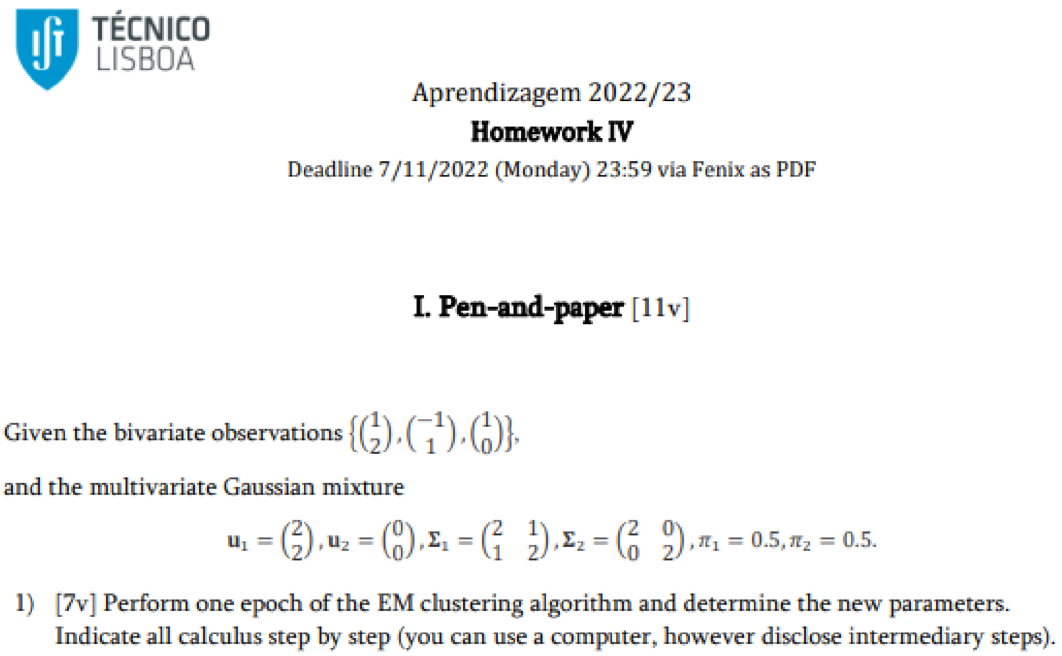


TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV

Deadline 7/11/2022 (Monday) 23:59 via Fenix as PDF

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).

E-Step:

For Cluster k, observation x_n:

$$p(x_n|c_k = 1) = N(x_n|\mu_k, \Sigma_k) = \frac{exp(-\frac{1}{2}.(x_n-\mu_k)^T \Sigma_k^{-1}.(x_n-\mu_k))}{(2\pi)^{\frac{K}{2}}\sqrt{|\Sigma_k|}}$$

$$\Sigma_1^{-1} = |\Sigma_1|^{-1}\begin{bmatrix} 2 & -1 \\ -1 & 2\end{bmatrix} = \frac{1}{2x2-1x1}\begin{bmatrix} 2 & -1 \\ -1 & 2\end{bmatrix} = \frac{1}{3}\begin{bmatrix} 2 & -1 \\ -1 & 2\end{bmatrix} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3}\end{bmatrix}$$

$$\Sigma_2^{-1} = |\Sigma_2|^{-1}\begin{bmatrix} 2 & 0 \\ 0 & 2\end{bmatrix} = \frac{1}{2x2}\begin{bmatrix} 2 & 0 \\ 0 & 2\end{bmatrix} = \frac{1}{4}\begin{bmatrix} 2 & 0 \\ 0 & 2\end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2}\end{bmatrix}$$

Cluster 1:

$$\text{p}(x_1|c_1 = 1) = \frac{exp(-\frac{1}{2}.(x_1-\mu_1)^T \Sigma_1^{-1}.(x_1-\mu_1))}{(2\pi)^{\frac{2}{2}}\sqrt{3}} = \frac{exp(-\frac{1}{2}.\frac{2}{3})}{(2\pi)^{\frac{2}{2}}\sqrt{3}} = 0.0658$$

$$\text{p}(x_2|c_1 = 1) = \frac{exp(-\frac{1}{2}.(x_2-\mu_1)^T \Sigma_1^{-1}.(x_2-\mu_1))}{(2\pi)^{\frac{2}{2}}\sqrt{3}} = \frac{exp(-\frac{1}{2}.\frac{14}{3})}{(2\pi)^{\frac{2}{2}}\sqrt{3}} = 0.00891$$

$$\text{p}(x_3|c_1 = 1) = \frac{exp(-\frac{1}{2}.(x_3-\mu_1)^T \Sigma_1^{-1}.(x_3-\mu_1))}{(2\pi)^{\frac{2}{2}}\sqrt{3}} = \frac{exp(-\frac{1}{2}.2)}{(2\pi)^{\frac{2}{2}}\sqrt{3}} = 0.0338$$

