Aprendizagem 2022/23

Homework IV – Group 010

I. Pen-and-paper

1)

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

$u_1 = \begin{pmatrix}2\\2\end{pmatrix};\ u_2 = \begin{pmatrix}0\\0\end{pmatrix};\ \Sigma_1 = \begin{pmatrix}2 & 1\\1 & 2\end{pmatrix};\ \Sigma_2 = \begin{pmatrix}2 & 0\\0 & 2\end{pmatrix};\ \pi_1 = 0.5;\ \pi_2 = 0.5$

<u>**E-step:**</u>

$$p(x_i|c=k) = \mathrm{N}(x_i|u_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2} * \sqrt{|\Sigma_k|}} e^{-\frac{1}{2}*(x_i-u_k)^T * \Sigma_k{}^{-1} * (x_i-u_k)}$$

$\boldsymbol{x_1}$:

$$p(x_1|c=1) = \frac{1}{2\pi * \sqrt{2*2-1*1}} * e^{-\frac{1}{2}*\left(\begin{pmatrix}1\\2\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right)^T * \begin{pmatrix}2 & 1\\1 & 2\end{pmatrix}^{-1} * \left(\begin{pmatrix}1\\2\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right)} = 0.06584$$

$$posterior = p(c=1|x_1) = p(x_1|c=1) * \pi_1 = 0.06584 * 0.5 = 0.03292$$

$$p(x_1|c=2) = \frac{1}{2\pi * \sqrt{2*2-0*0}} * e^{-\frac{1}{2}*\left(\begin{pmatrix}1\\2\end{pmatrix}-\begin{pmatrix}0\\0\end{pmatrix}\right)^T * \begin{pmatrix}2 & 0\\0 & 2\end{pmatrix}^{-1} * \left(\begin{pmatrix}1\\2\end{pmatrix}-\begin{pmatrix}0\\0\end{pmatrix}\right)} = 0.022799$$

$$posterior = p(c=2|x_1) = p(x_1|c=2) * \pi_2 = 0.022799 * 0.5 = 0.0113997$$

Normalized posteriors:

$$p(c=1|x_1) = \frac{0.03292}{0.03292 + 0.0113997} = 0.7428$$

$$p(c=2|x_1) = \frac{0.0113997}{0.03292 + 0.0113997} = 0.2572$$

$\boldsymbol{x_2}$:

$$p(x_2|c=1) = \frac{1}{2\pi * \sqrt{2*2-1*1}} * e^{-\frac{1}{2}*\left(\begin{pmatrix}-1\\1\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right)^T * \begin{pmatrix}2 & 1\\1 & 2\end{pmatrix}^{-1} * \left(\begin{pmatrix}-1\\1\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right)} = 0.008911$$

$$posterior = p(c=1|x_2) = p(x_2|c=1) * \pi_1 = 0.008911 * 0.5 = 0.004455287$$

$$p(x_2|c=2) = \frac{1}{2\pi * \sqrt{2*2-0*0}} * e^{-\frac{1}{2}*\left(\begin{pmatrix}-1\\1\end{pmatrix}-\begin{pmatrix}0\\0\end{pmatrix}\right)^T * \begin{pmatrix}2 & 0\\0 & 2\end{pmatrix}^{-1} * \left(\begin{pmatrix}-1\\1\end{pmatrix}-\begin{pmatrix}0\\0\end{pmatrix}\right)} = 0.048266$$

$$posterior = p(c=2|x_2) = p(x_2|c=2) * \pi_2 = 0.048266 * 0.5 = 0.024133$$

Normalized posteriors:

$$p(c=1|x_2) = \frac{0.004455287}{0.004455287 + 0.024133} = 0.155843$$

$$p(c=2|x_2) = \frac{0.024133}{0.004455287 + 0.024133} = 0.844157$$

$\boldsymbol{x_3}$:

$$p(x_3|c=1) = \frac{1}{2\pi * \sqrt{2*2-1*1}} * e^{-\frac{1}{2}*\left(\begin{pmatrix}1\\0\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right)^T * \begin{pmatrix}2 & 1\\1 & 2\end{pmatrix}^{-1} * \left(\begin{pmatrix}1\\0\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right)} = 0.03380376099$$

$$posterior = p(c=1|x_3) = p(x_3|c=1) * \pi_1 = 0.03380376099 * 0.5 = 0.0169018805$$

$$p(x_1|c=2) = \frac{1}{2\pi * \sqrt{2*2-0*0}} * e^{-\frac{1}{2}*\left(\binom{1}{0}-\binom{0}{0}\right)^T * \begin{pmatrix}2 & 0\\0 & 2\end{pmatrix}^{-1} * \left(\binom{1}{0}-\binom{0}{0}\right)} = 0.06197499715$$

$$posterior = p(c=2|x_3) = p(x_3|c=2) * \pi_2 = 0.06197499715 * 0.5 = 0.03098749858$$

