[0])
df['class'] = df['class'].str.decode('utf-8')

X = df.drop('class', axis=1)
y = df['class']
X_scaled = MinMaxScaler().fit_transform(X)

k_values = [2, 3, 4, 5]

for k in k_values:
    # Perform K-means clustering
    kmeans = KMeans(n_clusters=k, random_state=0).fit(X_scaled)
    if (k==3):
        kmeans3 = kmeans

    # Calculate silhouette scores and purity scores
    s_score = silhouette_score(X_scaled, kmeans.labels_)
    confusion_matrix = contingency_matrix(y, kmeans.labels_)
    p_score = np.sum(np.amax(confusion_matrix, axis=0)) / np.sum(confusion_matrix)

    print('For K =', k)
    print('Silhouete score:', s_score)
    print('Purity score:', p_score, '\n')
```

Output:

For K = 2
Silhouete score: 0.36044124340441114
Purity score: 0.632258064516129

For K = 3
Silhouete score: 0.29579055730002257
Purity score: 0.667741935483871

For K = 4
Silhouete score: 0.27442402122340176
Purity score: 0.6612903225806451

For K = 5
Silhouete score: 0.23823928397844843
Purity score: 0.6774193548387096

2) [2v] Consider the application of PCA after the data normalization:

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework III

i. Identify the variability explained by the top two principal components.

```
from sklearn.decomposition import PCA

pca = PCA(n_components=2)
principal_components = pca.fit(X_scaled)

# Compute the explained variance for the top two components
explained_variance = pca.explained_variance_ratio_
print('Explained variance of the top two components:', explained_variance)
```

Output:

Explained variance of the top two components: [0.56181445 0.20955953]

ii. For each one of these two components, sort the input variables by relevance by inspecting the absolute weights of the linear projection.

```
# Get the absolute weights
component_weights = abs(pca.components_)

# Sort the input variables by relevance
relevance_df = pd.DataFrame(component_weights, columns=X.columns, index=['Component 1', 'Component 2']).T
relevance_df = relevance_df.sort_values(by=['Component 1', 'Component 2'], ascending=False)

print('Sorted relevance of input variables in the top two components:')
print(relevance_df)
```

Output:

```
Sorted relevance of input variables in the top two components:
                          Component 1  Component 2
pelvic_incidence             0.591621     0.100037
lumbar_lordosis_angle        0.515085     0.080047
pelvic_tilt                  0.467039     0.670373
sacral_slope                 0.325689     0.443303
degree_spondylolisthesis     0.216930     0.004583
pelvic_radius                0.115824     0.581074
```

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework III

3) [2v] Visualize side-by-side the data using: i) the ground diagnoses, and ii) the *previously* learned $k = 3$ clustering solution. To this end, projected the normalized data onto a 2-dimensional data space using PCA and then color observations using the reference and cluster annotations.

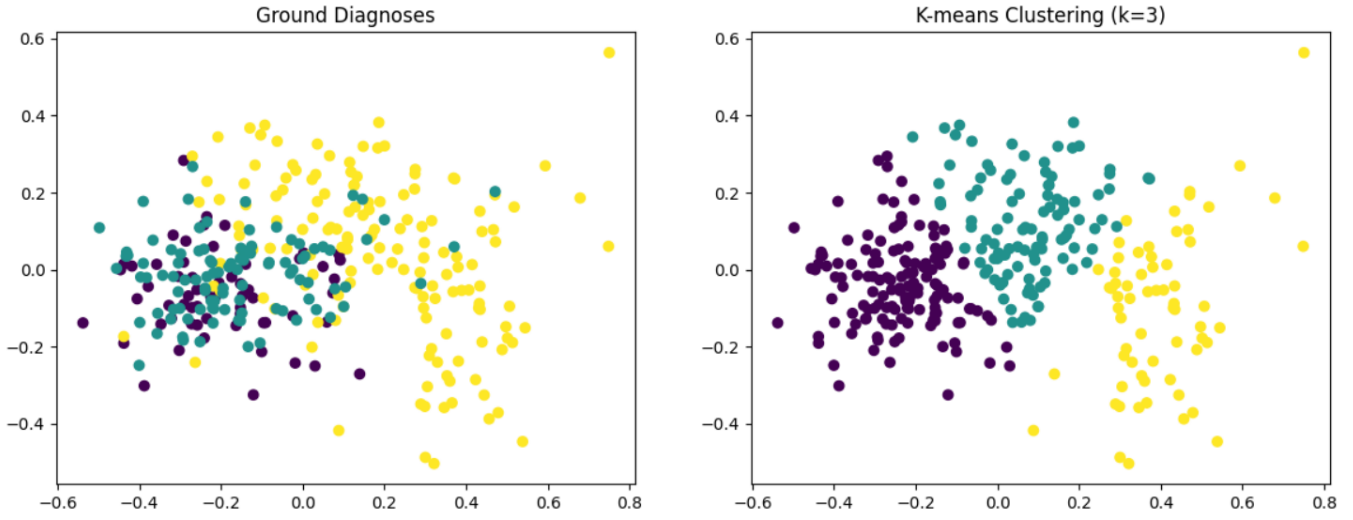
```
import matplotlib.pyplot as plt

# Reduce the dimensionality of data
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

# Real labels as inetegers to compare with kmeans labels
codes = {'Hernia':0, 'Normal':1, 'Spondylolisthesis':2}
y_pred = y.map(codes).tolist()

plt.figure(figsize=(14, 5))
plt.subplot(121)
plt.scatter(X_pca[:,0], X_pca[:,1], c=y_pred)
plt.title('Ground Diagnoses')

plt.subplot(122)
plt.scatter(X_pca[:,0], X_pca[:,1], c=kmeans3.labels_)
plt.title('K-means Clustering (k=3)')
plt.show()
```



4) [1v] Considering the results from questions (1) and (3), identify two ways on how clustering can be used to characterize the population of ill and healthy individuals.

Answer: Clustering can be used to identify groups of patients with similar medical characteristics to discover potential causes of the disease or define groups at higher risk of developing the disease.

END