Cluster 2:

$$p(x_1|c_2 = 1)= \frac{exp(-\frac{1}{2}.(x_1-\mu_2)^T\Sigma_2^{-1}.(x_1-\mu_2))}{(2\pi)^{\frac{2}{2}}\sqrt{4}} = \frac{exp(-\frac{1}{2}.\frac{2}{3})}{(2\pi)^{\frac{2}{2}}\sqrt{4}} = 0.0570$$

$$p(x_2|c_2 = 1)= \frac{exp(-\frac{1}{2}.(x_2-\mu_2)^T\Sigma_2^{-1}.(x_2-\mu_2))}{(2\pi)^{\frac{2}{2}}\sqrt{4}} = \frac{exp(-\frac{1}{2}.\frac{14}{3})}{(2\pi)^{\frac{2}{2}}\sqrt{4}} = 0.00772$$

$$p(x_3|c_2 = 1)= \frac{exp(-\frac{1}{2}.(x_3-\mu_2)^T\Sigma_2^{-1}.(x_3-\mu_2))}{(2\pi)^{\frac{2}{2}}\sqrt{4}} = \frac{exp(-\frac{1}{2}.2)}{(2\pi)^{\frac{2}{2}}\sqrt{4}} = 0.0293$$

$$p(c_k = 1|x_n) = \pi_k N(x_n|\mu_k, \Sigma_k)=\pi_k p(x_n|c_k = 1)$$

$$p(c_1 = 1|x_1) = 0.5 \times 0.0658 = 0.0329$$

$$p(c_1 = 1|x_2) = 0.5 \times 0.00891 = 0.004455$$

$$p(c_1 = 1|x_3) = 0.5 \times 0.0338 = 0.0169$$

$$p(c_2 = 1|x_1) = 0.5 \times 0.0570 = 0.0285$$

$$p(c_2 = 1|x_2) = 0.5 \times 0.00772 = 0.00386$$

$$p(c_2 = 1|x_3) = 0.5 \times 0.0293 = 0.01463$$

$$p(x_n) = \sum_{k=1}^{\#K} p(c_k = 1|x_n)$$

$$p(x_1) = 0.0329 + 0.0285 = 0.0614$$

$$p(x_2) = 0.004455 + 0.00386 = 0.008315$$

$$p(x_3) = 0.0169 + 0.01463 = 0.03153$$

Normalization:

$$\gamma_{nk} = \frac{p(c_k=1|x_n)}{p(x_n)}$$

$\gamma_{11} = \frac{p(c_1=1|x_1)}{p(x_1)} = \frac{0.0329}{0.0614}$ =0.5358

$\gamma_{21} = \frac{p(c_1=1|x_2)}{p(x_2)} = \frac{0.004455}{0.008315}$ =0.5358

$\gamma_{31} = \frac{p(c_1=1|x_3)}{p(x_3)} = \frac{0.0169}{0.03153}$ =0.5360

$\gamma_{12} = \frac{p(c_2=1|x_1)}{p(x_1)} = \frac{0.0285}{0.0614}$ =0.4642

$\gamma_{22} = \frac{p(c_2=1|x_2)}{p(x_2)} = \frac{0.00386}{0.008315}$ =0.4642

$\gamma_{32} = \frac{p(c_2=1|x_3)}{p(x_3)} = \frac{0.01463}{0.03153}$ =0.4640

M-Step:

$$N_k = \sum_{n=1}^{\#N} \gamma_{nk}$$

$$N_1 = \gamma_{11} + \gamma_{21} + \gamma_{31} = 1.6076$$

$$N_2 = \gamma_{12} + \gamma_{22} + \gamma_{32} = 1.3924$$

New mean vectors::

$$\mu_k = \frac{1}{N_k} \times \sum_{n=1}^{\#N} \gamma_{nk}.x_n$$

$$\mu_1 = \frac{1}{1.6076}.(0.5358.\begin{pmatrix} 1 \\ 2 \end{pmatrix} + 0.5358.\begin{pmatrix} -1 \\ 1 \end{pmatrix} + 0.5360.\begin{pmatrix} 1 \\ 0 \end{pmatrix}) = \begin{pmatrix} 0.333 \\ 1 \end{pmatrix}$$

$$\mu_2 = \frac{1}{1.3924}.(0.4642.\begin{pmatrix} 1 \\ 2 \end{pmatrix} + 0.4642.\begin{pmatrix} -1 \\ 1 \end{pmatrix} + 0.4640.\begin{pmatrix} 1 \\ 0 \end{pmatrix}) = \begin{pmatrix} 0.333 \\ 1 \end{pmatrix}$$