Normalized posteriors:

$$p(c=1|x_3) = \frac{0.0169018805}{0.0169018805 + 0.03098749858} = 0.3529358873$$

$$p(c=2|x_3) = \frac{0.03098749858}{0.0169018805 + 0.03098749858} = 0.6470641127$$

<u>**M-step:**</u>

$$u_k = \frac{\sum_{i=1}^{3} p(c=k|x_i) * x_i}{\sum_{i=0}^{3} p(c=k|x_i)}$$

$$\Sigma_k = \frac{\sum_{i=1}^{3} p(c=k|x_i) * (x_i - u_k) * (x_i - u_k)^T}{\sum_{i=0}^{3} p(c=k|x_i)}$$

$$\pi_k = \frac{\sum_{i=1}^{3} p(c=k|x_i)}{\sum_{j=1}^{3}\sum_{i=0}^{3} p(c=j|x_i)}$$

$\boldsymbol{c=1}$:

$$u_1 = \frac{0.7428 * \binom{1}{2} + 0.155843 * \binom{-1}{1} + 0.3529358873 * \binom{1}{0}}{0.7428 + 0.155843 + 0.3529358873} = \begin{pmatrix}0.75\\1.31\end{pmatrix}$$

$$\Sigma_1 = \frac{0.7428\binom{1-0.75}{2-1.31}(1-0.75 \quad 2-1.31) + 0.155843\binom{-1-0.75}{1-1.31}(-1-0.75 \quad 1-1.31) + 0.3529358873\binom{1-0.75}{0-1.31}(1-0.75 \quad 0-1.31)}{0.7428 + 0.155843 + 0.3529358873}$$

$$= \begin{pmatrix}0.436 & 0.0776\\0.0776 & 0.7785\end{pmatrix}$$

$$\pi_1 = \frac{0.7428 + 0.155843 + 0.3529358873}{3} = 0.417$$

$\boldsymbol{c=2}$:

$$u_2 = \frac{0.2572 * \binom{1}{2} + 0.844157 * \binom{-1}{1} + 0.6470641127 * \binom{1}{0}}{0.2572 + 0.844157 + 0.6470641127} = \begin{pmatrix}0.034\\0.777\end{pmatrix}$$

$$\Sigma_2 = \frac{0.2572\binom{1-0.034}{2-0.777}(1-0.034 \quad 2-0.777) + 0.844157\binom{-1-0.034}{1-0.777}(-1-0.034 \quad 1-0.777) + 0.6470641127\binom{1-0.034}{0-0.777}(1-0.034 \quad 0-0.777)}{0.2572 + 0.844157 + 0.6470641127}$$

$$= \begin{pmatrix}0.9988177 & -0.2153\\-0.2153 & 0.4675\end{pmatrix}$$

$$\pi_2 = \frac{0.2572 + 0.844157 + 0.6470641127}{3} = 0.5828$$

TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 010

2)

a.

$\boldsymbol{x_1}$:

$$p(c=1|x_1)=\pi_1 * p(x_1|c=1)$$

$$= 0.417 * \frac{1}{2\pi * \sqrt{0.436 * 0.7785 - 0.0776 * 0.0776}} * e^{-\frac{1}{2}*\left(\binom{1}{2}-\binom{0.75}{1.31}\right)^T * \begin{pmatrix} 0.436 & 0.0776 \\ 0.0776 & 0.7785 \end{pmatrix}^{-1} * \left(\binom{1}{2}-\binom{0.75}{1.31}\right)}$$

$$= 0.08164192$$

$$p(c=2|x_1)=\pi_2 * p(x_1|c=2)$$

$$= 0.5828 * \frac{1}{2\pi * \sqrt{0.9988177 * 0.4675 - (-0.2153)^2}} * e^{-\frac{1}{2}*\left(\binom{1}{2}-\binom{0.034}{0.777}\right)^T * \begin{pmatrix} 0.9988177 & -0.2153 \\ -0.2153 & 0.4675 \end{pmatrix}^{-1} * \left(\binom{1}{2}-\binom{0.034}{0.777}\right)}$$

$$= 0.00787938$$

Normalized:

$$p(c=1|x_1) = \frac{0.08164192}{0.08164192 + 0.00787938} = 0.91198316$$

$$p(c=2|x_1) = \frac{0.00787938}{0.08164192 + 0.00787938} = 0.08801684$$

$$p(c=1|x_1) > p(c=2|x_1) \Rightarrow \text{cluster } 1$$

$\boldsymbol{x_2}$:

$$p(c=1|x_2)=\pi_1 * p(x_2|c=1)$$

$$= 0.417 * \frac{1}{2\pi * \sqrt{0.436 * 0.7785 - 0.0776 * 0.0776}} * e^{-\frac{1}{2}*\left(\binom{-1}{1}-\binom{0.75}{1.31}\right)^T * \begin{pmatrix} 0.436 & 0.0776 \\ 0.0776 & 0.7785 \end{pmatrix}^{-1} * \left(\binom{-1}{1}-\binom{0.75}{1.31}\right)}$$

$$= 0.00341898$$

$$p(c=2|x_2)=\pi_2 * p(x_2|c=2)$$

$$= 0.5828 * \frac{1}{2\pi * \sqrt{0.9988177 * 0.4675 - (-0.2153)^2}} * e^{-\frac{1}{2}*\left(\binom{-1}{1}-\binom{0.034}{0.777}\right)^T * \begin{pmatrix} 0.9988177 & -0.2153 \\ -0.2153 & 0.4675 \end{pmatrix}^{-1} * \left(\binom{-1}{1}-\binom{0.034}{0.777}\right)}$$

$$= 0.08371795$$

Normalized:

$$p(c=1|x_2) = \frac{0.00341898}{0.00341898 + 0.08371795} = 0.03923683$$

$$p(c=2|x_2) = \frac{0.08371795}{0.00341898 + 0.08371795} = 0.96076317$$

$$p(c=2|x_2) > p(c=1|x_2) \Rightarrow \text{cluster } 2$$

$\boldsymbol{x_3}$:

$$p(c=1|x_3)=\pi_1 * p(x_3|c=1)$$

$$= 0.417 * \frac{1}{2\pi * \sqrt{0.436 * 0.7785 - 0.0776 * 0.0776}} * e^{-\frac{1}{2}*\left(\binom{1}{0}-\binom{0.75}{1.31}\right)^T * \begin{pmatrix} 0.436 & 0.0776 \\ 0.0776 & 0.7785 \end{pmatrix}^{-1} * \left(\binom{1}{0}-\binom{0.75}{1.31}\right)}$$

$$= 0.03219282$$

$$p(c=2|x_3) = \pi_2 * p(x_3|c=2)$$

$$= 0.5828 * \frac{1}{2\pi * \sqrt{0.9988177 * 0.467469 - (-0.2153)^2}} * e^{-\frac{1}{2} * \left(\binom{1}{0} - \binom{0.034}{0.777}\right)^T * \begin{pmatrix} 0.9988177 & -0.2153 \\ -0.2153 & 0.467469 \end{pmatrix}^{-1} * \left(\binom{1}{0} - \binom{0.034}{0.777}\right)}$$

$$= 0.06106939$$

Normalized:

$$p(c=1|x_3) = \frac{0.03219282}{0.03219282 + 0.06106939} = 0.3451861$$

$$p(c=2|x_3) = \frac{0.06106939}{0.03219282 + 0.06106939} = 0.6548139$$

$p(c=2|x_3) > p(c=1|x_3)$ => cluster 2

b. Being cluster 2 the larger cluster

$$S(x_2) = \frac{\|x_2 - x_1\|_2}{\|x_2 - x_3\|_2} - 1 = \frac{\sqrt{(-1-1)^2 + (1-2)^2}}{\sqrt{(-1-1)^2 + (1-0)^2}} - 1 = 0$$

$$S(x_3) = \frac{\|x_3 - x_1\|_2}{\|x_3 - x_2\|_2} - 1 = \frac{\sqrt{(1-1)^2 + (0-2)^2}}{\sqrt{\left(1-(-1)\right)^2 + (0-1)^2}} - 1 = -0.1055728$$

