TÉCNICO LISBOA

Aprendizagem 2021/22

Homework IV – Group 057

I. Pen-and-paper

1) $u_1 = x_1 = \binom{2}{4}$, $u_2 = x_2 = \binom{-1}{-4}$, $|\Sigma_1|=1$, $|\Sigma_2|=4$

Distribuição Normal com duas dimensões: $p(x) = \frac{1}{2\pi \times |\sum|^{1/2}} e^{-\frac{1}{2}(x-u)^T \Sigma^{-1}(x-u)}$

E-Step:

Para x_1:

Cluster 1:

Prior: $\pi_1 = p(c_1=1) = 0.7$

$p(x_1|c_1=1) = \frac{1}{2\pi} e^{-\frac{1}{2}\left(\binom{2}{4}-\binom{2}{4}\right)^T \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}^{-1} \left(\binom{2}{4}-\binom{2}{4}\right)} = \frac{1}{2\pi} e^0 = \frac{1}{2\pi}$

$p(c_1=1, x_1) = p(x_1|c_1=1)p(c_1=1) \approx 0.1114$

Cluster 2:

Prior: $\pi_2 = p(c_2=1) = 0.3$

$p(x_1|c_2=1) = \frac{1}{2\pi * 2} e^{-\frac{1}{2}\left(\binom{2}{4}-\binom{-1}{-4}\right)^T \begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}^{-1} \left(\binom{2}{4}-\binom{-1}{-4}\right)} = \frac{1}{4\pi} e^{-\frac{1}{8}(3 \quad 8)\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}\binom{3}{8}} = \frac{1}{4\pi} e^{-18,25}$

$p(c_2=1, x_1) = p(x_1|c_2=1)p(c_2=1) \approx 2.832\text{x}10^{-10}$

$p(c_1=1|x_1) = \frac{p(c_1=1,x_1)}{p(c_1=1,x_1)+p(c_2=1,x_1)} \approx \frac{0.1114}{0.1114 + 2.832\text{x}10^{-10}} \approx 0.9999999974578$

$p(c_2=1|x_1) = \frac{p(c_2=1,x_1)}{p(c_1=1,x_1)+p(c_2=1,x_1)} \approx \frac{2.832\text{x}10^{-10}}{0.1114 + 2.832\text{x}10^{-10}} \approx 2.5419\text{x}10^{-9}$

Para x_2:

Cluster 1:

$p(x_2|c_1=1) = \frac{1}{2\pi} e^{-\frac{1}{2}\left(\binom{-1}{-4}-\binom{2}{4}\right)^T \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}^{-1} \left(\binom{-1}{-4}-\binom{2}{4}\right)} = \frac{1}{2\pi} e^{-\frac{1}{2}(-3 \quad -8)\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\binom{-3}{-8}} = \frac{1}{2\pi} e^{-36,5}$

$p(c_1=1, x_2) = p(x_2|c_1=1)p(c_1=1) \approx 1.5674\times10^{-17}$

Cluster 2:

$p(x_2|c_2=1) = \frac{1}{2\pi * 2} e^{-\frac{1}{2}\left(\binom{-1}{-4}-\binom{-1}{-4}\right)^T \begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}^{-1} \left(\binom{-1}{-4}-\binom{-1}{-4}\right)} = \frac{1}{4\pi} e^0 = \frac{1}{4\pi}$

$p(c_2=1, x_2) = p(x_2|c_2=1)p(c_2=1) \approx 0.02387$

$p(c_1=1|x_2) = \frac{p(c_1=1,x_2)}{p(c_1=1,x_2)+p(c_2=1,x_2)} \approx \frac{1.5674\times10^{-17}}{0,02387 + 1.5674\times10^{-17}} \approx 0$

$p(c_2=1|x_2) = \frac{p(c_2=1,x_1)}{p(c_1=1,x_2)+p(c_2=1,x_2)} \approx \frac{0,02387}{0,02387 + 1.5674\times10^{-17}} \approx 1$

Para x_3:

Cluster 1:

$p(x_3|c_1=1) = \frac{1}{2\pi} e^{-\frac{1}{2}\left(\binom{-1}{2}-\binom{2}{4}\right)^T \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}^{-1} \left(\binom{-1}{2}-\binom{2}{4}\right)} = 0.00023$

$p(c_1=1, x_3) = p(x_3|c_1=1)p(c_1=1) \approx 0.00016$

Cluster 2:

