Aprendizagem 2021/22

Homework IV – Group 103

I. Pen-and-paper

Homework 4

Henrique Anjos 99081
Vasco Vaz 99125

I. Pen And Paper

1) $\left\{\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}-1\\1\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}\right\} \Rightarrow x_1 = \begin{bmatrix}1\\2\end{bmatrix}, x_2 = \begin{bmatrix}-1\\1\end{bmatrix}, x_3 = \begin{bmatrix}1\\0\end{bmatrix}$

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

$\pi_1 = 0.5 \quad \pi_2 = 0.5$

1) Assignment

	X_1	X_2	X_3
$-\frac{1}{2}(x_n-\mu_1)^T\Sigma_1^{-1}(x_n-\mu_1)$	-0.3333	-2.3333	-1
$-\frac{1}{2}(x_n-\mu_2)^T\Sigma_1^{-1}(x_n-\mu_2)$	-1.25	-0.5	-0.25
$P(x_n \mid k=1)$	0.0658	0.0089	0.0338
$P(x_n \mid k=2)$	0.0228	0.0483	0.0612
$P(k=1)\cdot P(x_n \mid k=1)$	0.0329	0.0045	0.0169
$P(k=2)\cdot P(x_n \mid k=2)$	0.0114	0.0241	0.0310
$P(x_n)$	0.0443	0.0286	0.0479
$P(k=1 \mid x_n)$	0.7428	0.1558	0.3529
$P(k=2 \mid x_n)$	0.2572	0.8442	0.6470

2) Re-estimate

$$\mu_1 = \frac{0.7428 \cdot \begin{bmatrix}1\\2\end{bmatrix} + 0.1558\begin{bmatrix}-1\\1\end{bmatrix} + 0.3529\begin{bmatrix}1\\0\end{bmatrix}}{0.7428 + 0.1558 + 0.3529} = \begin{bmatrix}0.7510\\1.3115\end{bmatrix}$$

$$\mu_2 = \frac{0.2572\begin{bmatrix}1\\2\end{bmatrix} + 0.8442\begin{bmatrix}-1\\1\end{bmatrix} + 0.6470\begin{bmatrix}1\\0\end{bmatrix}}{0.2572 + 0.8442 + 0.6470} = \begin{bmatrix}0.0345\\0.7770\end{bmatrix}$$

$$P(k=1) = \frac{0.7428 + 0.1558 + 0.3529}{3} = 0.4172$$

$$P(k=2) = \frac{0.2572 + 0.8442 + 0.6470}{3} = 0.5828$$

$$\Sigma_1 = \begin{bmatrix}\Sigma_1^{(11)} & \Sigma_1^{(12)}\\ \Sigma_1^{(21)} & \Sigma_1^{(22)}\end{bmatrix}$$

$$\Sigma_1^{(11)} = \frac{0.7428(1-0.7510)^2 + 0.1558(-1-0.7510)^2 + 0.3529(1-0.7510)^2}{0.7428 + 0.1558 + 0.3529} = 0.4361$$

$$\Sigma_1^{(22)} = \frac{0.7428(2-1.3115)^2 + 0.1558(1-1.3115)^2 + 0.3529(0-1.3115)^2}{0.7428 + 0.1558 + 0.3529} = 0.7785$$

$$\Sigma_1^{(12)} = \Sigma_1^{(21)} = \frac{0.7428(1-0.7510)(2-1.3115) + 0.1558(-1-0.7510)(1-1.3115) + 0.3529(1-0.7510)(0-1.3115)}{0.7428 + 0.1558 + 0.3529} = 0.0776$$

$$\Sigma_1 = \begin{bmatrix}0.4361 & 0.0776\\ 0.0776 & 0.7785\end{bmatrix}$$

$$\Sigma_2 = \begin{bmatrix} \Sigma_2^{(11)} & \Sigma_2^{(12)} \\ \Sigma_2^{(21)} & \Sigma_2^{(22)} \end{bmatrix}$$

$$\Sigma_2^{(11)} = \frac{0.2572(1-0.0345)^2 + 0.8442(-1-0.0345)^2 + 0.6470(1-0.0345)^2}{0.2572+0.8442+0.6470} = 0.9988$$

$$\Sigma_2^{(22)} = \frac{0.2572(2-0.777)^2 + 0.8442(1-0.777)^2 + 0.6470(0-0.7770)^2}{0.2572+0.8442+0.6470} = 0.4675$$

$$\Sigma_2^{(12)} = \Sigma_2^{(21)} = \frac{0.2572(1-0.0345)(2-0.777) + 0.8442(-1-0.0345)(1-0.777) + 0.6470(1-0.0345)(0-0.777)}{0.2572+0.8442+0.6470} = -0.2153$$

$$\Sigma_2 = \begin{bmatrix} 0.9988 & -0.2153 \\ -0.2153 & 0.4675 \end{bmatrix}$$