New covariance matrices:

$$\Sigma_k = \frac{1}{N_k} \times \sum_{n=1}^{\#N} \gamma_{nk}.(x_n - \mu_k)\ .(x_n - \mu_k)^T$$

$$\Sigma_1 = \frac{1}{1.6076}.(0.5358.(\begin{smallmatrix} 1-2 \\ 2-2 \end{smallmatrix}).(1 - 2 \quad 2 - 2) +$$
$$0.5358.(\begin{smallmatrix} -1-2 \\ 1-2 \end{smallmatrix}).(-1 - 2 \quad 1 - 2) +$$
$$0.5360.(\begin{smallmatrix} 1-2 \\ 0-2 \end{smallmatrix}).(1 - 2 \quad 0 - 2)) = (\begin{smallmatrix} 3.667 & 1.667 \\ 1.667 & 1.667 \end{smallmatrix})$$

$$\Sigma_2 = \frac{1}{1.3924}.(0.4642.(\begin{smallmatrix} 1-0 \\ 2-0 \end{smallmatrix}).(1 - 0 \quad 2 - 0) +$$
$$0.4642.(\begin{smallmatrix} -1-0 \\ 1-0 \end{smallmatrix}).(-1 - 0 \quad 1 - 0) +$$
$$0.4640.(\begin{smallmatrix} 1-0 \\ 0-0 \end{smallmatrix}).(1 - 0 \quad 0 - 0)) = (\begin{smallmatrix} 1 & 0.334 \\ 0.334 & 1.667 \end{smallmatrix})$$

New mixing parameters:

$$\pi_k = p(c_k = 1) = \frac{N_k}{\#N}$$

$$\pi_1 = p(c_1 = 1) = \frac{1.6076}{3} = 0.5359$$

$$\pi_2 = p(c_2 = 1) = \frac{1.3924}{3} = 0.4641$$

2)

a)

$$\Sigma_1^{-1} = |\Sigma_1|^{-1}\begin{pmatrix} 1.667 & -1.667 \\ -1.667 & 3.667 \end{pmatrix} = \frac{1}{3.334}\begin{pmatrix} 1.667 & -1.667 \\ -1.667 & 3.667 \end{pmatrix} = \begin{pmatrix} 0.5 & -0.5 \\ -0.5 & 1.1 \end{pmatrix}$$

$$\Sigma_2^{-1} = |\Sigma_2|^{-1}\begin{pmatrix} 1.667 & -0.334 \\ -0.334 & 1 \end{pmatrix} = \frac{1}{1.555}\begin{pmatrix} 1.667 & -0.334 \\ -0.334 & 1 \end{pmatrix} = \begin{pmatrix} 1.072 & -0.2148 \\ -0.2148 & 0.6431 \end{pmatrix}$$

$$p(x_n|c_k = 1) = N(x_n|\mu_k, \Sigma_k) = \frac{exp(-\frac{1}{2}.(x_n-\mu_k)^T\Sigma_k^{-1}.(x_n-\mu_k))}{(2\pi)^{\frac{K}{2}}\sqrt{|\Sigma_k|}}$$

Cluster 1:

$$\text{p}(x_1|c_1 = 1) = \frac{exp(-\frac{1}{2}.(x_1-\mu_1)^T\Sigma_1^{-1}.(x_1-\mu_1))}{(2\pi)^{\frac{2}{2}}\sqrt{3.334}} = \frac{exp(-\frac{1}{2}.\frac{2}{3})}{(2\pi)^{\frac{2}{2}}\sqrt{3.334}} = 0.0628$$

$$\text{p}(x_2|c_1 = 1) = \frac{exp(-\frac{1}{2}.(x_2-\mu_1)^T\Sigma_1^{-1}.(x_2-\mu_1))}{(2\pi)^{\frac{2}{2}}\sqrt{3.334}} = \frac{exp(-\frac{1}{2}.\frac{14}{3})}{(2\pi)^{\frac{2}{2}}\sqrt{3.334}} = 0.056$$

$$\text{p}(x_3|c_1 = 1) = \frac{exp(-\frac{1}{2}.(x_3-\mu_1)^T\Sigma_1^{-1}.(x_3-\mu_1))}{(2\pi)^{\frac{2}{2}}\sqrt{3.334}} = \frac{exp(-\frac{1}{2}.2)}{(2\pi)^{\frac{2}{2}}\sqrt{3.334}} = 0.0321$$