$p(x_3|c_2=1) = \frac{1}{2\pi * 2} e^{-\frac{1}{2}\left(\binom{-1}{2}-\binom{-1}{-4}\right)^T \begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}^{-1} \left(\binom{-1}{2}-\binom{-1}{-4}\right)}$

$p(c_2=1, x_3) = p(x_3|c_2=1)p(c_2=1) \approx 2.94619\times10^{-6}$

$p(c_1=1|x_3) = \frac{p(c_1=1,x_3)}{p(c_1=1,x_3)+p(c_2=1,x_3)} \approx \frac{0.00016}{0.00016 + 2.94619\times10^{-6}} \approx 0.9819$

$p(c_2=1|x_3) = \frac{p(c_2=1,x_3)}{p(c_1=1,x_3)+p(c_2=1,x_3)} \approx \frac{2.94619\times10^{-6}}{0.00016 + 2.94619\times10^{-6}} \approx 0.01808$

TÉCNICO LISBOA

Aprendizagem 2021/22

Homework IV – Group 057

Para x_4:

Cluster 1:

$$p(x_4|c_1=1) = \frac{1}{2\pi} e^{-\frac{1}{2}\left(\binom{4}{0}-\binom{2}{4}\right)^T \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}^{-1} \left(\binom{4}{0}-\binom{2}{4}\right)}$$

$$p(c_1=1, x_4) = p(x_4|c_1=1)p(c_1=1) \approx 5.05794\times10^{-6}$$

Cluster 2:

$$p(x_4|c_2=1) = \frac{1}{2\pi*2} e^{-\frac{1}{2}\left(\binom{4}{0}-\binom{-1}{-4}\right)^T \begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}^{-1} \left(\binom{4}{0}-\binom{-1}{-4}\right)}$$

$$p(c_2=1, x_4) = p(x_4|c_2=1)p(c_2=1) \approx 8.44098\times10^{-7}$$

$$p(c_1=1|x_4) = \frac{p(c_1=1,x_4)}{p(c_1=1,x_4)+p(c_2=1,x_4)} \approx \frac{5.05794\times10^{-6}}{5.05794\times10^{-6}+8.44098\times10^{-7}} \approx 0.85698$$

$$p(c_2=1|x_4) = \frac{p(c_2=1,x_4)}{p(c_1=1,x_4)+p(c_2=1,x_4)} \approx \frac{8.44098\times10^{-7}}{5.05794\times10^{-6}+8.44098\times10^{-7}} \approx 0.14302$$

M-Step:

Cluster 1:

$$\mu_1 = \frac{\sum_x p(c_1=1|x)x}{\sum_x p(c_1=1|x)} = \frac{0.9999999974578\binom{2}{4} + 0\binom{-1}{-4} + 0.9819\binom{-1}{2} + 0.85698\binom{4}{0}}{0.9999999974578 + 0 + 0.9819 + 0.85698} = \begin{pmatrix}1.56611\\ 2.100757\end{pmatrix}$$

$$\sum_{11}^1 = \frac{0.9999999974578(2-1.56611)^2+0+0.9819(-1-1.56611)^2+0.85698(4-1.56611)^2}{0.9999999974578 + 0 + 0.9819 + 0.85698} = 4.1321$$

$$\sum_{12}^1 = \frac{0.9999999974578(2-1.56611)(4-2.100757)+0+0.9819(-1-1.56611)(2-2.100757)+0.85698(4-1.56611)(0-2.100757)}{0.9999999974578 + 0 + 0.9819 + 0.85698} = -1.16377$$

$$\sum_{22}^1 = \frac{0.9999999974578(4-2.100757)^2+0+0.9819(2-2.100757)^2+0.85698(0-2.100757)^2}{0.9999999974578 + 0 + 0.9819 + 0.85698} = 2.60634$$

$$\sum_1 = \begin{pmatrix}4.1321 & -1.16377\\ -1.16377 & 2.60634\end{pmatrix}$$

$$\text{Prior: } p(c_1=1) = \frac{0.9999999974578+0+0.9819+0.85698}{(0.9999999974578+0+0.9819+0.85698)+(2.5419\times10^{-9}+1+0.01808+0.14302)} = 0.70972$$

