TÉCNICO LISBOA

Aprendizagem 2021/22

Homework IV – Group 105

I. Pen-and-paper

1)

$$x_1: \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \quad \mu_1 = \begin{bmatrix} 2 \\ 2 \end{bmatrix} \quad \pi_1 = 0.5$$

$$x_2: \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \pi_2 = 0.5$$

$$x_3: \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\sqrt{|\Sigma_1|} = \sqrt{2 \times 2 - 1 \times 1} = \sqrt{4-1} = \sqrt{3}$$

$$\sqrt{|\Sigma_2|} = \sqrt{2 \times 2 - 0 \times 0} = \sqrt{4} = 2$$

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

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

$$p(x_n | c=1) = \frac{1}{2\pi \times \sqrt{|\Sigma_1|}} \cdot \exp\left(-\frac{1}{2} \cdot (x_n - \mu_1)^T \cdot \Sigma_1^{-1} \cdot (x_n - \mu_1)\right)$$

$$= \frac{1}{2\pi \times \sqrt{3}} \cdot \exp\left(-\frac{1}{2} (x_n - \mu_1)^T \cdot \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} \end{bmatrix} \cdot (x_n - \mu_1)\right)$$

para x_1:

$$(x_1 - \mu_1) = \begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

$$(x_1 - \mu_1)^T = \begin{bmatrix} -1 & 0 \end{bmatrix}$$

$$p(x_1 \mid c=1) = \frac{1}{2\pi \times \sqrt{3}} \cdot \exp\left(-\frac{1}{2} \cdot \begin{bmatrix} -1 & 0 \end{bmatrix} \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} -1 \\ 0 \end{bmatrix}\right)$$

$$= \frac{e^{-\frac{1}{3}}}{2\pi \times \sqrt{3}} = 0.06584073599896272$$

$$p(x_2 \mid c=1) = 0.008910574654926666$$

$$p(x_3 \mid c=1) = 0.03380376099157291$$

$$p(x_n \mid c=2) = \frac{1}{2\pi \times \sqrt{|\Sigma_2|}} \cdot \exp\left(-\frac{1}{2} \cdot (x_n - \mu_2)^T \cdot \Sigma_2^{-1} \cdot (x_n - \mu_2)\right)$$

$$= \frac{1}{2\pi \times 2} \cdot \exp\left(-\frac{1}{2} (x_n - \mu_2)^T \cdot \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \cdot (x_n - \mu_2)\right)$$

para x_1:

$$(x_1 - \mu_2) = \begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$(x_1 - \mu_2)^T = \begin{bmatrix} 1 & 2 \end{bmatrix}$$

$$p(x_1 \mid c=2) = \frac{1}{4\pi} \cdot \exp\left(-\frac{1}{2} \cdot \begin{bmatrix} 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix}\right)$$

$$= \frac{e^{-\frac{5}{4}}}{4\pi} = 0.02279932731$$

$p(x_2 | c=2) = 0.04826617631502696$

$p(x_3 | c=2) = 0.06197499715482649$

$p(x_n, c=1) = p(x_n | c=1) . \pi_1$

$p(x_1, c=1) = 0.03292036799948136$

$p(x_2, c=1) = 0.00445528732746333$

$p(x_3, c=1) = 0.01690188049578645$

$p(x_n, c=2) = p(x_n | c=2) . \pi_2$

$p(x_1, c=2) = 0.01139966365995964 7$

$p(x_2, c=2) = 0.02413308815751348$

$p(x_3, c=2) = 0.03098749857711324 4$

$$p(c=1 | x_n) = \frac{p(x_n, c=1)}{p(x_n, c=1) + p(x_n, c=2)}$$

$p(c=1 | x_1) = 0.7427875560298409$

$p(c=1 | x_2) = 0.1558426196621275$

$p(c=1 | x_3) = 0.3529358873071136$

$$p(c=2 | x_n) = \frac{p(x_n, c=2)}{p(x_n, c=1) + p(x_n, c=2)}$$

$p(c=2 | x_1) = 0.2572124439701591$

$p(c=2 | x_2) = 0.8441573803378726$

$p(c=2 | x_3) = 0.6470641126928863$

Maximização

$$P_{c_1} = \begin{bmatrix} p(c=1 | x_1) \\ p(c=1 | x_2) \\ p(c=1 | x_3) \end{bmatrix} \qquad P_{c_2} = \begin{bmatrix} p(c=2 | x_1) \\ p(c=2 | x_2) \\ p(c=2 | x_3) \end{bmatrix}$$