Cluster 2:

$$\text{p}(x_1|c_2 = 1) = \frac{exp(-\frac{1}{2}.(x_1-\mu_2)^T\Sigma_2^{-1}.(x_1-\mu_2))}{(2\pi)^{\frac{2}{2}}\sqrt{1.555}} = \frac{exp(-\frac{1}{2}.\frac{2}{3})}{(2\pi)^{\frac{2}{2}}\sqrt{1.555}} = 0.084$$

$$\text{p}(x_2|c_2 = 1) = \frac{exp(-\frac{1}{2}.(x_2-\mu_2)^T\Sigma_2^{-1}.(x_2-\mu_2))}{(2\pi)^{\frac{2}{2}}\sqrt{1.555}} = \frac{exp(-\frac{1}{2}.\frac{14}{3})}{(2\pi)^{\frac{2}{2}}\sqrt{1.555}} = 0.0494$$

$$\text{p}(x_3|c_2 = 1) = \frac{exp(-\frac{1}{2}.(x_3-\mu_2)^T\Sigma_2^{-1}.(x_3-\mu_2))}{(2\pi)^{\frac{2}{2}}\sqrt{1.555}} = \frac{exp(-\frac{1}{2}.2)}{(2\pi)^{\frac{2}{2}}\sqrt{1.555}} = 0.0630$$

$$\text{p}(c_k = 1|x_n) = \pi_k N(x_n|\mu_k, \Sigma_k) = \pi_k p(x_n|c_k = 1)$$

$$\text{p}(c_1 = 1|x_1) = 0.5359 \times 0.0628 = 0.0336$$

$$\text{p}(c_1 = 1|x_2) = 0.5359 \times 0.056 = 0.03$$

p($c_1 = 1|x_3$) $= 0.5359 \times 0.0321 = 0.0172$

p($c_2 = 1|x_1$) $= 0.4641 \times 0.084 = 0.0390$

p($c_2 = 1|x_2$) $= 0.4641 \times 0.0494 = 0.0229$

p($c_2 = 1|x_3$) $= 0.4641 \times 0.0630 = 0.0292$

$p(c_1 = 1)$ =0.0336+0.03+0.0172=0.0808

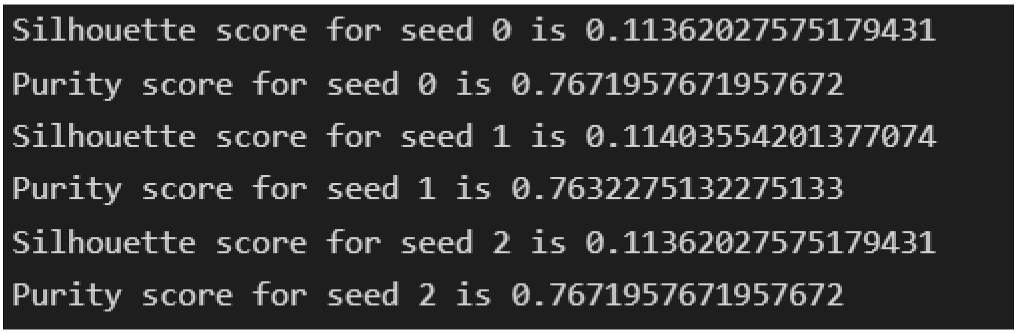
$p(c_2 = 1)$ =0.0390+0.0229+0.0292=0.0911

cluster 2 $p(c_2 = 1)$>$p(c_1 = 1)$.