Cluster 2:

$$\mu_2 = \frac{\sum_x p(c_2=1|x)x}{\sum_x p(c_2=1|x)} = \frac{2.5419\times10^{-9}\binom{2}{4} + 1\binom{-1}{-4} + 0.01808\binom{-1}{2} + 0.14302\binom{4}{0}}{2.5419\times10^{-9} + 1 + 0.01808 + 0.14302} = \begin{pmatrix}-0.38411\\ -3.41387\end{pmatrix}$$

$$\sum_{11}^2 = \frac{2.5419\times10^{-9}(2+0.38411)^2+1(-1+0.38411)^2+0.01808(-1+0.38411)^2+0.14302(4+0.38411)^2}{2.5419\times10^{-9} + 1 + 0.01808 + 0.14302} = 2.7001$$

$$\sum_{12}^2 = \frac{2.5419\times10^{-9}(2+0.38411)(4+3.41387)+1(-1+0.38411)(-4+3.41387)+0.01808(-1+0.38411)(2+3.41387)+0.14302(4+0.38411)(0+3.41387)}{2.5419\times10^{-9} + 1 + 0.01808 + 0.14302}$$

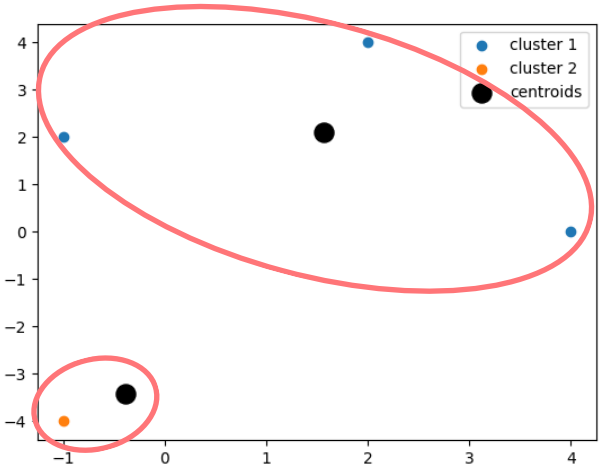
$$= 2.10254$$

$$\sum_{22}^2 = \frac{2.5419\times10^{-9}(4+3.41387)^2+1(-4+3.41387)^2+0.01808(2+3.41387)^2+0.14302(0+3.41387)^2}{2.5419\times10^{-9} + 1 + 0.01808 + 0.14302} = 2.18784$$

$$\sum_2 = \begin{pmatrix}2.7001 & 2.10254\\ 2.10254 & 2.18784\end{pmatrix}$$

$$\text{Prior: } p(c_2=1) = \frac{2.5419\times10^{-9}+1+0.01808+0.14302}{(0.9999999974578+0+0.9819+0.85698)+(2.5419\times10^{-9}+1+0.01808+0.14302)} = 0.29028$$

Com os novos valores para os priors e likelihoods poderíamos passar à próxima iteração.



TÉCNICO LISBOA

Aprendizagem 2021/22

Homework IV – Group 057

2) $a(i) = \frac{1}{|c_i|-1}\sum_{j\in c_i, j\neq i} d(i,j)$

$b(i) = \min_{k\neq i} \frac{1}{|c_k|}\sum_{j\in c_k} d(i,j)$

$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$, if $|c_i| > 1$, else 0

$a(x_1) = \frac{1}{3-1}\left(d(x1,x3)+d(x1,x4)\right) = \frac{1}{2}\left(\sqrt{(2-(-1))^2+(4-2)^2}+\sqrt{(2-4)^2+(4-0)^2}\right) = 3.6736$

$b(x_1) = \frac{d(x1,x2)}{1} = \sqrt{(2-(-1))^2+(4-(-4))^2} = 8.544$

$s(x_1) = \frac{8.544-3.6736}{8.544} = 0.570$

$s(x_2) = 0$

$a(x_3) = \frac{d(x3,x1)+d(x3,x4)}{2} = \frac{\sqrt{(-1-2)^2+(2-4)^2}+\sqrt{(-1+4)^2+(2-0)^2}}{2} = 3.60555$

$b(x_3) = \frac{d(x3,x2)}{1} = \sqrt{(-1-(-1))^2+(2-(-4))^2} = 6$

$s(x_3) = \frac{6-3.60555}{6} = 0.3991$

$a(x_4) = \frac{d(x4,x1)+d(x4,x3)}{2} = \frac{\sqrt{(2-4)^2+(4-0)^2}+\sqrt{(-1+4)^2+(2-0)^2}}{2} = 4.0388$