2) a)

	X_1	X_2	X_3
$-\frac{1}{2}(x_n-\mu_1)^T\Sigma_1^{-1}(x_n-\mu_1)$	-0.3425	-3.5555	-1.2731
$-\frac{1}{2}(x_n-\mu_2)^T\Sigma_2^{-1}(x_n-\mu_2)$	-2.8988	-0.5356	-0.8511
$P(x_n \mid k=1)$	0.1957	0.0082	0.0772
$P(x_n \mid k=2)$	0.0135	0.1436	0.1048
$P(k=1)\,P(x_n \mid k=1)$	0.08164	0.0034	0.0322
$P(k=2)\,P(x_n \mid k=2)$	0.0079	0.0837	0.0611

X_1: $\max(0.08164, 0.0079) = 0.08164 \Rightarrow X_1 \in C_1$

X_2: $\max(0.0034, 0.0837) = 0.0837 \Rightarrow X_2 \in C_2$

X_3: $\max(0.0079, 0.0837) = 0.0837 \Rightarrow X_3 \in C_2$

② b) the larger cluster is cluster 2 that has X_2 and X_3

$$\text{Silhouette}(C_2) = \frac{S(X_2) + S(X_3)}{2}$$

$$S(X) = 1 - \frac{a(x)}{b(x)}$$

	X_1	X_2	X_3
X_1	0	$\sqrt{5}$	2
X_2	$\sqrt{5}$	0	$\sqrt{5}$
X_3	2	$\sqrt{5}$	0

$a(X_2) = \sqrt{5}$
$b(X_2) = \sqrt{5}$
$a(X_3) = \sqrt{5}$
$b(X_3) = 2$

$$S(X_2) = 1 - \frac{\sqrt{5}}{\sqrt{5}} = 0$$

$$S(X_3) = 1 - \frac{\sqrt{5}}{2} = -0.118$$

$$S(C_2) = \frac{0 - 0.118}{2} = 0.059$$

Aprendizagem 2021/22

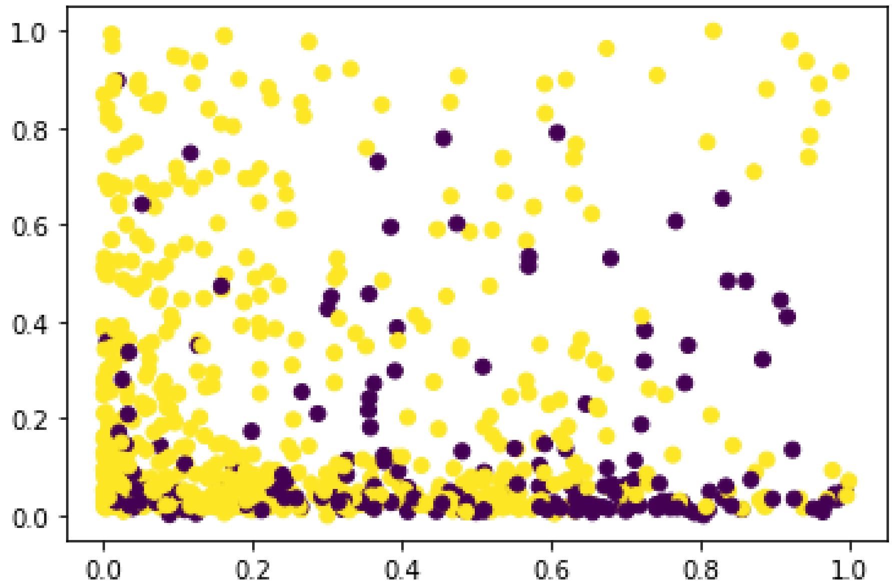
Homework IV – Group 103

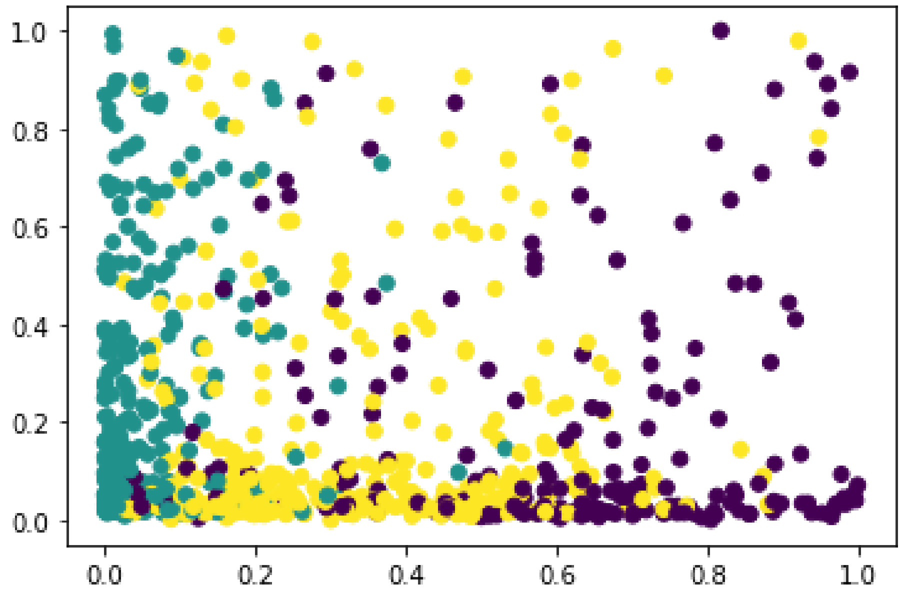
II. Programming and critical analysis

1)
```
Silhouette0: 0.11362027575179431
Silhouette1: 0.11403554201377074
Silhouette2: 0.11362027575179431
Purity0: 0.7671957671957672
Purity1: 0.7632275132275133
Purity2: 0.7671957671957672
```

2) O não-determinismo é causado pelas diferentes origens iniciais dos clusters

3)





4) `Number of components: 31`

III. APPENDIX

```
from sklearn import datasets, metrics, cluster, mixture
from scipy.io.arff import loadarff
import pandas as pd
import numpy as np
from sklearn.decomposition import PCA