$$N_1 = \sum P_{c_1} = 1.2515660629990821$$

$$N_2 = \sum P_{c_2} = 1.7484339370000918$$

Priors:

$$p(c=1) = \frac{N_1}{N_1 + N_2} = 0.4171886876663607$$

$$p(c=2) = \frac{N_2}{N_1 + N_2} = 0.5828113123336394$$

Centroids:

$$\mu_1 = \frac{p_{c_{1,1}} x_1 + p_{c_{1,2}} x_2 + p_{c_{1,3}} x_3}{N_1} =$$

$$\frac{0.7427875560298409 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 0.155842619662127 5 \begin{bmatrix} -1 \\ 1 \end{bmatrix} + 0.3529358873071136 \begin{bmatrix} 1 \\ 0 \end{bmatrix}}{1.2515660629990821}$$

$$= \begin{bmatrix} 0.75096381 \\ 1.31149108 \end{bmatrix}$$

$$\mu_2 = \frac{p_{c_{2,1}} x_1 + p_{c_{2,2}} x_2 + p_{c_{2,3}} x_3}{N_2} =$$

$$= \frac{0.2572124439701591 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 0.8441573803378726 \begin{bmatrix} -1 \\ 1 \end{bmatrix} + 0.6470641126928863 \begin{bmatrix} 1 \\ 0 \end{bmatrix}}{1.7484339370000918} =$$

$$= \begin{bmatrix} 0.03438459 \\ 0.77702808 \end{bmatrix}$$

Σ Novo:

$$\Sigma_{x_{n_{11}}} = (x_{n_1} - \mu_{1_1})^2$$

$$\Sigma_{x_{n_{22}}} = (x_{n_2} - \mu_{1_2})^2$$

$$\Sigma_{x_{n_{12}}} = \Sigma_{x_{n_{21}}} = (x_{n_1} - \mu_{1_1})(x_{n_2} - \mu_{1_2})$$

Σ_1

exemplo para x_1:

$$\Sigma_{x_{1_{11}}} = (1 - 0.7509638 1)^2 = 0.06201902 \qquad \Sigma_{x_{1_{22}}} = (2 - 1.31149108)^2 = 0.47404453$$

$$\Sigma_{x_{1_{12}}} = \Sigma_{x_{1_{21}}} = (1 - 0.75096381).(2 - 1.31149108) = 0.17146363$$

$$\Sigma_{x_1} = \begin{bmatrix} 0.06201902 & 0.17146363 \\ 0.17146363 & 0.47404453 \end{bmatrix}$$

$$\Sigma_{x_2} = \begin{bmatrix} 3.06587428 & 0.54540962 \\ 0.54540962 & 0.09702669 \end{bmatrix}$$

$$\Sigma_{x_3} = \begin{bmatrix} 0.06201902 & -0.32660874 \\ -0.32660874 & 1.72000886 \end{bmatrix}$$

$$\Sigma_1 = \frac{pc_{1_1}\Sigma_{x_1} + pc_{1_2}\Sigma_{x_2} + pc_{1_3}\Sigma_{x_3}}{N_1} = \begin{bmatrix} 0.43605335 & 0.07757255 \\ 0.07757255 & 0.77845521 \end{bmatrix}$$

Σ_2

exemplo para x_1:

$$\Sigma_{x_{1_{11}}} = (1 - 0.03438459)^2 = 0.93241313 \qquad \Sigma_{x_{1_{22}}} = (2 - 0.77702808)^2 = 1.49566031$$

$$\Sigma_{x_{1_{12}}} = \Sigma_{x_{1_{21}}} = (1 - 0.03438459).(2 - 0.77702808) = 1.18092054$$

$$\Sigma_{x_1} = \begin{bmatrix} 0.93241313 & 1.18092054 \\ 1.18092054 & 1.49566031 \end{bmatrix}$$

$$\Sigma_{x_2} = \begin{bmatrix} 1.06995147 & -0.23063872 \\ -0.23063872 & 0.04971648 \end{bmatrix}$$

$$\Sigma_{x_3} = \begin{bmatrix} 0.93241313 & -0.75031029 \\ -0.75031029 & 0.60377264 \end{bmatrix}$$

TÉCNICO LISBOA

Aprendizagem 2021/22

Homework IV – Group 105

$$\Sigma_2 = \frac{pc_{2_1}\Sigma_{x_1} + pc_{2_2}\Sigma_{x_2} + pc_{2_3}\Sigma_{x_3}}{N_2} = \begin{bmatrix} 0.9988177 & -0.21530512 \\ -0.21530512 & 0.46747582 \end{bmatrix}$$