$b(x_4) = \frac{d(x4,x2)}{1} = \sqrt{(-1-4)^2+(-4-0)^2} = 6.4031$

$s(x_4) = \frac{6.4031-4.0388}{6.4031} = 0.3692$

$s(c_1) = \frac{0.570+0.3991+0.3692}{3} = 0.4461$

$s(c_2) = s(x2) = 0$

3)

a)

i) $W^{[1]}=5\times5$, $b^{[1]}=5\times1$, $W^{[2]}=5\times5$, $b^{[2]}=5\times1$, $W^{[3]}=5\times5$, $b^{[3]}=5\times1$, $W^{[4]}=2\times5$, $b^{[4]}=2\times1$

$5\times5 = 25$ parâmetros, $5\times1=5$ parâmetros, $2\times5 = 10$ parâmetros, $2\times1 = 2$ parâmetros

Logo, no total teremos $25+5+25+5+25+5+10+2 = 102$ parâmetros, que é aproximadamente a VC dimension.

ii) $d_{VC}(\text{DecisionTree}) = 2^d$

$d = 2^5 = 32$

iii) 1 parâmetro para o prior (o prior é uma tabela com duas entradas, uma para cada classe). Para a likelihood, é preciso o vetor das médias, que é 5×1 logo 5 parâmetros; e a matriz de covariância 5×5, no entanto, como a matriz é simétrica, precisamos apenas da diagonal e parte superior da matriz, o que nos dá 15 parâmetros.

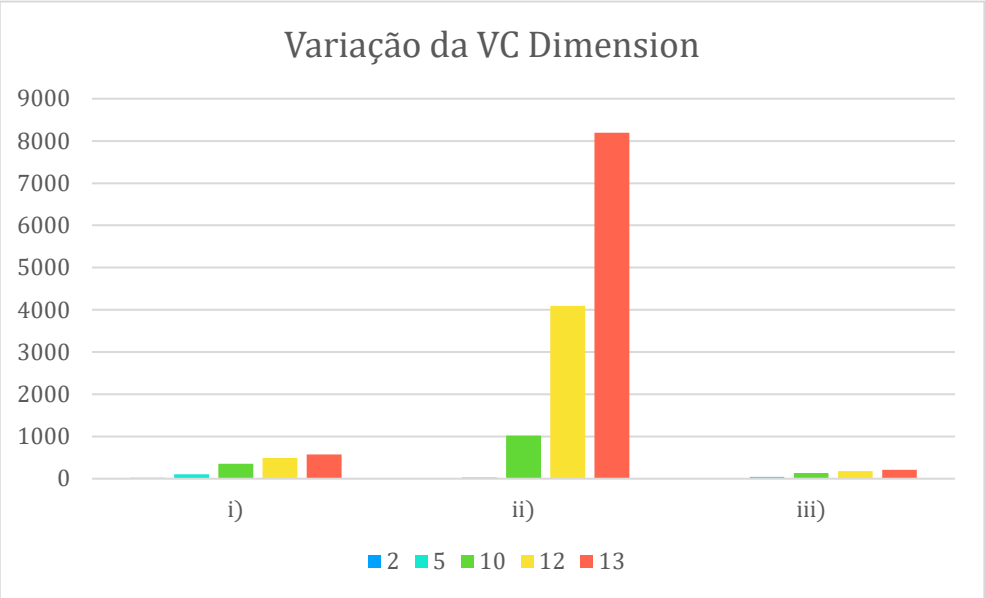
No total, teremos $2\times(5+15)+1 = 41$ parâmetros, ou seja, VC dimension de aproximadamente 41.

b)

m	2	5	10	12	13
VC dimension i)	24	102	352	494	574
VC dimension ii)	4	32	1024	4096	8192
VC dimension iii)	11	41	131	181	209