from sklearn.preprocessing import MinMaxScaler

# Reading the ARFF file
data = loadarff('pd_speech.arff')
df = pd.DataFrame(data[0])

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

# parameterize clustering
kmeans_algo1 = cluster.KMeans(n_clusters=3, random_state=0)
kmeans_algo2 = cluster.KMeans(n_clusters=3, random_state=1)
kmeans_algo3 = cluster.KMeans(n_clusters=3, random_state=2)

# learn the model
kmeans_model1 = kmeans_algo1.fit(X)
kmeans_model2 = kmeans_algo2.fit(X)
kmeans_model3 = kmeans_algo3.fit(X)

# return centroids
kmeans_model1.cluster_centers_
kmeans_model2.cluster_centers_
kmeans_model3.cluster_centers_

y_pred1 = kmeans_model1.labels_
y_pred2 = kmeans_model2.labels_
y_pred3 = kmeans_model3.labels_

# compute silhouette
print("Silhouette0:",metrics.silhouette_score(X, y_pred1, metric='euclidean'))
print("Silhouette1:",metrics.silhouette_score(X, y_pred2, metric='euclidean'))
print("Silhouette2:",metrics.silhouette_score(X, y_pred3, metric='euclidean'))


# compute purity
```

Aprendizagem 2021/22

Homework IV – Group 103

```
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
)

print("Purity0:",purity_score(y_true, y_pred1))
print("Purity1:",purity_score(y_true, y_pred2))
print("Purity2:",purity_score(y_true, y_pred3))
```

```
# scatter plot

variances = np.var(X, axis = 0)
idx = np.argsort(variances)[::-1]
X_new = X[:,idx[:2]]

plt.scatter(X_new[:,0], X_new[:,1], c=kmeans_model1.labels_)
```

```
plt.scatter(X_new[:,0], X_new[:,1], c=y_true)
```

```
#How many principal components are necessary to explain more than 80% of varia
bility?
pca = PCA(n_components=0.8)

pca.fit(X)
print("Number of components:",pca.n_components_)
```

END