2)

a.

$$P(c=1 \mid x_n) = \frac{P(c=1, x_n)}{P(x_n)}$$

$$P(x_n) = P(x_n, c=1) + P(x_n, c=2)$$

para $x = 1$:

$$P(x_1) = P(x_1, c=1) + P(x_1, c=2) =$$

$$= 0.0816419154145976 3 + 0.0078793820552040 85 =$$

$$= 0.08952129746980171$$

$$P(c=1 \mid x_1) = \frac{0.08164191541459763}{0.08952129746980171} = 0.9119831562 08219$$

$$P(c=1 \mid x_2) = 0.03923682864802956$$

$$P(c=1 \mid x_3) = 0.3451861042649158$$

$$P(c=2 \mid x_n) = \frac{P(c=2, x_n)}{P(x_n)}$$

$$P(x_n) = P(x_n, c=1) + P(x_n, c=2)$$

Aprendizagem 2021/22

Homework IV – Group 105

para $x = 1$:

$$P(x_1) = P(x_1, c=1) + P(x_1, c=2) =$$
$$= 0.0816419154149763 + 0.00787938205520408S =$$
$$= 0.0895212974698017 1$$

$$P(c=2 \mid x_1) = \frac{0.00787938205520408S}{0.08952129746980171} = 0.08801684378917814$$

$$P(c=2 \mid x_2) = 0.9607631713519704$$
$$P(c=2 \mid x_3) = 0.6548138957350843$$

$x_1 \in c=1$ pois $P(c=1 \mid x_1) > P(c=2 \mid x_1)$

$x_2 \in c=2$ pois $P(c=2 \mid x_2) > P(c=1 \mid x_2)$

$x_3 \in c=2$ pois $P(c=2 \mid x_3) > P(c=1 \mid x_3)$

b.

$a_i =$ distância média de x_i aos pontos no seu cluster

$b_i = \min($distância média de x_i aos pontos noutro cluster$)$

$s = 1 - \frac{a}{b}$ se $a < b$, $s = \frac{b}{a} - 1$ se não

$$dist(x_1, x_3) = dist(x_3, x_1) = \sqrt{(x_{1_1} - x_{3_1})^2 + (x_{1_2} - x_{3_2})^2} = \sqrt{4} = 2$$

$$dist(x_1, x_2) = \sqrt{(x_{1_1} - x_{2_1})^2 + (x_{1_2} - x_{2_2})^2} = \sqrt{5}$$

$$dist(x_3, x_2) = \sqrt{(x_{3_1} - x_{2_1})^2 + (x_{3_2} - x_{2_2})^2} = \sqrt{5}$$

$$s_{x_2} = 1 - \frac{dist(x_2, x_3)}{dist(x_1, x_2)} \text{ ou } \frac{dist(x_2, x_3)}{dist(x_1, x_2)} - 1 = 1 - \frac{\sqrt{5}}{\sqrt{5}} = 0$$

$$s_{x_3} = 1 - \frac{dist(x_3, x_2)}{dist(x_3, x_1)} \text{ ou } \frac{dist(x_3, x_2)}{dist(x_3, x_1)} - 1 = \frac{\sqrt{5}}{2} - 1 = 0.11803398875$$