Aprendizagem 2021/22

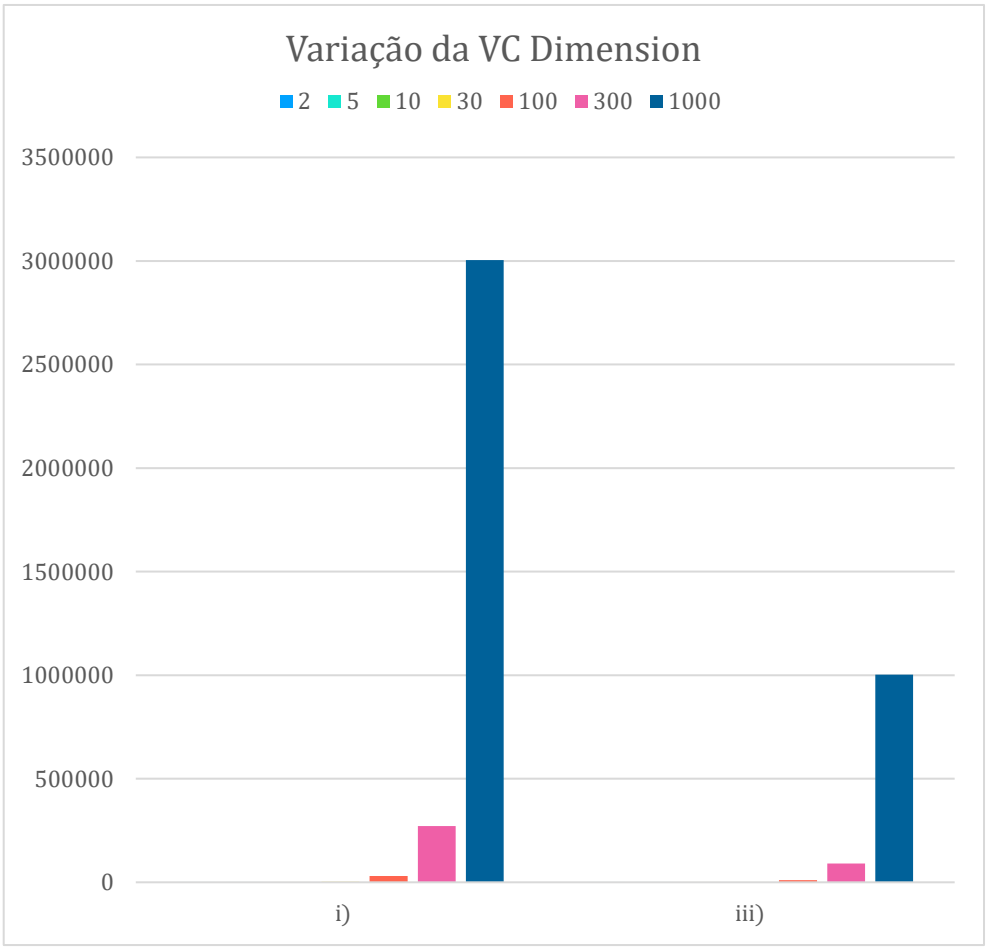
Homework IV – Group 057



Concluímos que na Decision Tree obtemos valores de VC Dimension elevados.

c)

m	2	5	10	30	100	300	1000
VC dimension i)	24	102	352	2 852	30 502	271 502	3 005 002
VC dimension iii)	11	41	131	991	10 301	90 901	1 003 001



Há uma diferença significativa nos resultados, mais visível com valores de m tão elevados como por exemplo 1000, sendo os valores no MLP muito superiores aos do Bayesian classifier.