II) Programming and critical analysis

1)

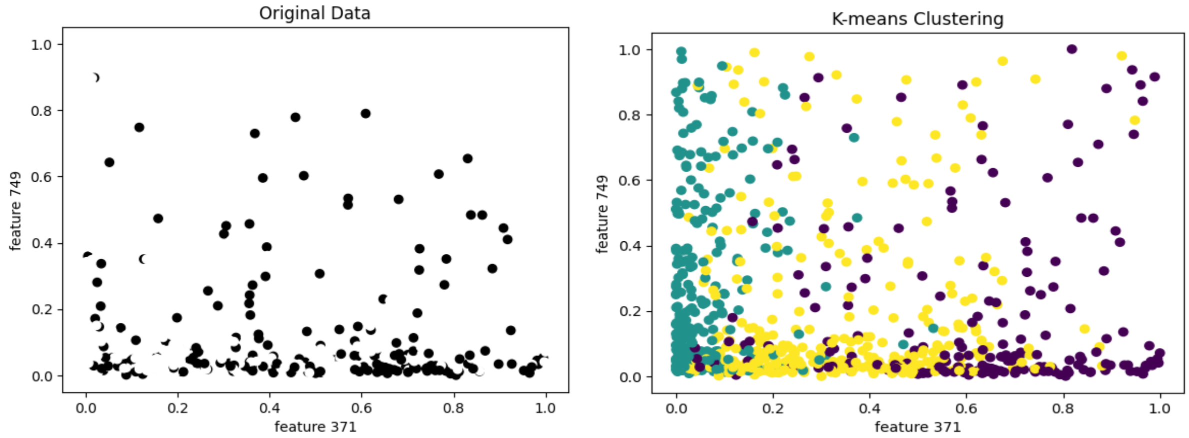
Using k = 3:



```
Silhouette score for seed 0 is 0.11362027575179431
Purity score for seed 0 is 0.7671957671957672
Silhouette score for seed 1 is 0.11403554201377074
Purity score for seed 1 is 0.7632275132275133
Silhouette score for seed 2 is 0.11362027575179431
Purity score for seed 2 is 0.7671957671957672
```

2)
The non-deterministic behavior is caused by the random initialization of the centroids. The centroids are initialized randomly, so the clustering results are different each time. Those are the points which are gonna tell us to where the cluster is gonna converge to, since they are updated by the median of the distances from the points to the center.

3)



4)
Number of principal components necessary to explain 80% of the variability in the data is 31

```
Number of principal components necessary to explain 80% of the variability in the data is 31
```

APPENDIX

```
# Import Wall
import pandas as pd
import numpy as np
```

```
from sklearn import metrics, datasets, tree
import matplotlib.pyplot as plt
from scipy.io.arff import loadarff
import seaborn as sns
from scipy import stats

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

#1)

#get features and targets
X = df.drop('class', axis=1)
Y = df['class']

# normalize the X using MinMaxScaler
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
X = scaler.fit_transform(X)


seeds = [0,1,2]
for seed in seeds:
    #apply k-means clustering fully unsupervised without targets with k=3
    from sklearn.cluster import KMeans
    kmeans = KMeans(n_clusters=3, random_state=seed).fit(X)
    labels = kmeans.labels_
    #calculate the silhouette score of the clustering
    from sklearn.metrics import silhouette_score
    silhouette_score(X, labels, metric='euclidean')
    print("Silhouette score for seed", seed, "is", silhouette_score(X, labels, metric='euclidean'))
    #calculate the purity score using contingency matrix
    # contingency matrix
    from sklearn.metrics.cluster import contingency_matrix
    contingency_matrix(Y, labels)
    # purity score
    purity = np.sum(np.amax(contingency_matrix(Y, labels), axis=0)) / np.sum(contingency_matrix(Y, labels))
    print("Purity score for seed", seed, "is", purity)
    if seed == 0:
        Y_pred_0 = labels

#3)
# list of inputs with highest variance on the normalized data
var = X.var(axis=0)
var = pd.DataFrame(var)
var = var.sort_values(by=0, ascending=False)
var = var.reset_index()
var = var.rename(columns={'index': 'feature', 0: 'variance'})
var = var['feature'].tolist()
var = var[0:5]

x_axis = var[0]
y_axis = var[1]
# plot the scatter plot of the two features with highest variance
plt.scatter(X[:,x_axis], X[:,y_axis], c=Y)
plt.title('Original Data')
plt.xlabel('feature ' + str(x_axis))
plt.ylabel('feature ' + str(y_axis))
plt.show()

#plot the scatter plot of the two features with highest variance
plt.scatter(X[:,x_axis], X[:,y_axis], c=Y_pred_0)
plt.title('K-means Clustering')
plt.xlabel('feature ' + str(x_axis))
plt.ylabel('feature ' + str(y_axis))
plt.show()

#4)
#how many principal components are necessary to explain 80% of the
variability in the data
from sklearn.decomposition import PCA
pca = PCA(n_components=0.8)
pca.fit(X)
print("Number of principal components necessary to explain 80% of the
variability in the data is", pca.n_components_)
```