$$C_2 s = \frac{s_{x_2} + s_{x_3}}{2} = 0.05901699437$$

Aprendizagem 2021/22

Homework IV – Group 105

II. Programming and critical analysis

3) Silhouette 0: 0.11362027575179426

Purity 0: 0.7671957671957672

Silhouette 1: 0.11403554201377068

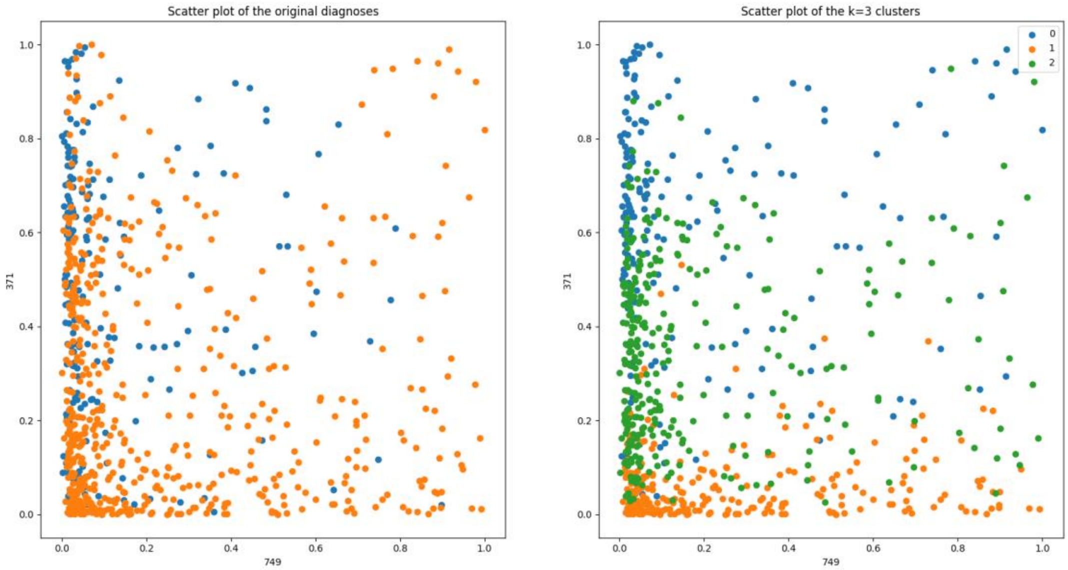
Purity 1: 0.7632275132275133

Silhouette 2: 0.11362027575179426

Purity 2: 0.7671957671957672

4) Como podemos observar no código, foram utilizadas várias seeds possíveis, o que afeta a inicialização dos centroides e, alterando estes valores, iremos obter resultados diferentes, ou seja, não determinísticos.

5)



6) Number of primary components to explain 80% variability: 31

III. APPENDIX

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

```
def warn(*args, **kwargs):
    pass
warnings.warn = warn

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

# Scale the dataframe
scaler = MinMaxScaler()
df = scaler.fit_transform(df)
X_list = df[:, :-1]
df = pd.DataFrame(df)

X, y = df[list(df.columns[:-1])], df[[752]]

temp_y = y.to_numpy()
y_true = [int(x) for sublist in temp_y for x in sublist]

kmeans = []
kmeans_model = []
silhouettes = []
purities = []

for i in range(3):
    # Generate 3 KMeans clusterings with 3 different seeds (0, 1, 2)
    kmeans.append(cluster.KMeans(n_clusters=3, random_state=i))
    kmeans_model.append(kmeans[i].fit(X))

    y_pred = kmeans_model[i].labels_
    confusion_matrix = metrics.cluster.contingency_matrix(y_true, y_pred)

    # Calculate silhouette and purity for the model
    silhouette = metrics.silhouette_score(X, y_pred)
    silhouettes.append(silhouette)
    print("Silhouette", str(i) + ":", silhouette)
    purity = np.sum(np.amax(confusion_matrix, axis=0)) / np.sum(confusion_matrix)
    purities.append(purity)
    print("Purity", str(i) + ":", purity)

# Fix random = 0
y_pred = kmeans_model[0].labels_

# Get the indexes for the 2 features with the biggest variances
variances = X.var().to_numpy()
indexes = np.argpartition(variances, -2)[-2:]
```

```
scatter_X = X_list[:, indexes[0]]
scatter_Y = X_list[:, indexes[1]]

fig = plt.figure()
ax1 = fig.add_subplot(121)
ax2 = fig.add_subplot(122)

ax1.set_title("Scatter plot of the original diagnoses")
for g in np.unique(y_true):
    # Select the indexes where we find the specified label
    ix = np.where(y_true == g)
    ax1.scatter(scatter_X[ix], scatter_Y[ix], label=g)
ax1.set_xlabel(X.columns[indexes[0]])
ax1.set_ylabel(X.columns[indexes[1]])

ax2.set_title("Scatter plot of the k=3 clusters")
for g in np.unique(y_pred):
    ix = np.where(y_pred == g)
    ax2.scatter(scatter_X[ix], scatter_Y[ix], label=g)
ax2.set_xlabel(X.columns[indexes[0]])
ax2.set_ylabel(X.columns[indexes[1]])
plt.legend()
plt.show()

# Calculate number of primary components needed
components = 0
size = len(X_list[0])
for i in range(size):
    pca = PCA(n_components=i)
    pca.fit(X)
    if sum(pca.explained_variance_ratio_) > 0.8:
        components = i
        break

print("Number of primary components to explain 80% variability:", components)
```

END