Aprendizagem 2021/22

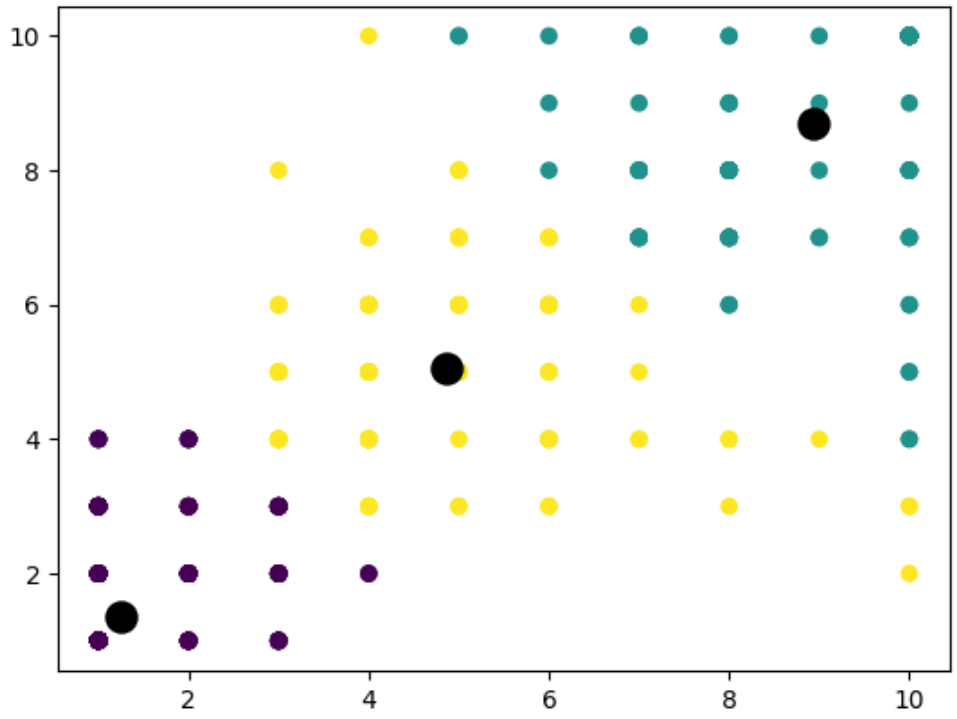
Homework IV – Group 057

II. Programming and critical analysis

4) **Para k=2:**
Silhouette score: 0.597
ECR: 13.5
Para k=3:
Silhouette score: 0.524
ECR: 6.667

Observamos que tanto os valores obtidos pela silhueta como pelo ECR são mais elevados para k=2 do que para k=3.

5)



6) Observando os clusters, representados na alínea anterior, verificamos que os 3 clusters se encontram bem separados, sem que haja pontos de um cluster dentro de outro. No entanto, também não há uma distância muito grande entre os diferentes clusters, o que demonstra uma boa separação, mas não é ainda a ideal.
Também notamos que em geral os pontos de um mesmo cluster se encontram próximos uns dos outros, ou seja, os clusters são compactos, há boa coesão, principalmente nos clusters com os pontos representados a roxo e azul (o cluster amarelo já não é tão coeso).
Calculámos o coeficiente de silhueta e obtivemos um valor de 0.707, o que vai de encontro ao que observamos empíricamente: é um valor positivo, acima de 0, o que significa que não há clusters sobrepostos, mas ainda não está muito próximo do valor ideal de 1.
Podemos assim concluir que aplicando k-means com k=3 a este data-set é possível e fornece bons resultados, mas não excelentes.

TÉCNICO LISBOA

Aprendizagem 2021/22

Homework IV – Group 057

III. APPENDIX

```
# ex 1
from matplotlib import pyplot as plt

# cluster 1
c1 = plt.scatter([2, -1, 4], [4, 2, 0])
# cluster 2
c2 = plt.scatter([-1], [-4])
# centroids
c = plt.scatter([1.56611, -0.38411], [2.100757, -3.41387], color = 'black', s=150)

plt.legend((c1, c2, c), ('cluster 1', 'cluster 2', 'centroids'))

plt.show()
```

```
# ex 4 and 5
from scipy.io import arff
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.feature_selection import mutual_info_classif
from sklearn.feature_selection import SelectKBest

data = arff.loadarff('breast.w_modified.arff')
df = pd.DataFrame(data[0])

X = df.iloc[:, 0:9]
y = df.iloc[:, -1]
y = y.astype('string')

def scores(k):
    km = KMeans(n_clusters=k)
    km.fit_predict(X)
    score = silhouette_score(X, km.labels_)
    print("silhouette k=" + str(k) + " score: " + str(score))

    mydict = {i: np.where(km.labels_ == i)[0] for i in range(km.n_clusters)}
    print("ecr k=" + str(k) + " score: " + str(ecr(k, mydict)))


def ecr(k, clusters):
    sum = 0

    for i in range(k):
        benigns = 0
        malignants = 0
        for j in clusters[i]:
            if y[j] == "b'benign'":
                benigns += 1
            elif y[j] == "b'malignant'":
                malignants += 1
        sum += len(clusters[i]) - max(malignants, benigns)

    return sum/k
```

TÉCNICO LISBOA

Aprendizagem 2021/22

Homework IV – Group 057

```
def plot():
    X_new = SelectKBest(mutual_info_classif, k=2).fit_transform(X, y)
    km = KMeans(n_clusters=3)
    km.fit(X_new)
    y_kmeans = km.predict(X_new)

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

    centers = km.cluster_centers_
    plt.scatter(centers[:, 0], centers[:, 1], c='black', s=150);

    plt.show()


scores(2)
scores(3)
plot()
```

END