$$S(c_2) = \frac{S(x_2)+S(x_3)}{2} = \frac{0-0.1055728}{2} = -0.0527864$$

II. Programming and critical analysis

1) Silhouette score for k-means with random_state = 0 : 0.1136202757517943

Purity score for k-means with random_state = 0 : 0.7671957671957672

Silhouette score for k-means with random_state = 1 : 0.11403554201377072

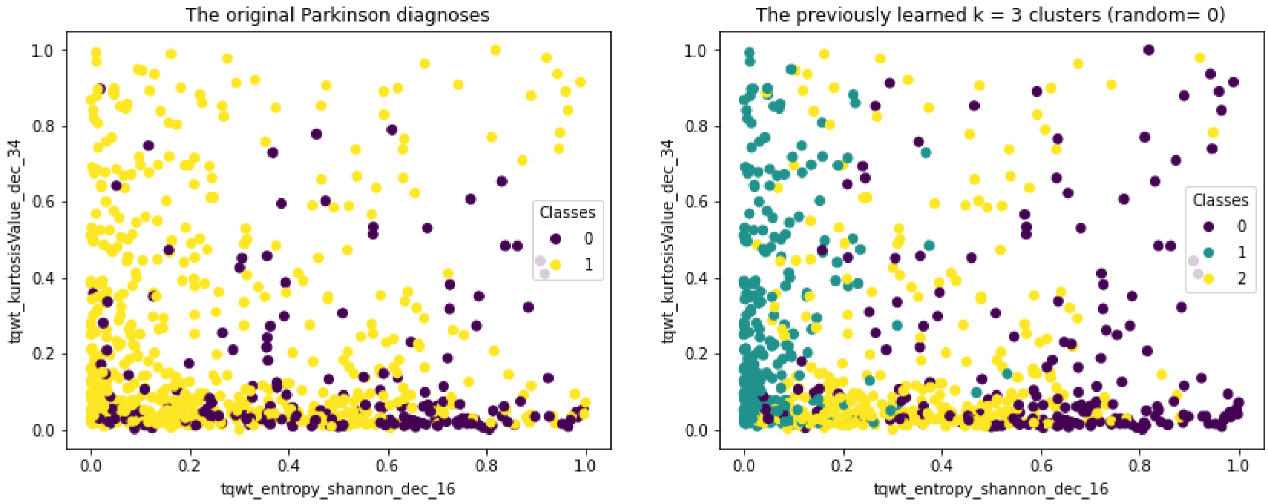
Purity score for k-means with random_state = 1 : 0.7632275132275133

Silhouette score for k-means with random_state = 2 : 0.1136202757517943

Purity score for k-means with random_state = 2 : 0.7671957671957672

2) The non-determinism is caused by the random initialization of the centroids. The algorithm will converge to a local minimum, and because the initial centroids are different, the algorithm will converge to different local minimums.

3)



4) 31 principal components are needed to explain more than 80% of variability.

Aprendizagem 2022/23

Homework IV – Group 010

III. APPENDIX

```python
import pandas as pd, numpy as np
from scipy.io.arff import loadarff
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
from sklearn import metrics, cluster
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA

# Load the data
data = loadarff('pd_speech.arff')
df = pd.DataFrame(data[0])
y = pd.to_numeric(df['class'])
df = df.drop('class', axis=1)

df_norm = pd.DataFrame(MinMaxScaler().fit_transform(df), columns=df.columns)

# 1)

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

# K-means
kmeans = []
for i in range(3):
    kmeans += [KMeans(n_clusters=3, random_state=i)]
    kmeans[i].fit(df_norm)
    y_pred = kmeans[i].labels_
    print("Silhouette score for k-means with random_state =", i, ":",
metrics.silhouette_score(df_norm, y_pred, metric='euclidean'))
    print("Purity score for k-means with random_state =", i, ":", purity_score(y,
y_pred))


# 3)
sorted_by_variance = df_norm.var().sort_values(ascending=False)
features = sorted_by_variance[:2].index

plt.figure(figsize=(14, 5))
plt.subplot(121)
scatter = plt.scatter(df_norm[features[0]], df_norm[features[1]], c = y)
plt.xlabel(features[0])
plt.ylabel(features[1])
plt.legend(*scatter.legend_elements(), loc="best", title="Classes")
```

```
plt.title('The original Parkinson diagnoses')

plt.subplot(122)
scatter = plt.scatter(df_norm[features[0]], df_norm[features[1]],
c=kmeans[0].labels_)
plt.xlabel(features[0])
plt.ylabel(features[1])
plt.legend(*scatter.legend_elements(), loc="best", title="Classes")
plt.title('The previously learned k=3 clusters (random=0)')


plt.show()

# 4) PCA
pca = PCA(n_components=0.8)
pca.fit(df_norm)
print(pca.n_components_, "principal components are needed to explain more than 80% of
variability.")
```

END